CLC **Drug Discovery** Workbench

USER MANUAL

Manual for

CLC Drug Discovery Workbench 3.0

Windows, Mac OS X and Linux

May 10, 2016

This software is for research purposes only.

QIAGEN Aarhus A/S
Silkeborgvej 2
Prismet
DK-8000 Aarhus C
Denmark

Contents

Part I

Introduction

Chapter 1

Introduction to *CLC Drug Discovery Workbench*

Contents

CLC Drug Discovery Workbench is a virtual lab bench. It gives you access to atomic level insights in protein-ligand interaction, and allows new ideas for improved binders to be quickly tested and visualized.

CLC Drug Discovery Workbench comes with drug design and sequence analysis tools that allow you to analyze and visualize protein targets and ligands binding to them. The interface is designed to communicate with all chemists, with no assumptions about their level of theoretical training.

We strongly encourage you to read this user manual in order to get the best possible basis for working with the software package.

This software is for research purposes only.

1.1 Contact information

The *CLC Drug Discovery Workbench* is developed by:

QIAGEN Aarhus
Silkeborgvej 2
Prismet
8000 Aarhus C
Denmark

http://www.clcbio.com
http://www.qiagenbioinformatics.com

VAT no.: DK 28 30 50 87

Email: support-clcbio@qiagen.com
Telephone: +45 70 22 32 44

If you have questions or comments regarding the program, you can contact us through the support team as described here: http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Getting_help.html.

1.2 Download and installation

The *CLC Drug Discovery Workbench* is developed for Windows, Mac OS X and Linux. The software for either platform can be downloaded from http://www.clcbio.com/download.

1.2.1 Program download

The program is available for download on http://www.clcbio.com/download.

Before you download the program you are asked to fill in the **Download** dialog.

In the dialog you must choose:

- Which operating system you use
- Whether you would like to receive information about future releases

Depending on your operating system and your Internet browser, you are taken through some download options.

When the download of the installer (an application which facilitates the installation of the program) is complete, follow the platform specific instructions below to complete the installation procedure.

1.2.2 Installation on Microsoft Windows

Starting the installation process is done in one of the following ways:

When you have downloaded an installer:
Locate the downloaded installer and double-click the icon.
The default location for downloaded files is your desktop.

Installing the program is done in the following steps:

- On the welcome screen, click **Next**.
- Read and accept the License agreement and click **Next**.
- Choose where you would like to install the application and click **Next**.
- Choose a name for the Start Menu folder used to launch *CLC Drug Discovery Workbench* and click **Next**.
- Choose if *CLC Drug Discovery Workbench* should be used to open CLC files and click **Next**.
- Choose where you would like to create shortcuts for launching *CLC Drug Discovery Workbench* and click **Next**.
- Choose if you would like to associate .clc files to *CLC Drug Discovery Workbench*. If you check this option, double-clicking a file with a "clc" extension will open the *CLC Drug Discovery Workbench*.
- Wait for the installation process to complete, choose whether you would like to launch *CLC Drug Discovery Workbench* right away, and click **Finish**.

When the installation is complete the program can be launched from the Start Menu or from one of the shortcuts you chose to create.

1.2.3 Installation on Mac OS X

Starting the installation process is done in the following way:

When you have downloaded an installer:
Locate the downloaded installer and double-click the icon.
The default location for downloaded files is your desktop.
Launch the installer by double-clicking on the "*CLC Drug Discovery Workbench*" icon.

Installing the program is done in the following steps:

- On the welcome screen, click **Next**.
- Read and accept the License agreement and click **Next**.
- Choose where you would like to install the application and click **Next**.
- Choose if *CLC Drug Discovery Workbench* should be used to open CLC files and click **Next**.
- Choose whether you would like to create desktop icon for launching *CLC Drug Discovery Workbench* and click **Next**.
- Choose if you would like to associate .clc files to *CLC Drug Discovery Workbench*. If you check this option, double-clicking a file with a "clc" extension will open the *CLC Drug Discovery Workbench*.

- Wait for the installation process to complete, choose whether you would like to launch *CLC Drug Discovery Workbench* right away, and click **Finish**.

When the installation is complete the program can be launched from your Applications folder, or from the desktop shortcut you chose to create. If you like, you can drag the application icon to the dock for easy access.

1.2.4 Installation on Linux with an installer

Navigate to the directory containing the installer and execute it. This can be done by running a command similar to:

```
# sh CLCDrugDiscoveryWorkbench_3_0_64.sh.sh
```

Installing the program is done in the following steps:

- On the welcome screen, click **Next**.
- Read and accept the License agreement and click **Next**.
- Choose where you would like to install the application and click **Next**.
  For a system-wide installation you can choose for example /opt or /usr/local. If you do not have root privileges you can choose to install in your home directory.
- Choose where you would like to create symbolic links to the program
  DO NOT create symbolic links in the same location as the application.
  Symbolic links should be installed in a location which is included in your environment PATH. For a system-wide installation you can choose for example /usr/local/bin. If you do not have root privileges you can create a 'bin' directory in your home directory and install symbolic links there. You can also choose not to create symbolic links.
- Wait for the installation process to complete and click **Finish**.

If you choose to create symbolic links in a location which is included in your PATH, the program can be executed by running the command:

```
# clcdrugdiscoverywb2
```

Otherwise you start the application by navigating to the location where you choose to install it and running the command:

```
# ./clcdrugdiscoverywb2
```

1.3 System requirements

- Windows Vista, Windows 7, Windows 8, Windows 10, Windows Server 2008, or Windows Server 2012
- Mac OS X 10.7 or later

- Linux: RHEL 5.0 or later. SUSE 10.2 or later. Fedora 6 or later
- 2 GB RAM required
- 4 GB RAM recommended
- 1024 x 768 display required
- 1600 x 1200 display recommended
- Intel or AMD CPU required

- **3D Graphics Requirements**
  - A graphics card capable of supporting OpenGL 2.0. Note that *CLC Drug Discovery Workbench* only uses the GPU for the OpenGL 3D rendering. The GPU is not used to speed up molecular simulations.
  - Updated graphics drivers. Please make sure the latest driver for the graphics card is installed .
- **3D Graphics Recommendations**
  - A discrete graphics card from either Nvidia or AMD/ATI. Modern integrated graphics cards (such as the Intel HD Graphics series) may also be used, but these are usually slower than the discrete cards.
  - A 64-bit workbench version is recommended for working with large complexes.

1.3.1 Limitations on maximum number of cores

Most modern CPUs implements hyper threading or a similar technology which makes each physical CPU core appear as two logical cores on a system. In this manual the therm "core" always refer to a logical core unless otherwise stated.

For static licenses, there is a limitation on the number of logical cores on the computer. If there are more than 64 logical cores, the *CLC Drug Discovery Workbench* cannot be started. In this case, a network license is needed (read more at http://www.clcbio.com/desktop-applications/licensing/).

1.4 Workbench Licenses

When you have installed the *CLC Drug Discovery Workbench*, and start it for the first time or after installing a new major release, you will meet the license assistant, shown in figure 1.1. The **License Manager** can also be accessed from the menu bar in the Workbench:

Help | **License Manager**

This can be useful if you wish to use a different license or want to view information about the license(s) the Workbench is currently using. The **License Manager** is described in detail in section 1.4.5 and can be seen in figure 1.23.

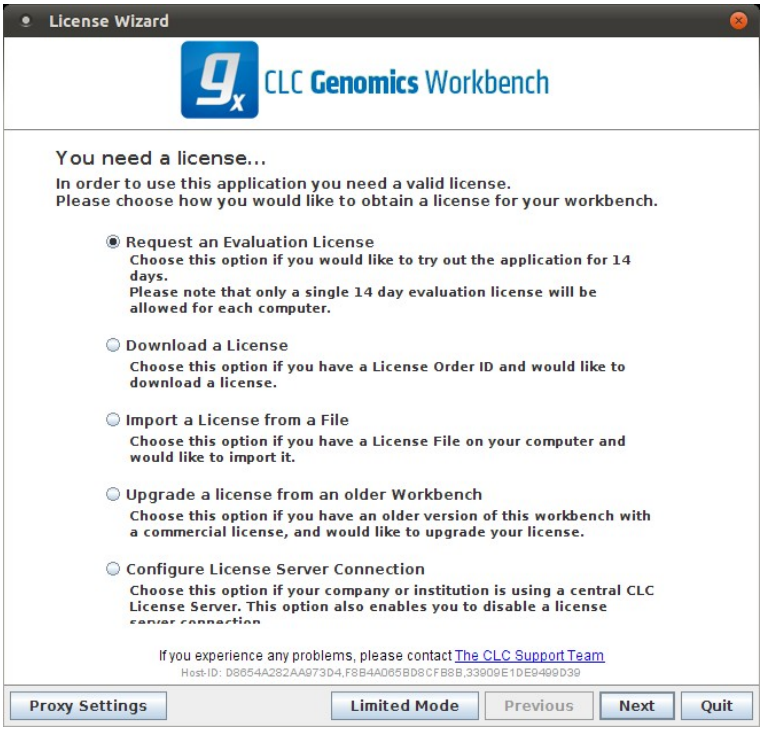


Figure 1.1: *The license assistant showing you the options for getting started.*

To install a license, you must be running the program in administrative mode [1].

The following options are available. They are described in detail in the sections that follow.

- **Request an evaluation license**. Request a fully functional, time-limited license (see below).
- **Download a license**. Use the license order ID received when you purchase the software to download and install a license file.
- **Import a license from a file**. Import an existing license file, for example a file downloaded from the web-based licensing system.
- **Upgrade license**. If you have used a previous version of the *CLC Drug Discovery Workbench*, and you are entitled to upgrade to a new major version, select this option to upgrade your license file.
- **Configure license server connection**. If your organization has a CLC License Server, select this option to configure the connection to it.

Select the appropriate option and click on button labeled **Next**.

To use the Download option in the License Manager, your machine must be able to access the external network. If this is not the case, please see section 1.4.6.

If for some reason you don't have a license order ID or access to a license, you can click the **Limited Mode** button (see section 1.4.7).

[1] How to do this differs for different operating systems. To run the program in administrator mode on Windows Vista, or 7, right-click the program shortcut and choose "Run as Administrator.

1.4.1 Request an evaluation license

We offer a fully functional version of the *CLC Drug Discovery Workbench* for evaluation purposes, free of charge.

Each user is entitled to a 14-day trial of *CLC Drug Discovery Workbench*.

If you are unable to complete your assessment in the available time, please send an email to bioinformaticssales@qiagen.com to request an additional evaluation period.

When you choose the option **Request an evaluation license**, you will see the dialog shown in figure 1.2.

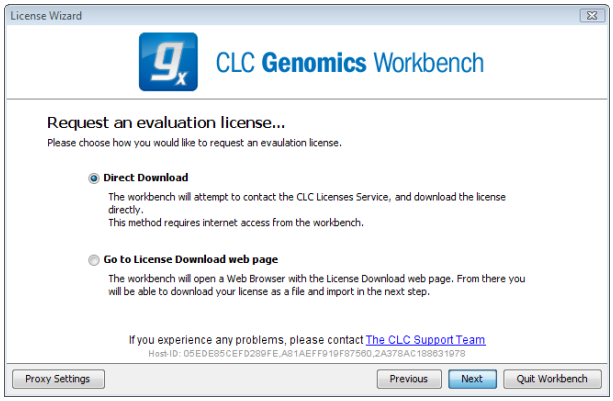


Figure 1.2: *Choosing between direct download or going to the license download web page.*

In this dialog, there are two options:

- **Direct download**. Download the license directly from CLC bio. This method requires that the Workbench has access to the external network.
- **Go to license download web page**. In a browser window, show the license download web page, which can be used to download a license file. This option is suitable in situations where, for example, you are working behind a proxy, so that the Workbench does not have direct access to the CLC Licenses Service.

If you select the option to download a license directly and it turns out that the Workbench does not have direct access to the external network, (because of a firewall, proxy server etc.), you can click **Previous** button to try the other method.

After selection on your method of choice, click on the button labeled **Next**.

Direct download

After choosing the **Direct Download** option and clicking on the button labeled **Next**, the dialog shown in figure 1.3 appears.

A progress for getting the license is shown, and when the license is downloaded, you will be able to click **Next**.

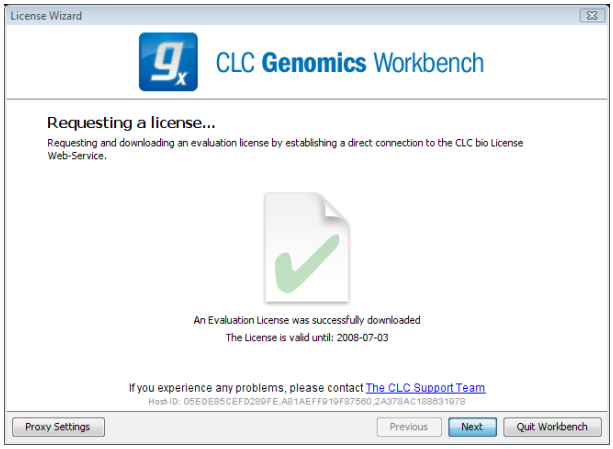


Figure 1.3: *A license has been downloaded.*

Go to license download web page

After choosing the *Go to license download web page* option and clicking on the button labeled **Next**, the license download web page appears in a browser window, as shown in 1.4.

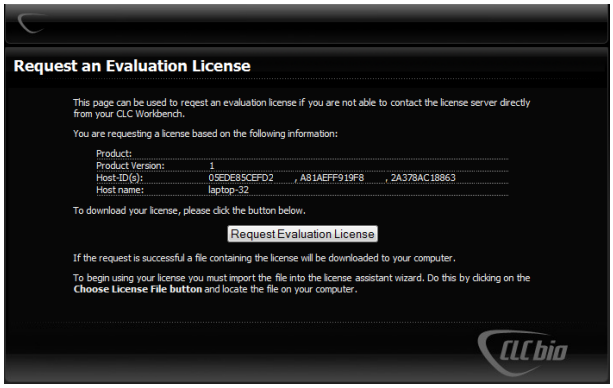


Figure 1.4: *The license download web page.*

Click the **Request Evaluation License** button. You can then save the license on your system.

Back in the Workbench window, you will now see the dialog shown in 1.5.

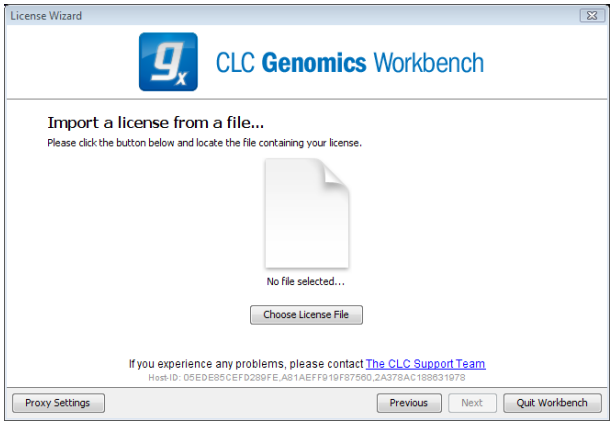


Figure 1.5: *Importing the license file downloaded from the web page.*

Click the **Choose License File** button and browse to find the license file you saved. When you have selected the file, click on the button labeled **Next**.

Accepting the license agreement

Part of the installation of the license involves checking and accepting the end user license agreement (EULA). You should now see the a window like that in figure 1.6.

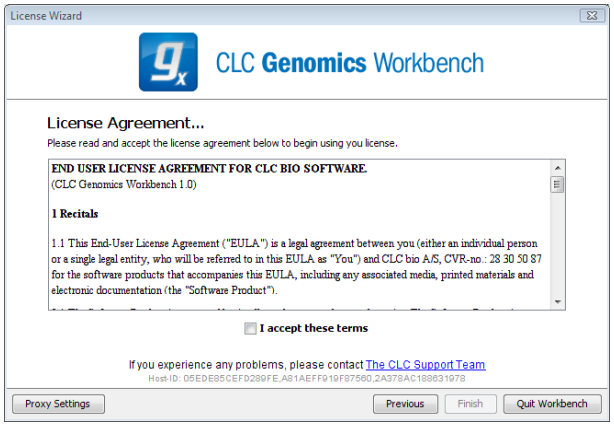


Figure 1.6: *Read the license agreement carefully.*

Please read the EULA text carefully before clicking in the box next to the text **I accept these terms** to accept, and then clicking on the button labeled **Finish**.

1.4.2 Download a license using a license order ID

Using a license order ID, you can download a license file via the Workbench or using an online form. When you have chosen this option and clicked **Next** button, you will see the dialog shown in 1.7. Enter your license order ID into the text field under the title License Order-ID. (The ID can be pasted into the box after copying it and then using menus or key combinations like Ctrl+V on some system or ⌘ + V on Mac).

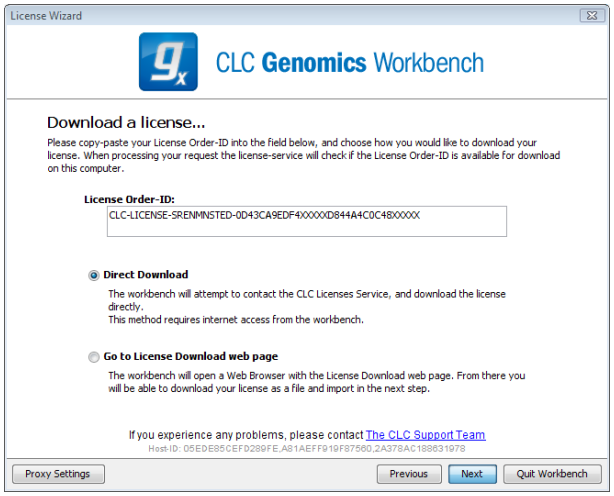


Figure 1.7: *Enter a license order ID for the software.*

In this dialog, there are two options:

- **Direct download**. Download the license directly from CLC bio. This method requires that the Workbench has access to the external network.

- **Go to license download web page**. In a browser window, show the license download web page, which can be used to download a license file. This option is suitable in situations where, for example, you are working behind a proxy, so that the Workbench does not have direct access to the CLC Licenses Service.

If you select the option to download a license directly and it turns out that the Workbench does not have direct access to the external network, (because of a firewall, proxy server etc.), you can click **Previous** button to try the other method.

After selection on your method of choice, click on the button labeled **Next**.

Direct download

After choosing the **Direct Download** option and clicking on the button labeled **Next**, the dialog shown in figure 1.8 appears.

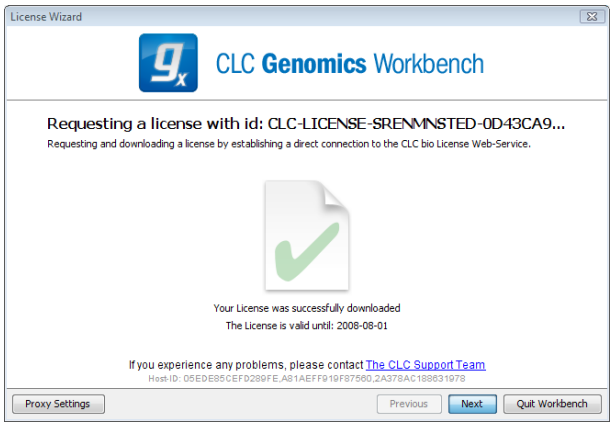


Figure 1.8: *A license has been downloaded.*

A progress for getting the license is shown, and when the license is downloaded, you will be able to click **Next**.

Go to license download web page

After choosing the *Go to license download web page* option and clicking on the button labeled **Next**, the license download web page appears in a browser window, as shown in 1.9.

Click the **Request Evaluation License** button. You can then save the license on your system.

Back in the Workbench window, you will now see the dialog shown in 1.10.

Click the **Choose License File** button and browse to find the license file you saved. When you have selected the file, click on the button labeled **Next**.

Accepting the license agreement

Part of the installation of the license involves checking and accepting the end user license agreement (EULA). You should now see the a window like that in figure 1.11.

Please read the EULA text carefully before clicking in the box next to the text **I accept these terms** to accept, and then clicking on the button labeled **Finish**.

Figure 1.9: *The license download web page.*

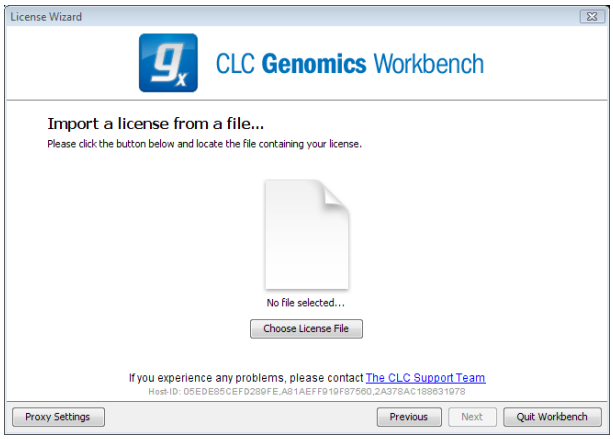


Figure 1.10: *Importing the license file downloaded from the web page.*

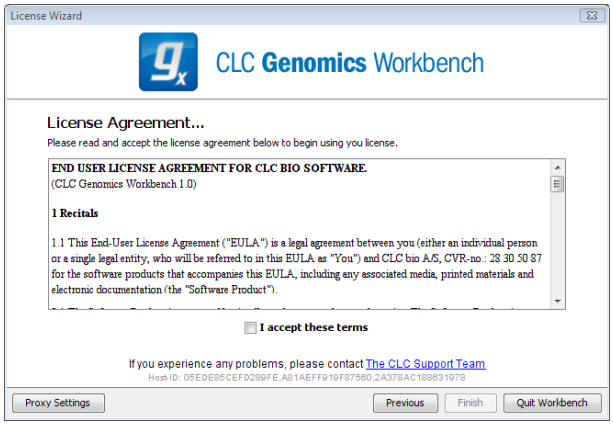


Figure 1.11: *Read the license agreement carefully.*

1.4.3 Import a license from a file

If you already have a license file associated with the host ID of your machine,it can be imported using this option.

When you have clicked on the **Next** button, you will see the dialog shown in 1.12.

Click the **Choose License File** button and browse to find the license file. When you have selected the file, click on the **Next** button.

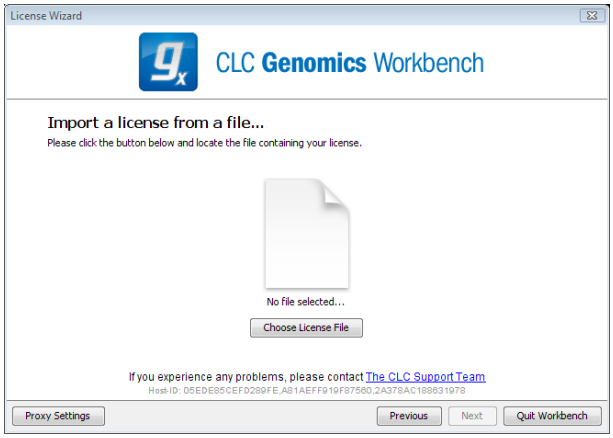


Figure 1.12: *Selecting a license file .*

Accepting the license agreement

Part of the installation of the license involves checking and accepting the end user license agreement (EULA). You should now see the a window like that in figure 1.13.

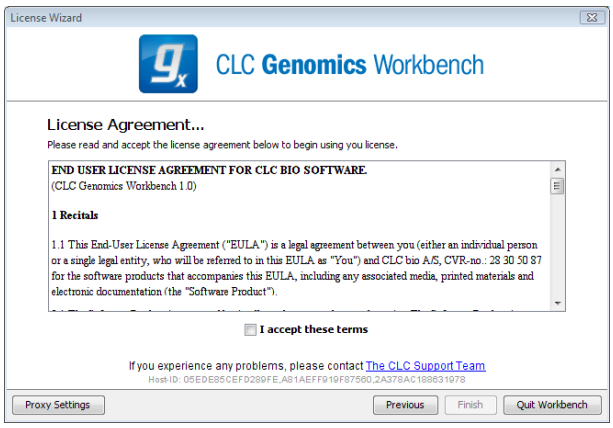


Figure 1.13: *Read the license agreement carefully.*

Please read the EULA text carefully before clicking in the box next to the text **I accept these terms** to accept, and then clicking on the button labeled **Finish**.

1.4.4 Upgrade license

This option is used when you already have used a previous version of *CLC Drug Discovery Workbench*, and you are entitled to upgrade to a new major version. The Workbench will need direct access to the external network to use this option.

When you click on the **Next** button, the Workbench will search for a previous installation of *CLC Drug Discovery Workbench*. It will then locate the old license.

If the Workbench finds an existing license file, the next dialog will look like figure 1.14.

When you click on the **Next** button, the Workbench checks if you are entitled to upgrade your license. This is done by contacting CLC bio's servers.

If the Workbench cannot connect to the external network directly, please see the section on downloading a license for non-networked machines. You will need your license order ID for this.

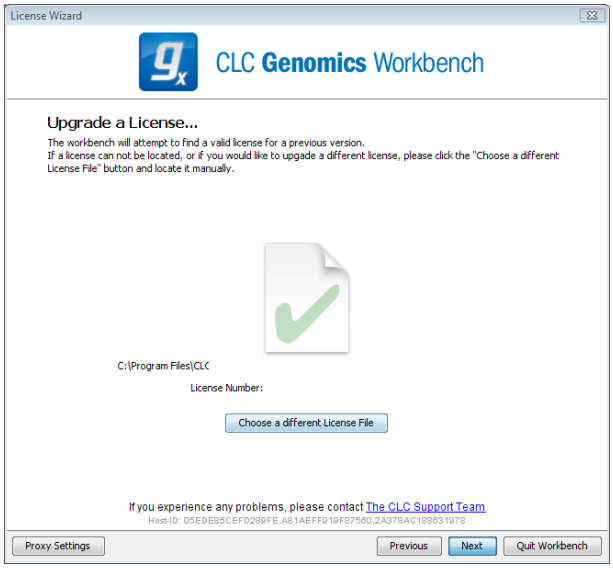


Figure 1.14: *An license from an older installation is found.*

Your license must be covered by our Maintenance, Upgrades and Support (MUS) program to be eligible to upgrade your license. If the license is covered for upgrades and there are any problems with this, please contact licenses@clcbio.com.

In this dialog, there are two options:

- **Direct download**. Download the license directly from CLC bio. This method requires that the Workbench has access to the external network.
- **Go to license download web page**. In a browser window, show the license download web page, which can be used to download a license file. This option is suitable in situations where, for example, you are working behind a proxy, so that the Workbench does not have direct access to the CLC Licenses Service.

If you select the option to download a license directly and it turns out that the Workbench does not have direct access to the external network, (because of a firewall, proxy server etc.), you can click **Previous** button to try the other method.

After selection on your method of choice, click on the button labeled **Next**.

Direct download

After choosing the **Direct Download** option and clicking on the button labeled **Next**, the dialog shown in figure 1.15 appears.

A progress for getting the license is shown, and when the license is downloaded, you will be able to click **Next**.

Go to license download web page

After choosing the *Go to license download web page* option and clicking on the button labeled **Next**, the license download web page appears in a browser window, as shown in 1.16.

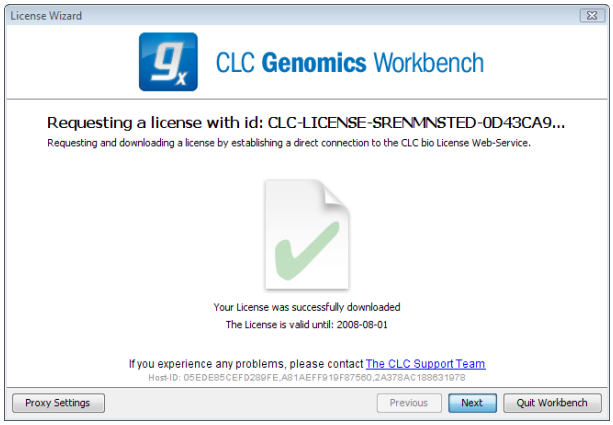


Figure 1.15: *A license has been downloaded.*

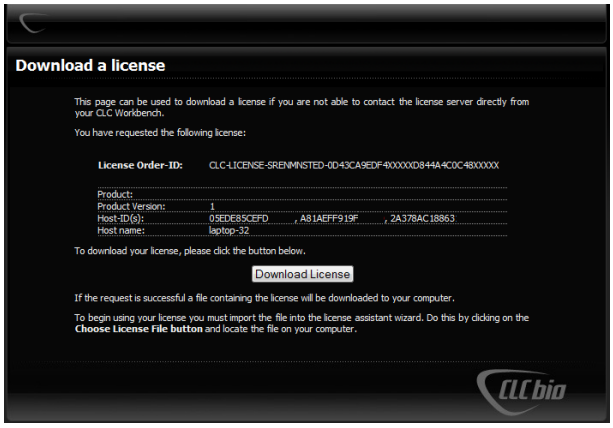


Figure 1.16: *The license download web page.*

Click the **Request Evaluation License** button. You can then save the license on your system.

Back in the Workbench window, you will now see the dialog shown in 1.17.

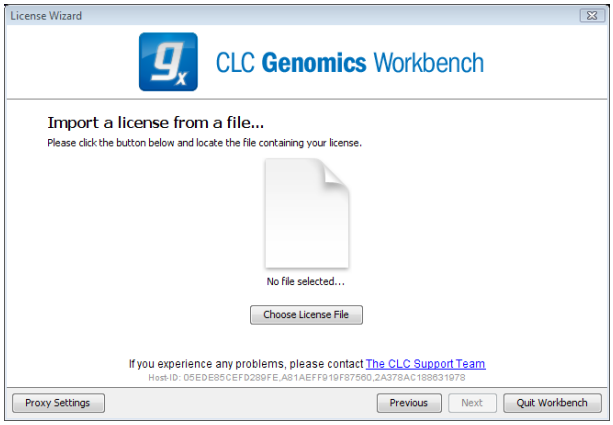


Figure 1.17: *Importing the license file downloaded from the web page.*

Click the **Choose License File** button and browse to find the license file you saved. When you have selected the file, click on the button labeled **Next**.

Accepting the license agreement

Part of the installation of the license involves checking and accepting the end user license agreement (EULA). You should now see the a window like that in figure 1.18.

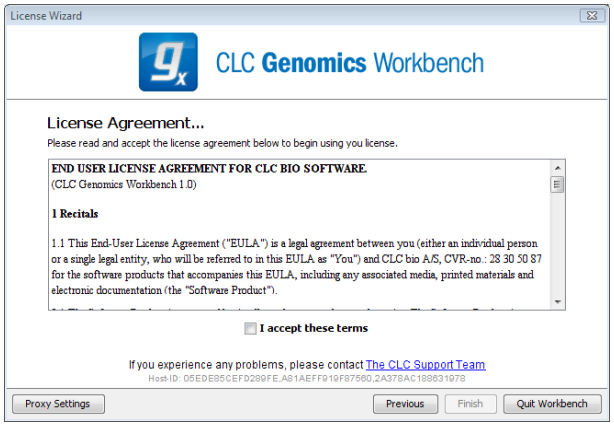


Figure 1.18: *Read the license agreement carefully.*

Please read the EULA text carefully before clicking in the box next to the text **I accept these terms** to accept, and then clicking on the button labeled **Finish**.

1.4.5 Configure license server connection

If your organization is running a CLC License Server, you can configure your Workbench to connect to it to get a license.

To do this, select this option and click on the **Next** button. A dialog like that shown in figure 1.19 then appears. Here, you configure how to connect to the CLC License Server.

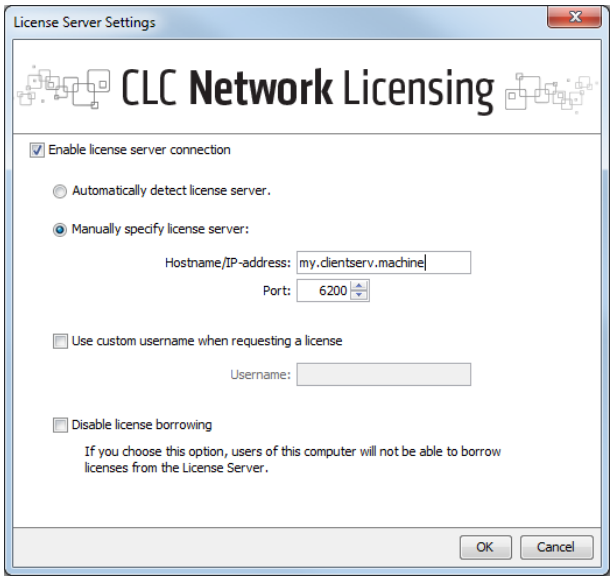


Figure 1.19: *Connecting to a CLC License Server.*

- **Enable license server connection**. This box must be checked for the Workbench is to contact the CLC License Server to get a license for *CLC Drug Discovery Workbench*.

- **Automatically detect license server**. By checking this option the Workbench will look for a CLC License Server accessible from the Workbench[2].
- **Manually specify license server**. If there are technical limitations such that the CLC License Server cannot be detected automatically, use this option to provides details of machine the CLC License Server software is on, and the port used by the software to receive requests. After selecting this option, please enter:
  - **Host name**. The address for the machine the CLC Licenser Server software is running on.
  - **Port**. The port used by the CLC License Server to receive requests.
- **Use custom username when requesting a license**. A username entered here will be passed to the CLC License Server instead of the username of your account this machine.
- **Disable license borrowing on this computer**. If you do not want users of the computer to borrow a license from the set of licenses available, then (see section 1.4.5), select this option.

Borrowing a license

A network license can only be used when you are connected to the a license server. If you wish to use the *CLC Drug Discovery Workbench* when you are not connected to the CLC License Server, you can *borrow* an available license for a period of time. During this time, there will be one less network license available on the for other users. The Workbench must have a connection to the CLC License Server at the point in time when you wish to borrow a license.

The procedure for borrowing a license is:

1. Go to the Workbench menu option:

   Help | **License Manager**
2. Click on the "Borrow License" tab to display the dialog shown in figure 1.20.
3. Use the checkboxes at the right hand side of the table in the License overview section of the window to select the license(s) that you wish to borrow.
4. Select the length of time you wish to borrow the license(s).
5. Click on the button labeled **Borrow Licenses**.
6. Close the License Manager when you are done.

You can now go offline and work with the *CLC Drug Discovery Workbench*. When the time period you borrowed the license for has elapsed, the network license you borrowed is made available

[2]Automatic server discovery sends UDP broadcasts from the Workbench on a fixed port, 6200. Available license servers respond to the broadcast. The Workbench then uses TCP communication for to get a license, assuming one is available. Automatic server discovery works only on local networks and will not work on WAN or VPN connections. Automatic server discovery is not guaranteed to work on all networks. If you are working on an enterprise network on where local firewalls or routers cut off UDP broadcast traffic, then you may need to configure the details of the CLC License server manually instead.

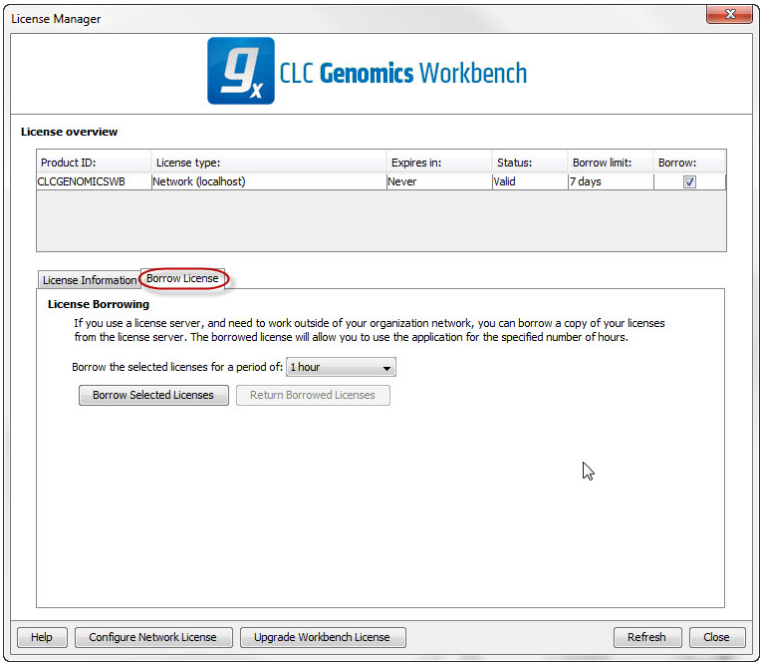


Figure 1.20: *Borrow a license.*

again for other users to access. To continue using the *CLC Drug Discovery Workbench* with a license, you will need to connect the Workbench to the network again so it can contact the CLC Licene Server to obtain one.

Note! Your CLC License Server administrator can choose to disable to the option allowing the borrowing of licenses. If this has been done, you will not be able to borrow a network license using your Workbench.

Common issues when using a network license

No license available at the moment If all the network licenses or *CLC Drug Discovery Workbench*are in use, you will see a dialog like that shown in figure 1.21 when you start up the Workbench.

This means others are using the network licenses. You will need to wait for them to return their licenses before you can continue to work with a fully functional copy of software. If this is a frequent issue, you may wish to discuss this with your CLC License Server administrator.

Clicking on the **Limited Mode** button in the dialog allows you to start the Workbench with functionality equivalent to the CLC Sequence Viewer. This includes the ability to access your CLC data.

Lost connection to the CLC License Server If the Workbench connection to the CLC License Server is lost, you will see a dialog as shown in figure 1.22.

If you have chosen the option to **Automatically detect license server** and you have not succeeded in connecting to the License Server before, please check with your local IT support that automatic detection will be possible to do at your site. If it is not possible at your site, you will need to manually configure the CLC License Server settings using the License Manager, as described

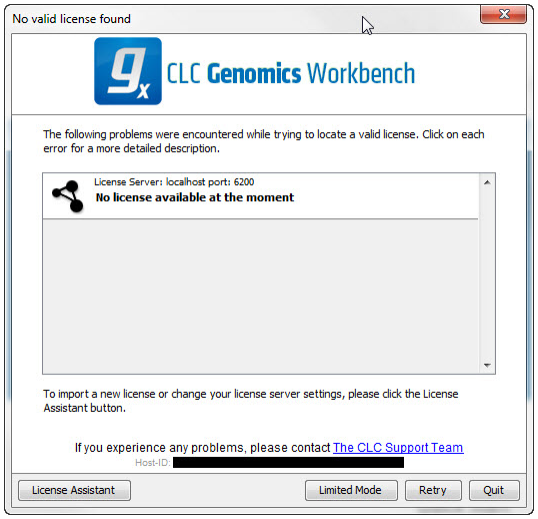


Figure 1.21: *This window appears when there are no available network licenses for the software you are running.*

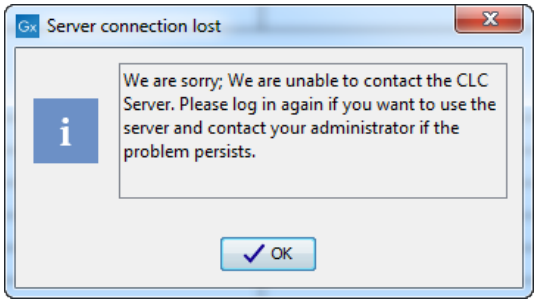


Figure 1.22: *This message appears if the Workbench is unable to establish a connection to a CLC License server.*

earlier in this section.

If you have successfully contacted the CLC License Server from your Workbench previously, please consider discussing this issue with your CLC License Server administrator or your local IT support, to make sure that the CLC License Server is running and that your Workbench can connect to it. There may be situations where you wish to use a different license or view information about the license(s) the Workbench is currently using. To do this, open the License Manager using the menu option:

Help | License Manager ( )

The license manager is shown in figure 1.23.

This dialog can be used to:

- See information about the license (e.g. what kind of license, when it expires)
- Configure how to connect to a license server (**Configure License Server** the button at the lower left corner). Clicking this button will display a dialog similar to figure 1.19.
- Upgrade from an evaluation license by clicking the **Upgrade license** button. This will display the dialog shown in figure 1.1.
- Export license information to a text file.

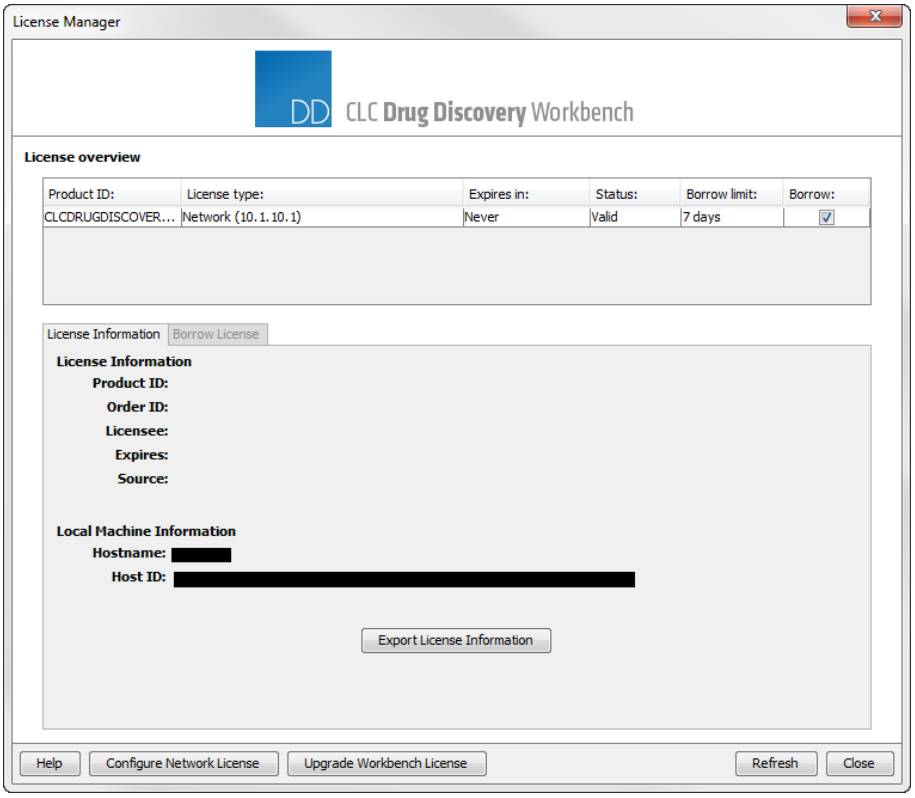


Figure 1.23: *The license manager.*

- Borrow a license

If you wish to switch away from using a network license, click on the button to **Configure License Server** and uncheck the box beside the text **Enable license server connection** in the dialog. When you restart the Workbench, you can set up the new license as described in section 1.4.

1.4.6 Download a static license on a non-networked machine

To download a static license for a machine that does not have direct access to the external network, you can follow the steps below:

- Install the *CLC Drug Discovery Workbench* on the machine you wish to run the software on.
- Start up the software as an administrative user and find the host ID of the machine that you will run the CLC Workbench on. You can see the host ID the machine reported at the bottom of the License Manager window in grey text.
- Make a copy of this host ID such that you can use it on a machine that has internet access.
- Go to a computer with internet access, open a browser window and go to the relevant network license download web page:
- For Workbenches released from January 2013 and later, (e.g. the Genomics Workbench version 6.0 or higher, and the Main Workbench, version 6.8 or higher), please go to:

  `https://secure.clcbio.com/LmxWSv3/GetLicenseFile`

For earlier Workbenches, including any DNA, Protein or RNA Workbench, please go to:

`http://licensing.clcbio.com/LmxWSv1/GetLicenseFile`

It is vital that you choose the license download page appropriate to the version of the software you plan to run.

- Paste in your license order ID and the host ID that you noted down in the relevant boxes on the webpage.
- Click 'download license' and save the resulting .lic file.
- Open the Workbench on your non-networked machine. In the Workbench license manager choose 'Import a license from a file'. In the resulting dialog click 'choose license file' to browse the location of the .lic file you have just downloaded.

  If the License Manager does not start up by default, you can start it up by going to the Help menu and choosing License Manager.
- Click on the **Next** button and go through the remaining steps of the license manager wizard.

1.4.7 Limited mode

We have created the limited mode to prevent a situation where you are unable to access your data because you do not have a license. When you run in limited mode, a lot of the tools in the Workbench are not available, but you still have access to your data . To get out of the limited mode and run the Workbench normally, restart the Workbench. When you restart the Workbench will try to find a proper license and if it does, it will start up normally. If it can't find a license, you will again have the option of running in limited mode.

1.5 About CLC Workbenches

In November 2005 CLC bio released two Workbenches: *CLC Free Workbench* and *CLC Protein Workbench*. *CLC Protein Workbench* is developed from the free version, giving it the well-tested user friendliness and look & feel. However, the *CLC Protein Workbench* includes a range of more advanced analyses.

In March 2006, *CLC DNA Workbench* (formerly *CLC Gene Workbench*) and *CLC Main Workbench* were added to the product portfolio of CLC bio. Like *CLC Protein Workbench*, *CLC DNA Workbench* builds on *CLC Free Workbench*. It shares some of the advanced product features of *CLC Protein Workbench*, and it has additional advanced features. *CLC Main Workbench* holds all basic and advanced features of the *CLC Workbenches*.

In June 2007, *CLC RNA Workbench* was released as a sister product of *CLC Protein Workbench* and *CLC DNA Workbench*. *CLC Main Workbench* now also includes all the features of *CLC RNA Workbench*.

In March 2008, the *CLC Free Workbench* changed name to *CLC Sequence Viewer*.

In June 2008, the first version of the *CLC Genomics Workbench* was released due to an extraordinary demand for software capable of handling sequencing data from all new high-throughput sequencing platforms such as Roche-454, Illumina and SOLiD in addition to Sanger reads and hybrid data.

For an overview of which features all the applications include, see http://www.clcbio.com/features.

In December 2006, CLC bio released a **Software Developer Kit** which makes it possible for anybody with a knowledge of programming in Java to develop plugins. The plugins are fully integrated with the CLC Workbenches and the Viewer and provide an easy way to customize and extend their functionalities.

In April 2012, *CLC Protein Workbench*, *CLC DNA Workbench* and *CLC RNA Workbench* were discontinued. All customers with a valid license for any of these products were offered an upgrade to the *CLC Main Workbench*.

In February 2014, CLC bio expanded the product repertoire with the release of *CLC Drug Discovery Workbench*, a product that enables studies of protein-ligand interactions for drug discovery.

In April 2014, CLC bio released the CLC Cancer Research Workbench, a product that containing streamlined data analysis workflows with integrated trimming and quality control tailored to meet the requirements of clinicians and researchers working within the cancer field.

In April 2015, the CLC Cancer Research Workbench was renamed to Biomedical Genomics Workbench to reflect the inclusion of tools addressing the requirements of clinicians and researchers working within the hereditary disease field in addition to the tools designed for those working within the cancer field.

1.5.1 New program feature request

The CLC team is continuously improving the *CLC Drug Discovery Workbench* with our users' interests in mind. We welcome all requests and feedback from users, as well as suggestions for new features or more general improvements to the program. To contact us via the Workbench, please go to the menu option:

Help | **Contact Support**

1.5.2 Getting help

If you encounter a problem or need help understanding how the *CLC Drug Discovery Workbench* works, and the license you are using is covered by our Maintenance, Upgrades and Support (MUS) program (https://www.clcbio.com/support/maintenance-support-program/), you can contact our customer support via the Workbench by going to the menu option:

Help | **Contact Support**

This will open a dialog to enter your contact information and a text field for entering the question or problem you have.

You can also attach small datasets, if this helps explain the problem or you believe it will help in troubleshooting the problem.

When you send a support request this way, it will include technical information about your installation that usually helps when troubleshooting. It also includes your license information so that you do not have to look this up yourself. Our support staff will reply to you by email.

Further information about Maintenance, Upgrades and Support (MUS) program can be found online at https://www.clcbio.com/support/maintenance-support-program/.

Information about how to to find your license information is included in the licenses section of our Frequently Asked Questions (FAQ) area: https://secure.clcbio.com/helpspot/index.php?pg=kb

Information about MUS cover on particular licenses can be found by https://secure.clcbio.com/myclc/login.

Start in safe mode

If the program becomes unstable on start-up, you can start it in **Safe mode**. This is done by pressing and holding down the Shift button while the program starts.

When starting in safe mode, the user settings (e.g. the settings in the **Side Panel**) are deleted and cannot be restored. Your data stored in the **Navigation Area** is not deleted. When started in safe mode, some of the functionalities are missing, and you will have to restart the *CLC Drug Discovery Workbench* again (without pressing Shift.

1.6 When the program is installed: Getting started

CLC Drug Discovery Workbench includes an extensive **Help** function, which can be found in the **Help** menu of the program's **Menu bar**. The **Help** can also be shown by pressing F1. The help topics are sorted in a table of contents and the topics can be searched.

Tutorials describing hands-on examples of how to use the individual tools and features of the *CLC Drug Discovery Workbench* can be found at http://www.clcbio.com/support/tutorials/. We also recommend our **Online presentations** where a product specialist from CLC bio demonstrates our software. This is a very easy way to get started using the program. Read more about video tutorials and other online presentations here: http://www.clcbio.tv/.

1.6.1 Quick start

When the program opens for the first time, the background of the workspace is visible. In the background are two quick start shortcuts, which will help you getting started. These can be seen in figure 1.24.

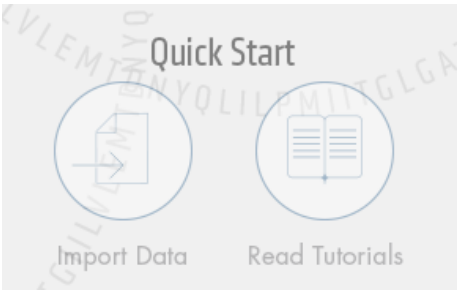


Figure 1.24: *Quick start short cuts, available in the background of the workspace.*

The function of the quick start shortcuts is explained here:

- **Import data.** Opens the **Import** dialog, which you let you browse for, and import data from your file system.

- **Read tutorials.** Opens the tutorials menu with a number of tutorials. These are also available from the **Help** menu in the **Menu bar**.

Below these three quick start shortcuts, you will see a text: "Looking for more features?" Clicking this text will take you to a page on http://www.clcbio.com where you can read more about how to get more functionalities into *CLC Drug Discovery Workbench*.

1.6.2 Import of example data

It might be easier to understand the logic of the program by trying to do simple operations on existing data. Therefore *CLC Drug Discovery Workbench* includes an example data set.

When downloading *CLC Drug Discovery Workbench* you are asked if you would like to import the example data set. If you accept, the data is downloaded automatically and saved in the program. If you didn't download the data, or for some other reason need to download the data again, you have two options:

You can click **Import Example Data** ( ) in the **Help** menu of the program. This imports the data automatically. You can also go to http://www.clcbio.com/download and download the example data from there.

If you download the file from the website, you need to import it into the program. See chapter 6 for more about importing data.

1.7 Plugins

When you install *CLC Drug Discovery Workbench*, it has a standard set of features. However, you can upgrade and customize the program using a variety of plugins.

As the range of plugins is continuously updated and expanded, they will not be listed here. Instead we refer to http://www.clcbio.com/plugins for a full list of plugins with descriptions of their functionalities.

1.7.1 Installing plugins

Plugins are installed using the plugin manager. In order to install plugins on Windows, the Workbench must be run in administrator mode: Right-click the program shortcut and choose "Run as Administrator". Then follow the procedure described below.

Help in the Menu Bar | **Plugins... ( )**

or **Plugins ( ) in the Toolbar**

The plugin manager has two tabs at the top:

- **Manage Plugins.** This is an overview of plugins that are installed.
- **Download Plugins.** This is an overview of available plugins on CLC bio's server.

To install a plugin, click the **Download Plugins** tab. This will display an overview of the plugins that are available for download and installation (see figure 1.25).

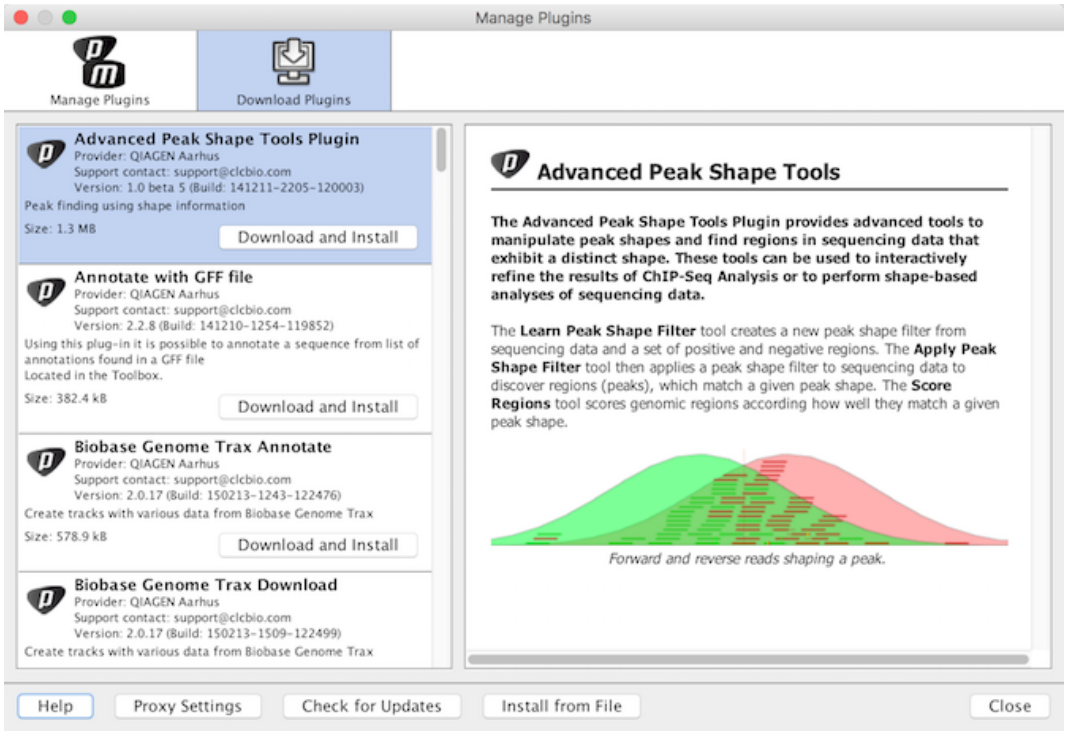


Figure 1.25: *The plugins that are available for download.*

Clicking a plugin will display additional information at the right side of the dialog. This will also display a button: **Download and Install**.

Click the plugin and press **Download and Install**. A dialog displaying progress is now shown, and the plugin is downloaded and installed.

If the plugin is not shown on the server, and you have it on your computer (for example if you have downloaded it from our website), you can install it by clicking the **Install from File** button at the bottom of the dialog. This will open a dialog where you can browse for the plugin. The plugin file should be a file of the type ".cpa".

When you close the dialog, you will be asked whether you wish to restart the *CLC Drug Discovery Workbench*. The plugin will not be ready for use until you have restarted.

1.7.2 Uninstalling plugins

Plugins are uninstalled using the plugin manager:

Help in the Menu Bar | **Plugins... (** **)**

or **Plugins (** **) in the Toolbar**

This will open the dialog shown in figure 1.26.

The installed plugins are shown in this dialog. To uninstall:

Click the plugin | **Uninstall**

If you do not wish to completely uninstall the plugin but you don't want it to be used next time you start the Workbench, click the **Disable** button.

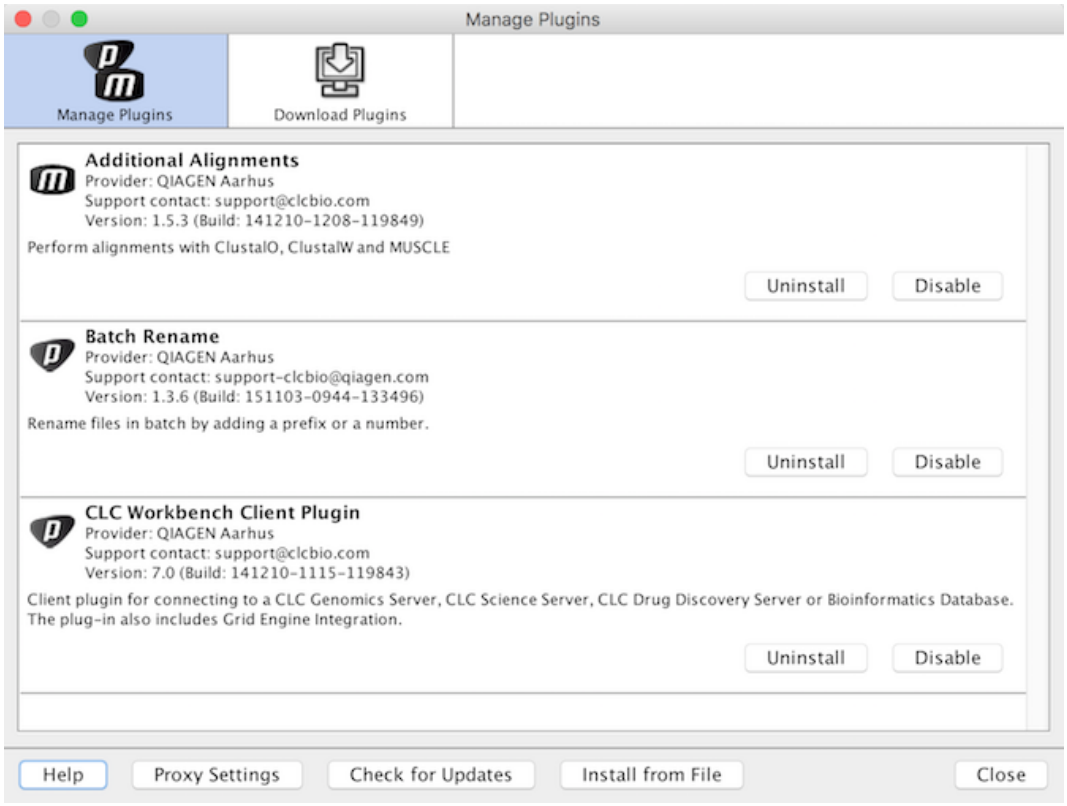


Figure 1.26: *The plugin manager with plugins installed.*

When you close the dialog, you will be asked whether you wish to restart the workbench. The plugin will not be uninstalled until the workbench is restarted.

1.7.3 Updating plugins

If a new version of a plugin is available, you will get a notification during start-up as shown in figure 1.27.

In this list, select which plugins you wish to update, and click **Install Updates**. If you press **Cancel** you will be able to install the plugins later by clicking **Check for Updates** in the Plugin manager (see figure 1.26).

1.8 Network configuration

If you use a proxy server to access the Internet you must configure *CLC Drug Discovery Workbench* to use this. Otherwise you will not be able to perform any online activities (e.g. searching for structures at NCBI).

CLC Drug Discovery Workbench supports the use of an HTTP-proxy and an anonymous SOCKS-proxy.

To configure your proxy settings, open *CLC Drug Discovery Workbench*, and go to the **Advanced**-tab of the **Preferences** dialog (figure 1.28) and enter the appropriate information. The **Preferences** dialog is opened from the **Edit** menu.

You have the choice between an HTTP-proxy and a SOCKS-proxy. *CLC Drug Discovery Workbench*

Figure 1.27: *Plugin updates.*

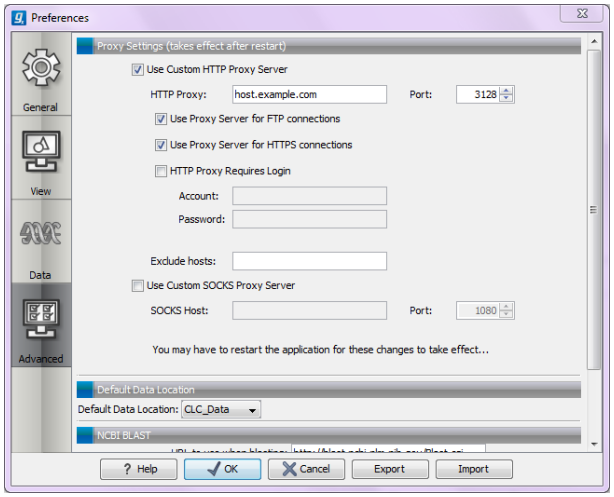


Figure 1.28: *Adjusting proxy preferences.*

only supports the use of a SOCKS-proxy that does not require authorization.

You can select whether the proxy should be used also for FTP and HTTPS connections.

Exclude hosts can be used if there are some hosts that should be contacted directly and not through the proxy server. The value can be a list of hosts, each separated by a `|`, and in addition a wildcard character `*` can be used for matching. For example: `*.foo.com|localhost`.

If you have any problems with these settings you should contact your systems administrator.

1.9 The format of the user manual

This user manual offers support to Windows, Mac OS X and Linux users. The software is very similar on these operating systems. In areas where differences exist, these will be described separately. However, the term "right-click" is used throughout the manual, but some Mac users may have to use Ctrl+click in order to perform a "right-click" (if they have a single-button mouse).

The most recent version of the user manuals can be downloaded from http://www.clcbio.com/usermanuals.

The user manual consists of four parts.

- The **first part** includes the introduction to the *CLC Drug Discovery Workbench*.
- The **second part** describes in detail how to operate all the program's basic functionalities.
- The **third part** digs deeper into some of the molecular modeling and bioinformatic features of the program. In this part, you will also find our "Bioinformatics explained" sections. These sections elaborate on the algorithms and analyses of *CLC Drug Discovery Workbench* and provide more general knowledge of molecular modeling and bioinformatic concepts.
- The **fourth part** is the Appendix and Index.

Each chapter includes a short table of contents.

1.9.1 Text formats

In order to produce a clearly laid-out content in this manual, different formats are applied:

- A feature in the program is in bold starting with capital letters. (Example: **Navigation Area**)
- An explanation of how a particular function is activated, is illustrated by "|" and bold. (E.g.: **select the element** | **Edit** | **Rename**)

Part II

Core Functionalities

Chapter 2

User interface

Contents

This chapter provides an overview of the different areas in the user interface of *CLC Drug Discovery Workbench*. As can be seen from figure 2.1 this includes a **Navigation Area**, **View Area**, **Menu Bar**, **Toolbar**, **Status Bar** and **Toolbox**.

A description of the **Navigation Area** is tightly connected to the data organization features of *CLC Drug Discovery Workbench* and can be found in section 3.1.

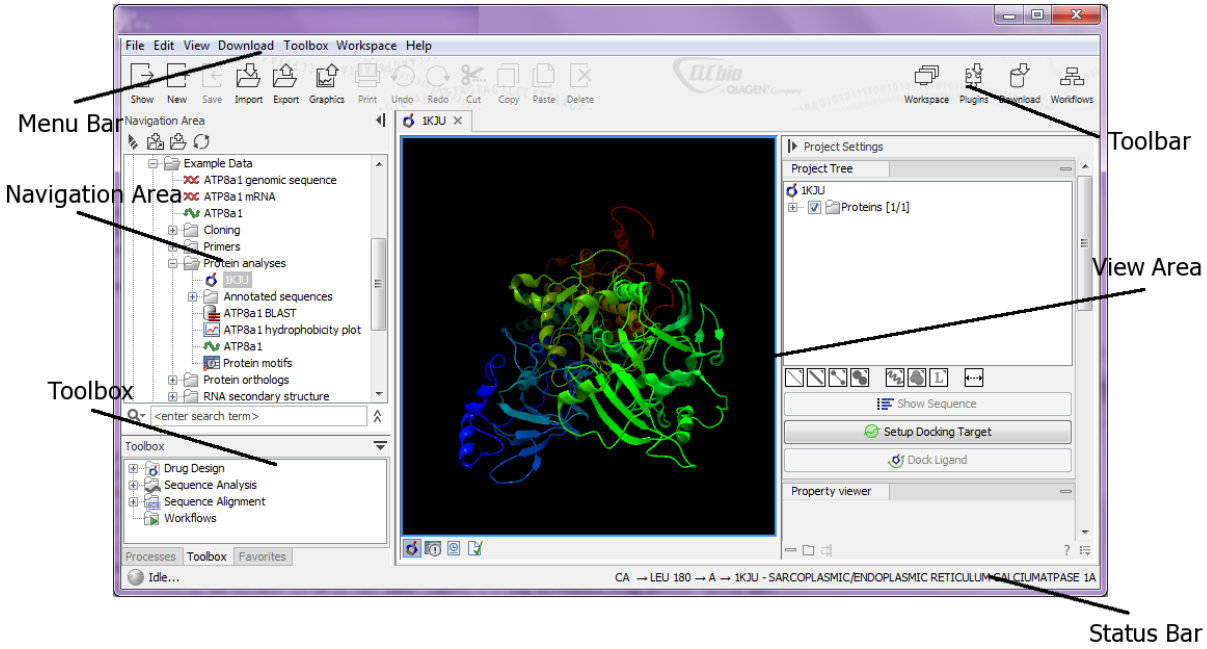


Figure 2.1: *The user interface consists of the Menu Bar, Toolbar, Status Bar, Navigation Area, Toolbox, and View Area.*

2.1 View Area

The **View Area** is the right-hand part of the screen, displaying your current work. The **View Area** may consist of one or more **Views**, represented by tabs at the top of the **View Area**.

This is illustrated in figure 2.2.

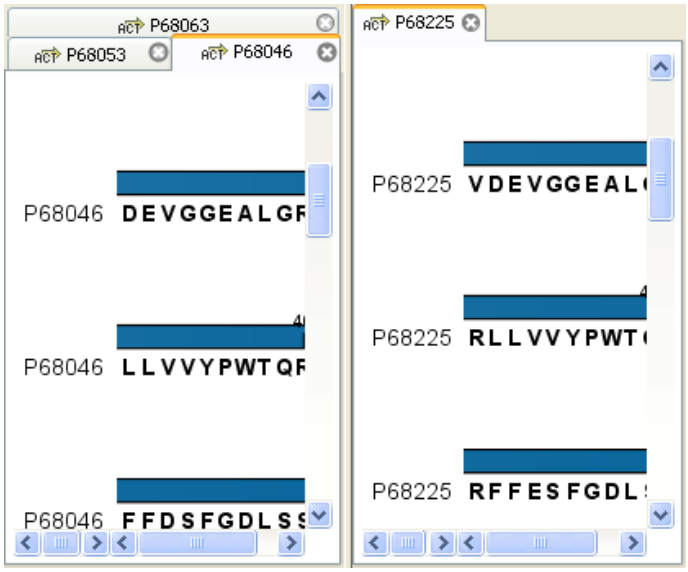


Figure 2.2: *A View Area can enclose several views, each view is indicated with a tab (see right view, which shows protein P68225). Furthermore, several views can be shown at the same time (in this example, four views are displayed).*

The tab concept is central to working with *CLC Drug Discovery Workbench*, because several operations can be performed by dragging the tab of a view, and extended right-click menus can

be activated from the tabs.

This chapter deals with the handling of views inside a **View Area**. Furthermore, it deals with rearranging the views.

Section 2.2 deals with the zooming and selecting functions.

2.1.1 Open view

Opening a view can be done in a number of ways:

double-click an element in the Navigation Area

or **select an element in the Navigation Area** | **File** | **Show** | **Select the desired way to view the element**

or **select an element in the Navigation Area** | **Ctrl + O (⌘ + B on Mac)**

Opening a view while another view is already open, will show the new view in front of the other view. The view that was already open can be brought to front by clicking its tab.

Note! If you right-click an open tab of any element, click **Show**, and then choose a different view of the same element, this new view is automatically opened in a split-view, allowing you to see both views.

See section 3.1.5 for instructions on how to open a view using drag and drop.

2.1.2 Open additional views with the toolbar

Each element can be shown in different ways. A sequence, for example, can be shown as linear, circular, text etc.

In the following example, you want to see a sequence in a circular view. If the sequence is already open in a view, you can change the view to a circular view:

Click Show As Circular ( ) at the lower left part of the view

The buttons used for switching views are shown in figure 2.3).

Figure 2.3: *The buttons shown at the bottom of a view of a nucleotide sequence. You can click the buttons to change the view to e.g. a circular view or a history view.*

If the sequence is already open in a linear view ( ), and you wish to see both a circular and a linear view, you can split the views very easily:

Press Ctrl (⌘ on Mac) while you | **Click Show As Circular ( ) at the lower left part of the view**

This will open a split view with a linear view at the bottom and a circular view at the top (see 10.5).

You can also show a circular view of a sequence without opening the sequence first:

Select the sequence in the Navigation Area | **Show ( )** | **As Circular ( )**

History and Info views The two buttons to the right hand side of the toolbar are **Show History ( )** and **Show Element Info ( )**.

The History view is a textual log of all operations you make in the program. If for example you rename a sequence, align sequences, create a phylogenetic tree or translate a sequence, you can always go back and check what you have done. In this way, you are able to document and reproduce previous operations.

When an element's history is opened, the newest change is submitted in the top of the view (figure 2.4).

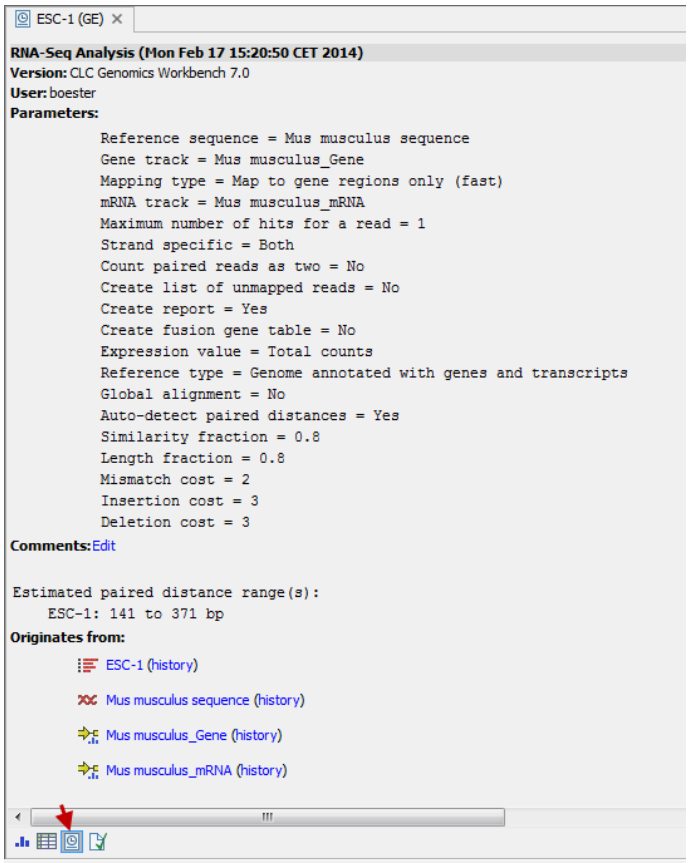


Figure 2.4: *An element's history.*

The following information is available:

- **Title**. The action that the user performed.
- **Date and time**. Date and time for the operation. The date and time are displayed according to your locale settings (see section 4.1).
- **Version**. The Workbench type and version that has been used.
- **User**. The user who performed the operation. If you import some data created by another person in a CLC Workbench, that person's name will be shown.
- **Parameters**. Details about the action performed. This could be the parameters that were chosen for an analysis.

- **Comments**. By clicking **Edit** you can enter your own comments regarding this entry in the history. These comments are saved.
- **Originates from**. This information is usually shown at the bottom of an element's history. Here, you can see which elements the current element originates from. For example, if you have created an alignment of three sequences, the three sequences are shown here. Clicking the element selects it in the **Navigation Area**, and clicking the 'history' link opens the element's own history.

When an element's info is open you can check current information about the element, and in particular the potential association of the data you are looking at with metadata. To learn more about the **Show Element Info** button, see section 10.4 and see section 3.2.3.

2.1.3 Close views

When a view is closed, the **View Area** remains open as long as there is at least one open view.

A view is closed by:

right-click the tab of the View | Close

or **select the view | Ctrl + W**

or **hold down the Ctrl-button | Click the tab of the view while the button is pressed**

By right-clicking a tab, the following close options exist (figure 2.5).

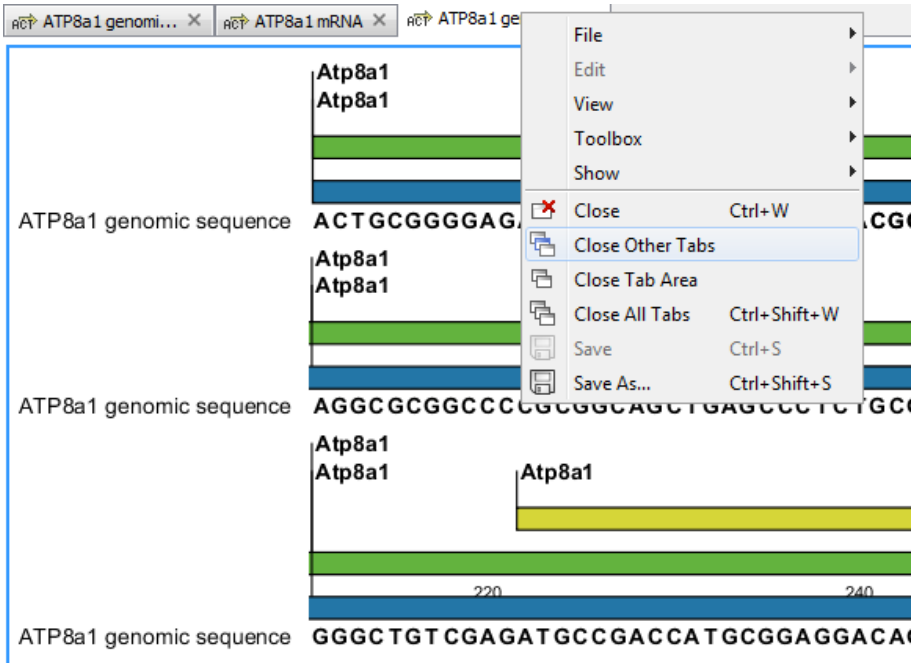


Figure 2.5: *By right-clicking a tab, several close options are available.*

- **Close.** See above.
- **Close Other Tabs.** Closes all other tabs, in all tab areas, except the one that is selected.

- **Close Tab Area.** Closes all tabs in the tab area.
- **Close All Tabs.** Closes all tabs, in all tab areas. Leaves an empty workspace.

2.1.4 Save changes in a view

When changes to an element are made in a view, the text on the tab appears *bold and italic* (on Mac it is indicated by an * before the name of the tab). This indicates that the changes are not saved. The **Save** function may be activated in two ways:

Click the tab of the view you want to save | **Save (** **) in the toolbar.**

or **Click the tab of the view you want to save** | **Ctrl + S (⌘ + S on Mac)**

If you close a tab of a view containing an element that has been changed since you opened it, you are asked if you want to save.

When saving an element from a new view that has not been opened from the **Navigation Area** (e.g. when opening a sequence from a list of search hits), a save dialog appears (figure 2.6).

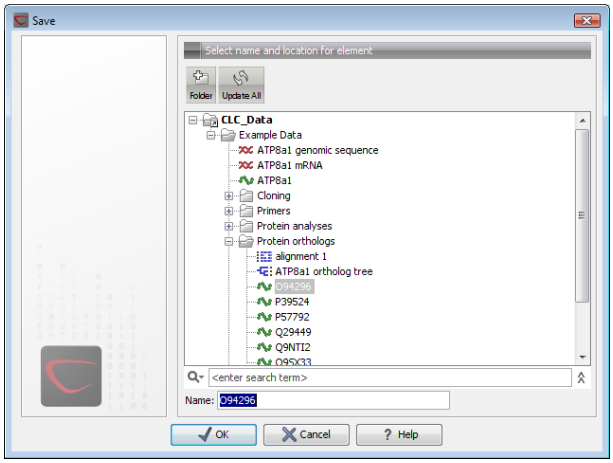


Figure 2.6: *Save dialog.*

In the dialog you select the folder in which you want to save the element.

After naming the element, press **OK**

2.1.5 Undo/Redo

If you make a change to an element in a view, e.g. remove an annotation in a sequence or modify a tree, you can undo the action. In general, **Undo** applies to all changes you can make when right-clicking in a view. **Undo** is done by:

Click undo (**) in the Toolbar**

or **Edit** | **Undo (** **)**

or **Ctrl + Z**

If you want to undo several actions, just repeat the steps above. To reverse the undo action:

Click the redo icon in the Toolbar

or **Edit** | **Redo (**)

or **Ctrl + Y**

Note! Actions in the **Navigation Area**, e.g. renaming and moving elements, cannot be undone. However, you can restore deleted elements (see section 3.1.7).

You can set the number of possible undo actions in the Preferences dialog (see section 4).

2.1.6 Arrange views in View Area

To provide more space for viewing data, you can hide **Navigation Area** and the **Toolbox** by clicking the hide icon ( ) at the top of the **Navigation Area**.

Views are arranged in the **View Area** by their tabs. The order of the **views** can be changed using drag and drop. E.g. drag the tab of one view onto the tab of a another. The tab of the first view is now placed at the right side of the other tab.

If a tab is dragged into a view, an area of the view is made gray (see fig. 2.7) illustrating that the view will be placed in this part of the **View Area**.

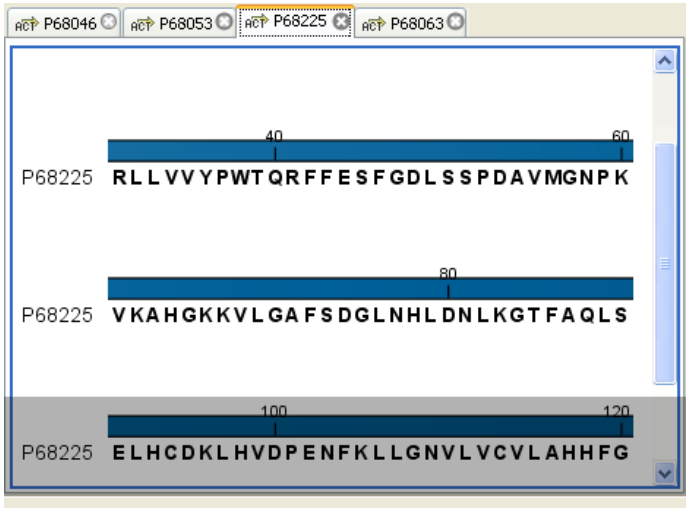


Figure 2.7: *When dragging a view, a gray area indicates where the view will be shown.*

The results of this action is illustrated in figure 2.8.

You can also split a **View Area** horizontally or vertically using the menus.

Splitting horizontally may be done this way:

right-click a tab of the view | **View** | **Split Horizontally (**)

This action opens the chosen view below the existing view. (See figure 2.9). When the split is made vertically, the new view opens to the right of the existing view.

Splitting the **View Area** can be undone by dragging e.g. the tab of the bottom view to the tab of the top view. This is marked by a gray area on the top of the view.

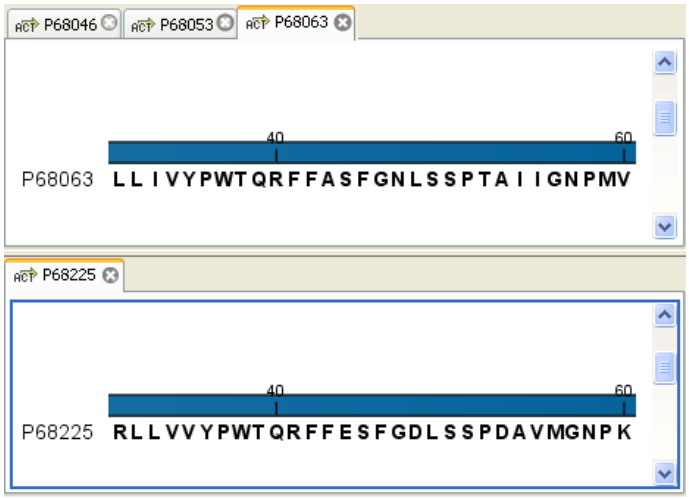


Figure 2.8: *A horizontal split-screen. The two views split the View Area.*

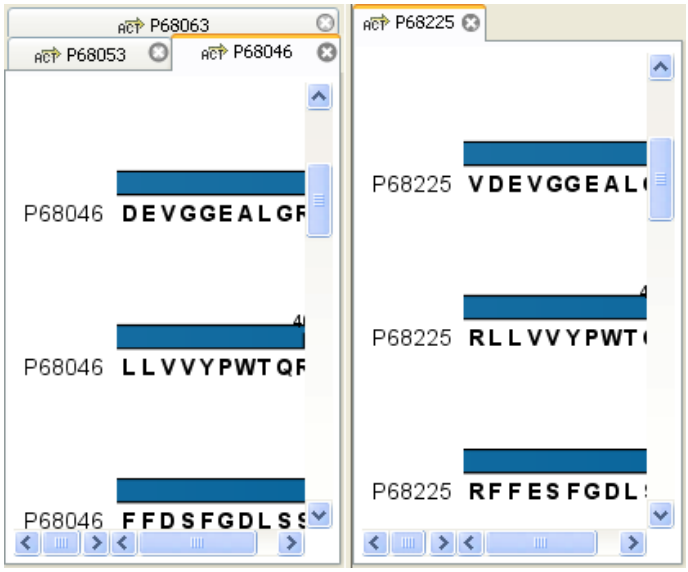


Figure 2.9: *A vertical split-screen.*

Maximize/Restore size of view

The **Maximize/Restore View** function allows you to see a view in maximized mode, meaning a mode where no other **views** nor the **Navigation Area** is shown.

Maximizing a view can be done in the following ways:

select view | **Ctrl + M**

or **select view** | **View** | **Maximize/restore View (▭)**

or **select view** | **right-click the tab** | **View** | **Maximize/restore View (▭)**

or **double-click the tab of view**

The following restores the size of the view:

Ctrl + M

or **View** | **Maximize/restore View (▭)**

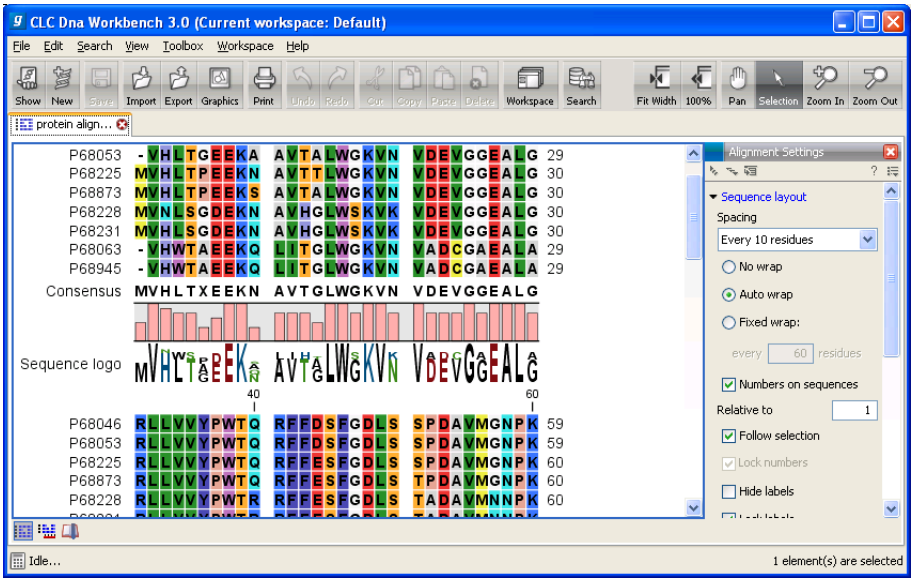


Figure 2.10: *A maximized view. The function hides the Navigation Area and the Toolbox.*

or **double-click title of view**

Please note that you can also hide **Navigation Area** and the **Toolbox** by clicking the hide icon ( ) at the top of the **Navigation Area**

2.1.7 Moving a view to a different screen

Using multiple screens can be a great benefit when analyzing data with the *CLC Drug Discovery Workbench*. You can move a view to another screen by dragging the tab of the view and dropping it outside the workbench window. Alternatively, you can right-click in the view area or on the tab itself and select **View | Move to New Window** from the context menu.

An example is shown in figure 2.11, where the main Workbench window shows a table of open reading frames, and the screen to the right is used to display the sequence and annotations.

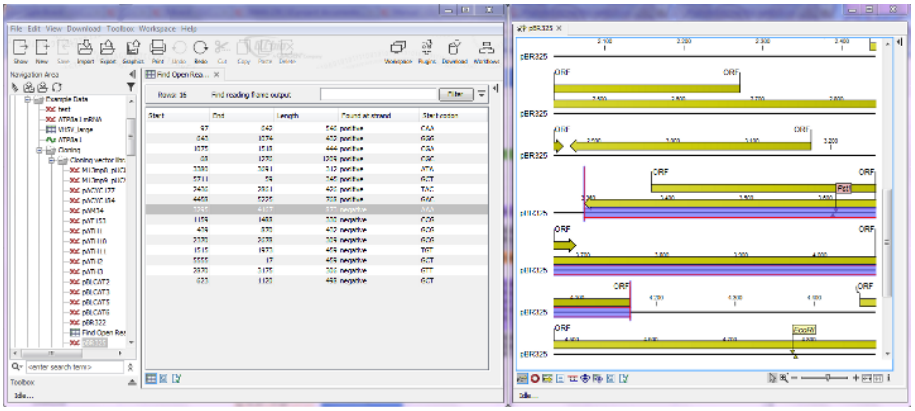


Figure 2.11: *Showing the table on one screen while the sequence is displayed on another screen. Clicking the table of open reading frames causes the view on the other screen to follow the selection. Note that the screen resolution in this figure is kept low in order to include it in the manual; in a real scenario, the resolution will be much higher.*

You can make more detached windows, by dropping tabs outside the open workbench windows,

or you can drag more tabs to a detached window. To get a tab back to the main workbench window, just drag the detached tab back, and drop it next to the other tabs in the top of the view area. **Note:** You should not drag the detached window header, just the tab itself.

You can also split the view area in the detached windows as described in section 2.1.6.

2.1.8 Side Panel

The **Side Panel** allows you to change the way the contents of a view are displayed. The options in the **Side Panel** depend on the kind of data in the view, and they are described in the relevant sections about sequences, alignments, trees etc.

Figure 2.12 shows the default **Side Panel** for a protein sequence. It is organized into *palettes*.

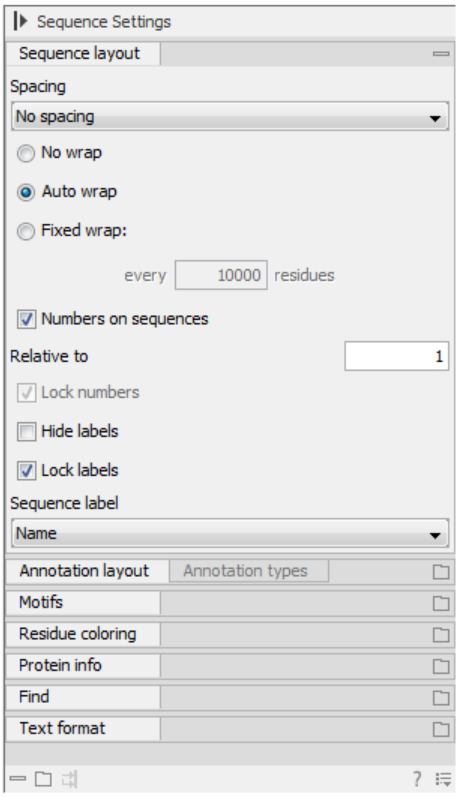


Figure 2.12: *The default view of the Side Panel when opening a protein sequence.*

In this example, there is one for Sequence layout, one for Annotation Layout etc. These palettes can be re-organized by dragging the palette name with the mouse and dropping it where you want it to be. They can either be situated next to each other, so that you can switch between them, or they can be listed on top of each other, so that expanding one of the palettes will push the palettes below further down.

In addition, they can be moved away from the **Side Panel** and placed anywhere on the screen as shown in figure2.13.

In this example, the **Motifs** palette has been placed on top of the sequence view together with the **Protein info** and the **Residue coloring** palettes. In the **Side Panel** to the right, the **Find** palette has been put on top.

In order to make all palettes dock in the **Side Panel** again, click the **Dock Side Panel** icon ( ).

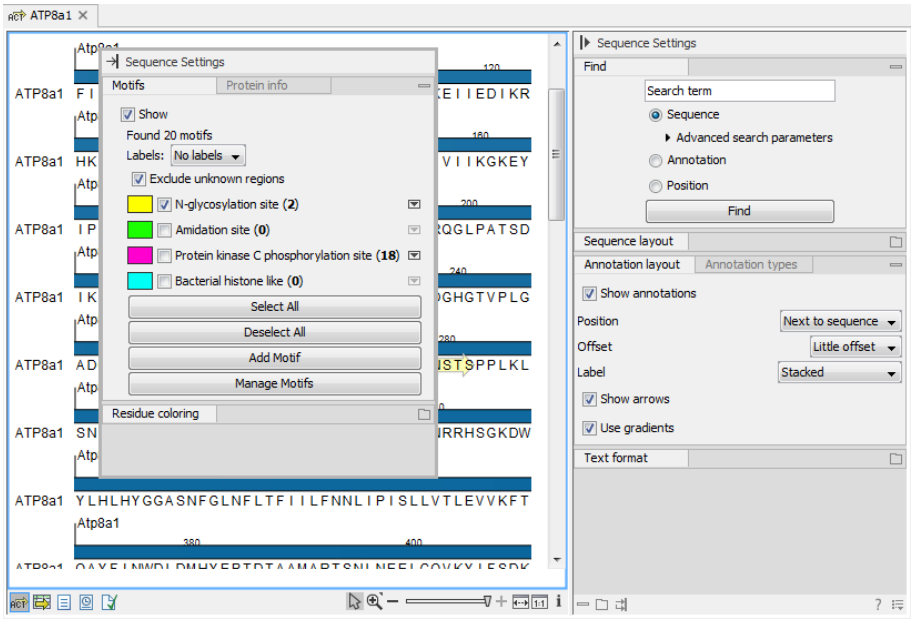


Figure 2.13: *Palettes can be organized in the Side Panel as you like or placed anywhere on the screen.*

You can completely hide the **Side Panel** by clicking the **Hide Side Panel** icon ( ).

At the bottom of the **Side Panel** (see figure 2.14) there are a number of icons used to:

- Expand all settings ( ).
- Collapse all settings ( ).
- Dock all palettes ( )
- Get **Help** for the particular view and settings ( )
- Save the settings of the **Side Panel** or apply already saved settings. Read more in section 4.5

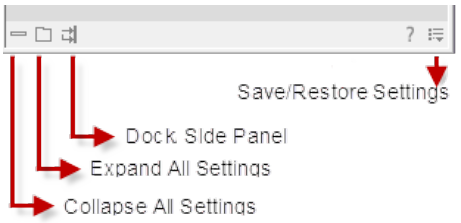


Figure 2.14: *Controlling the Side Panel at the bottom*

Note! Changes made to the **Side Panel**, including the organization of palettes will not be saved when you save the view. See how to save the changes in section 4.5

2.2 Zoom and selection in View Area

All views except tabular and text views support zooming. Figure 2.15 shows the zoom tools, located at the bottom right corner of the view.

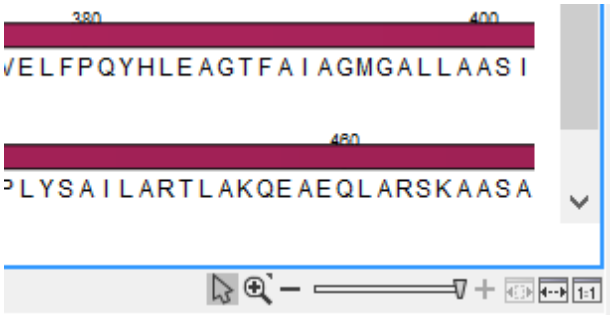


Figure 2.15: *The zoom tools are located at the bottom right corner of the view.*

The zoom tools consist of some shortcuts for zooming to fit the width of the view (), zoom to 100 % to see details (), zoom to a selection (), a zoom slider, and two mouse mode buttons () ().

The slider reflects the current zoom level and can be used to quickly adjust this. For more fine-grained control of the zoom level, move the mouse upwards while sliding.

The sections below describes how to use these tools as well as other ways of zooming and navigating data.

Please note that when working with protein 3D structures, there are specific ways of controlling zooming and navigation as explained in section 9.1.

2.2.1 Zoom in

There are six ways of **zooming in**:

Click Zoom in mode () in the zoom tools (or press Ctrl+2) | click the location in the view that you want to zoom in on

or **Click Zoom in mode () in the zoom tools | click-and-drag a box around a part of the view | the view now zooms in on the part you selected**

or **Press '+' on your keyboard**

or **Move the zoom slider located in the zoom tools**

or **Click the plus icon in the zoom tools**

The last option for zooming in is only available if you have a mouse with a scroll wheel:

or **Press and hold Ctrl (⌘ on Mac) | Move the scroll wheel on your mouse forward**

Note! You might have to click in the view before you can use the keyboard or the scroll wheel to zoom.

If you press the **Shift** button on your keyboard while in zoom mode, the zoom function is reversed.

If you want to zoom in to 100 % to see the data at base level, click the **Zoom to base level** () icon.

2.2.2 Zoom out

It is possible to zoom out in different ways:

Click Zoom out mode ( ) in the zoom tools (or press Ctrl+3) | click in the view

or **Press '-' on your keyboard**

or **Move the zoom slider located in the zoom tools**

or **Click the minus icon in the zoom tools**

The last option for zooming out is only available if you have a mouse with a scroll wheel:

or **Press and hold Ctrl (⌘ on Mac)** | **Move the scroll wheel on your mouse backwards**

Note! You might have to click in the view before you can use the keyboard or the scroll wheel to zoom.

If you want to zoom out to see all the data, click the **Zoom to Fit** ( ) icon.

If you press **Shift** while clicking in a **View**, the zoom function is reversed. Hence, clicking on a sequence in this way while the **Zoom out** mode toolbar item is selected, zooms in instead of zooming out.

2.2.3 Selecting, panning and zooming

In the zoom tools, you can control which mouse mode to use. The default is **Selection mode** ( ) which is used for selecting data in a view. Next to the selection mode, you can select the **Zoom in mode** as described in section 2.2.1. If you press and hold this button, two other modes become available as shown in figure 2.16:

- **Panning** ( ) is used for dragging the view with the mouse as a way of scrolling.
- **Zoom out** ( ) is used to change the mouse mode so that whenever you click the view, it zooms out.

Figure 2.16: *Additional mouse modes can be found in the zoom tools.*

If you hold the mouse over the selection and zoom tools, tooltips will appear that provide further information about how to use the tools.

The mouse modes only apply when the mouse is within the view where they are selected.

The **Selection mode** can also be invoked with the keyboard shortcut Ctrl+1, while the **Panning mode** can be invoked with Ctrl+4.

For some views, if you have made a selection, there is a **Zoom to Selection** ( ) button, which allows you to zoom and scroll directly to fit the view to the selection.

2.3 Toolbox and Status Bar

The **Toolbox** is placed in the left side of the user interface of *CLC Drug Discovery Workbench* below the **Navigation Area**.

The **Toolbox** shows a **Processes tab**, **Favorites tab** and a **Toolbox tab**.

The **Toolbox** can be hidden, so that the **Navigation Area** is enlarged and thereby displays more elements:

View | **Show/Hide Toolbox** | **Show/Hide Toolbox**

You can also click the **Hide Toolbox** ( ) button.

2.3.1 Processes

By clicking the **Processes** tab, the **Toolbox** displays previous and running processes, e.g. an NCBI search or a calculation of an alignment. The running processes can be stopped, paused, and resumed by clicking the small icon ( ) next to the process (see figure 2.17).

Running and paused processes are not deleted.

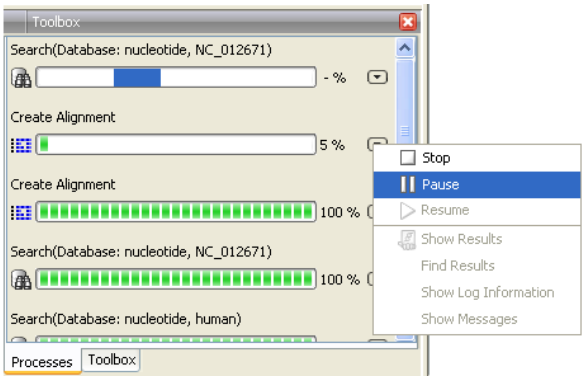


Figure 2.17: *A database search and an alignment calculation are running. Clicking the small icon next to the process allow you to stop, pause and resume processes.*

Besides the options to stop, pause and resume processes, there are some extra options for *a selected number* of the tools running from the Toolbox:

- **Show results**. If you have chosen to save the results (see section 7.2), you will be able to open the results directly from the process by clicking this option.
- **Find results**. If you have chosen to save the results (see section 7.2), you will be able to highlight the results in the Navigation Area.
- **Show Log Information**. This will display a log file showing progress of the process. The log file can also be shown by clicking **Show Log** in the "handle results" dialog where you choose between saving and opening the results.
- **Show Messages**. Some analyses will give you a message when processing your data. The messages are the black dialogs shown in the lower left corner of the Workbench that disappear after a few seconds. You can reiterate the messages that have been shown by clicking this option.

The terminated processes can be removed by:

View | **Remove Finished Processes (✕)**

If you close the program while there are running processes, a dialog will ask if you are sure that you want to close the program. Closing the program will stop the process, and it cannot be restarted when you open the program again.

2.3.2 Toolbox

The content of the **Toolbox** tab in the **Toolbox** corresponds to **Toolbox** in the **Menu Bar**.

The tools in the toolbox can be accessed by double-clicking or by dragging elements from the **Navigation Area** to an item in the **Toolbox**.

Quick access to tools

To enable quick launch of tools in *CLC Drug Discovery Workbench*, press Ctrl + Shift + T (⌘ + Shift + T on Mac) to show the quick launch dialog (see figure 2.18).

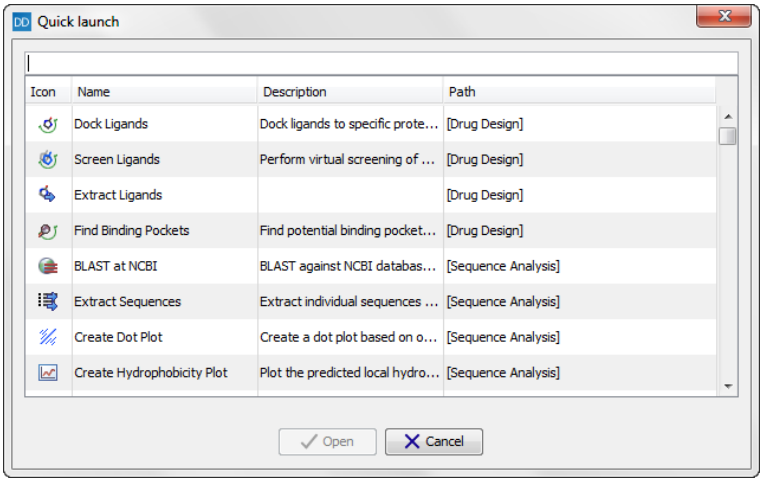


Figure 2.18: *Quick access to all tools in* **CLC Drug Discovery Workbench**.

When the dialog is opened, you can start typing search text in the text field at the top. This will bring up the list of tools that match this text either in the name, description or location in the Toolbox. In the example shown in figure 2.19, typing `create` shows a list of tools involving the word "create", and the arrow keys or mouse can be used for selecting and starting a tool.

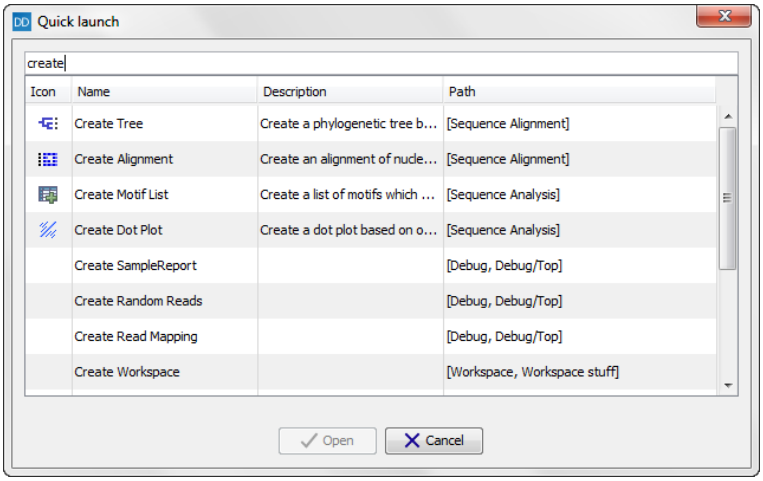


Figure 2.19: *Typing in the search field at the top will filter the list of tools to launch.*

Favorites toolbox

Next to the **Toolbox** tab, you find the **Favorites** tab. This can be used for organizing and getting quick access to the tools you use the most. It consists of two parts as shown in figure 2.20.

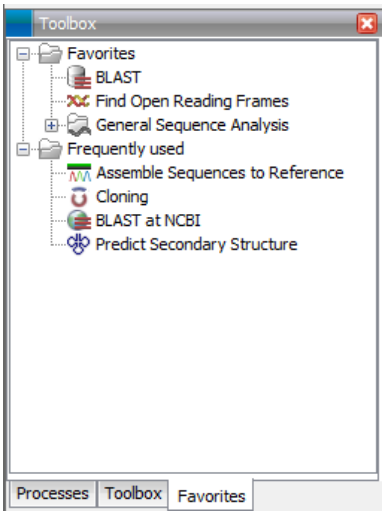


Figure 2.20: *Favorites toolbox.*

Favorites You can manually add tools to the favorites menu simply by right-clicking the tool in the Toolbox. You can also right-click the Favorites folder itself and select **Add Tool**. To remove a tool, right-click and select **Remove from Favorites**. Note that you can also add complete folders to the favorites.

Frequently used The list of tools in this folder is automatically populated as you use the Workbench. The most frequently used tools are listed at the top.

2.3.3 Status Bar

As can be seen from figure 2.1, the **Status Bar** is located at the bottom of the window. In the left side of the bar is an indication of whether the computer is making calculations or whether it is idle. The right side of the **Status Bar** indicates the range of the selection of a sequence. (See chapter 2.2.3 for more about the Selection mode button.)

2.4 Workspace

If you are working on a project and have arranged the views for this project, you can save this arrangement using **Workspaces**. A Workspace remembers the way you have arranged the views, and you can switch between different workspaces.

The **Navigation Area** always contains the same data across **Workspaces**. It is, however, possible to open different folders in the different **Workspaces**. Consequently, the program allows you to display different clusters of the data in separate **Workspaces**.

All **Workspaces** are automatically saved when closing down *CLC Drug Discovery Workbench*. The next time you run the program, the **Workspaces** are reopened exactly as you left them.

Note! It is not possible to run more than one version of *CLC Drug Discovery Workbench* at a time. Use two or more **Workspaces** instead.

2.4.1 Create Workspace

When working with large amounts of data, it might be a good idea to split the work into two or more **Workspaces**. As default the *CLC Drug Discovery Workbench* opens one **Workspace**. Additional **Workspaces** are created in the following way:

Workspace in the Menu Bar) | Create Workspace | enter name of Workspace | OK

When the new **Workspace** is created, the heading of the program frame displays the name of the new **Workspace**. Initially, the selected elements in the **Navigation Area** is collapsed and the **View Area** is empty and ready to work with. (See figure 2.21).

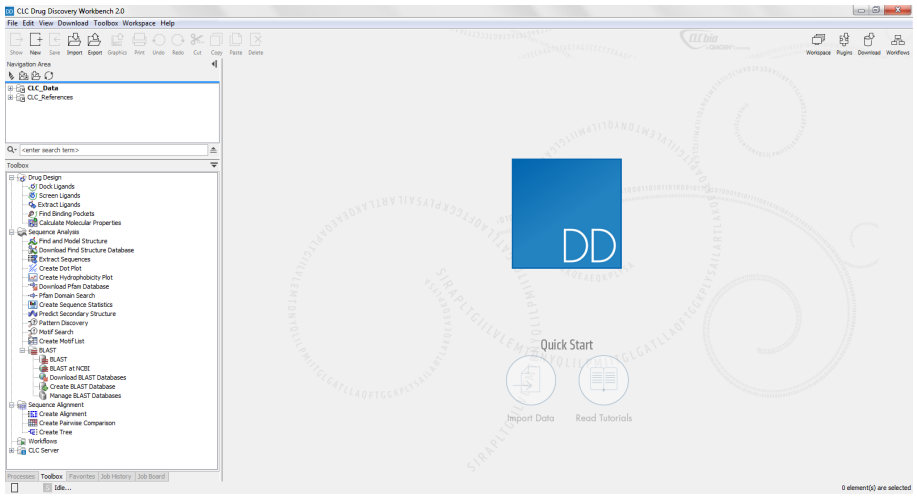


Figure 2.21: *An empty Workspace.*

2.4.2 Select Workspace

When there is more than one **Workspace** in the *CLC Drug Discovery Workbench*, there are two ways to switch between them:

Workspace ( ) in the Toolbar | Select the Workspace to activate

or **Workspace in the Menu Bar | Select Workspace ( ) | choose which Workspace to activate | OK**

The name of the selected **Workspace** is shown after "*CLC Drug Discovery Workbench*" at the top left corner of the main window, in figure 2.21 it says: (default).

2.4.3 Delete Workspace

Deleting a **Workspace** can be done in the following way:

Workspace in the Menu Bar | Delete Workspace | choose which Workspace to delete | OK

Note! Be careful to select the right **Workspace** when deleting. The delete action cannot be undone. (However, no data is lost, because a workspace is only a representation of data.)

It is not possible to delete the default workspace.

Chapter 3

Data management and search

Contents

This chapter explains the data management features of *CLC Drug Discovery Workbench*. The first section explains the basics of the data organization and the **Navigation Area**. The next section explains how to set up custom attributes for the data that can be used for more advanced data management. Finally, there is a section about how to search through local data.

3.1 Navigation Area

The **Navigation Area** is located in the left side of the screen, under the **Toolbar** (see figure 3.1). It is used for organizing and navigating data. Its behavior is similar to the way files and folders are usually displayed on your computer.

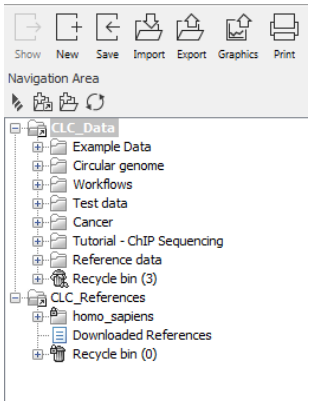


Figure 3.1: *The Navigation Area.*

To provide more space for viewing data, you can hide **Navigation Area** and the **Toolbox** by clicking the hide icon ( ) at the top.

3.1.1 Data structure

The data in the **Navigation Area** is organized into a number of **Locations**. When the *CLC Drug Discovery Workbench* is started for the first time, there is one location called *CLC_Data* (unless your computer administrator has configured the installation otherwise).

A location represents a folder on the computer: The data shown under a location in the **Navigation Area** is stored on the computer in the folder which the location points to.

This is explained visually in figure 3.2. The full path to the system folder can be located by mousing over the data location as shown in figure 3.3.

Adding locations

Per default, there is one location in the **Navigation Area** called CLC_Data. It points to the following folder:

- On Windows: C:\Users\<your_username>\CLC_Data
- On Mac: ~/CLC_Data
- On Linux: /homefolder/CLC_Data

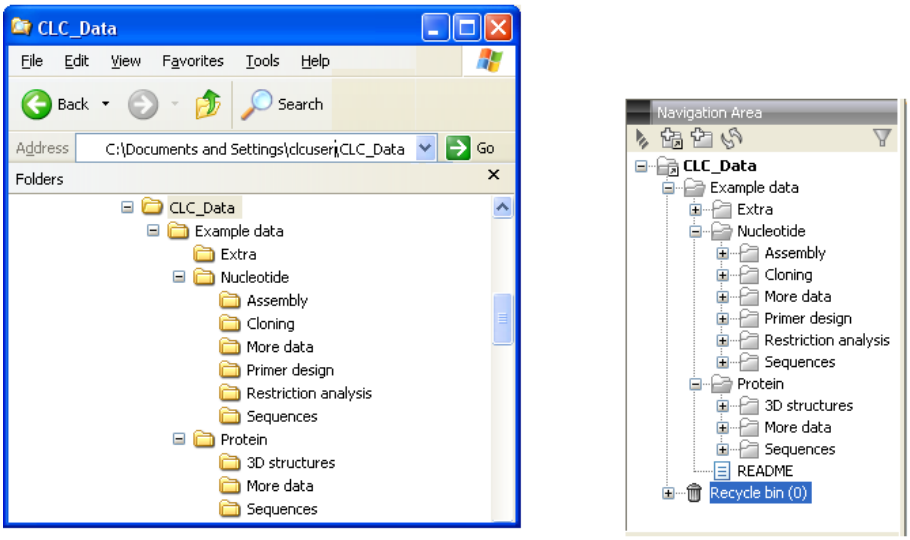


Figure 3.2: *In this example the location called 'CLC_Data' points to the folder at C:\Documents and settings\clcuser\CLC_Data.*

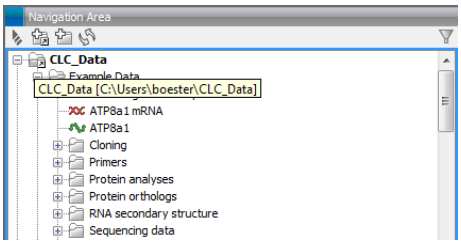


Figure 3.3: *Mousing over the location called 'CLC_Data' shows the full path to the system folder, which in this case is C:\Users\boester\CLC_Data.*

You can easily add more locations to the **Navigation Area**:

File | **New** | **Location (** **)**

This will bring up a dialog where you can navigate to the folder you wish to use as your new location (see figure 3.4).

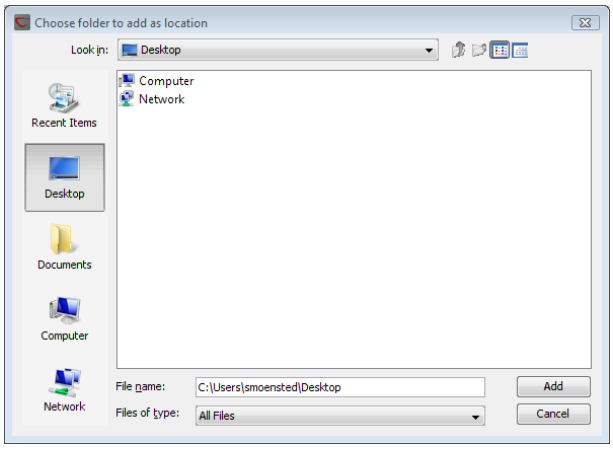


Figure 3.4: *Navigating to a folder to use as a new location.*

When you click **Open**, the new location is added to the **Navigation Area** as shown in figure 3.5.

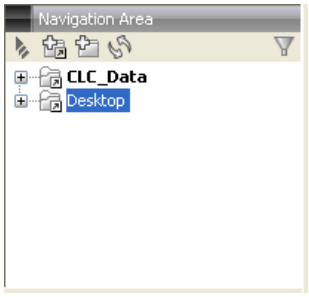


Figure 3.5: *The new location has been added.*

The name of the new location will be the name of the folder selected for the location. To see where the folder is located on your computer, place your mouse cursor on the location icon ( ) for a second. This will show the path to the location.

You can use a folder on a network drive or a removable drive as a Data Location. Such a location will appear inactive if the relevant drive is not available when you start up the Workbench. Once the drive is available, click on Update All (Image updates) and the relevant Data Location will become active (note that there might be a few seconds delay from the moment you connect).

Sharing data is possible when a network drive is available to multiple Workbenches. In this case, you can add the same folder as a Data Location on each Workbench. However, it is important to note that data sharing is not actively supported: we do not support concurrent alteration of data and while the software will often detect this situation and handle it appropriately, by for example only allowing read access to all but the one party editing the file, we do not guarantee this. In addition, any functionality that involves using the data search indices, (e.g. search functionality, associating metadata with data), will not work properly for shared data locations. Re-indexing a Data Location can help in the short term, but as soon as a new file is created by another piece of software, the index will be out of date. If you decide to share data via Workbenches this way, it is vital that any Workbench that adds a Data Location already used by other Workbenches uses as a Data Location **the exact same folder from the network drive file system hierarchy as the other Workbenches have used**. Indicating a folder higher up or lower down in the hierarchy will cause problems with the indexing of the files, meaning that newly created objects by Workbench A will not be found by Workbench B and vice versa.

Opening data

The elements in the **Navigation Area** are opened by:

Double-clicking on the element

or **Clicking once on the element** | **Show ( ) in the Toolbar**

or **Clicking once on the element** | **Right-click on the element** | **Show ( )**

or **Clicking once on the element** | **Right-click on the element** | **Show (the one without an icon)** | **Select the desired way to view the element from the menu that appears when mousing over "Show"**

This will open a view in the **View Area**, which is described in section 2.1.

Adding data

Data can be added to the **Navigation Area** in a number of ways. Files can be imported from the file system (see chapter 6). Furthermore, an element can be added by dragging it into the **Navigation Area**. This could be views that are open, elements on lists, e.g. search hits or sequence lists, and files located on your computer. Finally, you can add data by adding a new location (see section 3.1.1).

If a file or another element is dropped on a folder, it is placed at the bottom of the folder. If it is dropped on another element, it will be placed just below that element.

If the element already exists in the **Navigation Area** a copy will be created with the name extension "-1", "-2" etc. if more than one copy exist.

3.1.2 Create new folders

In order to organize your files, they can be placed in folders. Creating a new folder can be done in two ways:

right-click an element in the Navigation Area | **New** | **Folder (**  **)**

or **File** | **New** | **Folder (**  **)**

If a folder is selected in the **Navigation Area** when adding a new folder, the new folder is added at the bottom of this folder. If an element is selected, the new folder is added right above that element.

You can move the folder manually by selecting it and dragging it to the desired destination.

3.1.3 Sorting folders

You can sort the elements in a folder alphabetically:

right-click the folder | **Sort Folder**

On Windows, subfolders will be placed at the top of the folder, and the rest of the elements will be listed below in alphabetical order. On Mac, both subfolders and other elements are listed together in alphabetical order.

3.1.4 Multiselecting elements

Multiselecting elements means that you select more than one element at the same time. This can be done in the following ways:

- Holding down the <Ctrl> key (⌘ on Mac) while clicking on multiple elements selects the elements that have been clicked.
- Selecting one element, and selecting another element while holding down the <Shift> key selects all the elements listed between the two locations (the two end locations included).
- Selecting one element, and moving the curser with the arrow-keys while holding down the <Shift> key, enables you to increase the number of elements selected.

3.1.5 Moving and copying elements

Elements can be moved and copied in several ways:

- Using **Copy** (), **Cut** () and **Paste** () from the **Edit** menu.
- Using Ctrl + C (⌘ + C on Mac), Ctrl + X (⌘ + X on Mac) and Ctrl + V (⌘ + V on Mac).
- Using **Copy** (), **Cut** () and **Paste** () in the **Toolbar**.
- Using drag and drop to move elements.
- Using drag and drop while pressing Ctrl / Command to copy elements.

In the following, all of these possibilities for moving and copying elements are described in further detail.

Copy, cut and paste functions

Copies of elements and folders can be made with the copy/paste function which can be applied in a number of ways:

select the files to copy | **right-click one of the selected files** | **Copy ()** | **right-click the location to insert files into** | **Paste ()**

or **select the files to copy** | **Ctrl + C (⌘ + C on Mac)** | **select where to insert files** | **Ctrl + P (⌘ + P on Mac)**

or **select the files to copy** | **Edit in the Menu Bar** | **Copy ()** | **select where to insert files** | **Edit in the Menu Bar** | **Paste ()**

If there is already an element of that name, the pasted element will be renamed by appending a number at the end of the name.

Elements can also be moved instead of copied. This is done with the cut/paste function:

select the files to cut | **right-click one of the selected files** | **Cut** () | **right-click the location to insert files into** | **Paste ()**

or **select the files to cut** | **Ctrl + X (⌘ + X on Mac)** | **select where to insert files** | **Ctrl + V (⌘ + V on Mac)**

When you have cut the element, it is "grayed out" until you activate the paste function. If you change your mind, you can revert the cut command by copying another element.

Note that if you move data between locations, the original data is kept. This means that you are essentially doing a copy instead of a move operation.

Move using drag and drop

Using drag and drop in the **Navigation Area**, as well as in general, is a four-step process:

click the element | **click on the element again, and hold left mouse button** | **drag the element to the desired location** | **let go of mouse button**

This allows you to:

- Move elements between different folders in the **Navigation Area**
- Drag from the **Navigation Area** to the **View Area**: A new view is opened in an existing **View Area** if the element is dragged from the **Navigation Area** and dropped next to the tab(s) in that **View Area**.
- Drag from the **View Area** to the **Navigation Area**: The element, e.g. a sequence, alignment, search report etc. is saved where it is dropped. If the element already exists, you are asked whether you want to save a copy. You drag from the **View Area** by dragging the tab of the desired element.

Use of drag and drop is supported throughout the program, also to open and re-arrange views (see section 2.1.6).

Note that if you move data between locations, the original data is kept. This means that you are essentially doing a copy instead of a move operation.

Copy using drag and drop

To copy instead of move using drag and drop, hold the Ctrl (⌘ on Mac) key while dragging:

> **click the element | click on the element again, and hold left mouse button | drag the element to the desired location | press Ctrl (⌘ on Mac) while you let go of mouse button release the Ctrl/⌘ button**

3.1.6 Change element names

This section describes two ways of changing the names of sequences in the **Navigation Area**. In the first part, the sequences themselves are not changed - it's their representation that changes. The second part describes how to change the name of the element.

Change how sequences are displayed

Sequence elements can be displayed in the **Navigation Area** with different types of information:

- Name (this is the default information to be shown).
- Accession (sequences downloaded from databases like GenBank have an accession number).
- Latin name.
- Latin name (accession).
- Common name.
- Common name (accession).

Whether sequences can be displayed with this information depends on their origin. Sequences that you have created yourself or imported might not include this information, and you will only be able to see them represented by their name. However, sequences downloaded from databases like GenBank will include this information. To change how sequences are displayed:

right-click any element or folder in the Navigation Area | Sequence Representation | select format

This will only affect sequence elements, and the display of other types of elements, e.g. alignments, trees and external files, will be not be changed. If a sequence does not have this information, there will be no text next to the sequence icon.

Rename element

Renaming a folder or an element in the **Navigation Area** can be done in two different ways:

select the element | Edit in the Menu Bar | Rename

or **select the element | F2**

When you can rename the element, you can see that the text is selected and you can move the cursor back and forth in the text. When the editing of the name has finished; press **Enter** or select another element in the **Navigation Area**. If you want to discard the changes instead, press the **Esc**-key.

For renaming annotations instead of folders or elements, see section 10.3.3.

3.1.7 Delete, restore and remove elements

When one deletes data from a data folder in the Workbench, it is moved to the recycle bin in that data location. Each data location has its own recycle bin. From the recycle bin, data can then be restored, or completely removed. Removal of data from the recycle bin frees disk space.

Deleting a folder or an element from a Workbench data location can be done in two ways:

right-click the element | Delete ( **)**

or **select the element | press Delete key**

This will cause the element to be moved to the **Recycle Bin** () where it is kept until the recycle bin is emptied or until you choose to restore the data object to your data location.

For deleting annotations instead of folders or elements, see section 10.3.4.

Items in a recycle bin can be restored in two ways:

Drag the elements with the mouse into the folder where they used to be.

or **select the element | right click and choose the option Restore**.

Once restored, you can continue to work with that data.

All contents of the recycle bin can be removed by choosing to empty the recycle bin:

Edit in the Menu Bar | Empty Recycle Bin ( **)**

This deletes the data and frees up disk space.

Note! This cannot be undone. Data is not recoverable after it is removed by emptying the recycle bin.

3.1.8 Show folder elements in a table

A location or a folder might contain large amounts of elements. It is possible to view their elements in the **View Area**:

> **select a folder or location** | **Show (→) in the Toolbar**

or

> **select a folder or location** | right click on the folder and select **Show (→)** | **Contents (📁)**

An example is shown in figure 3.6.

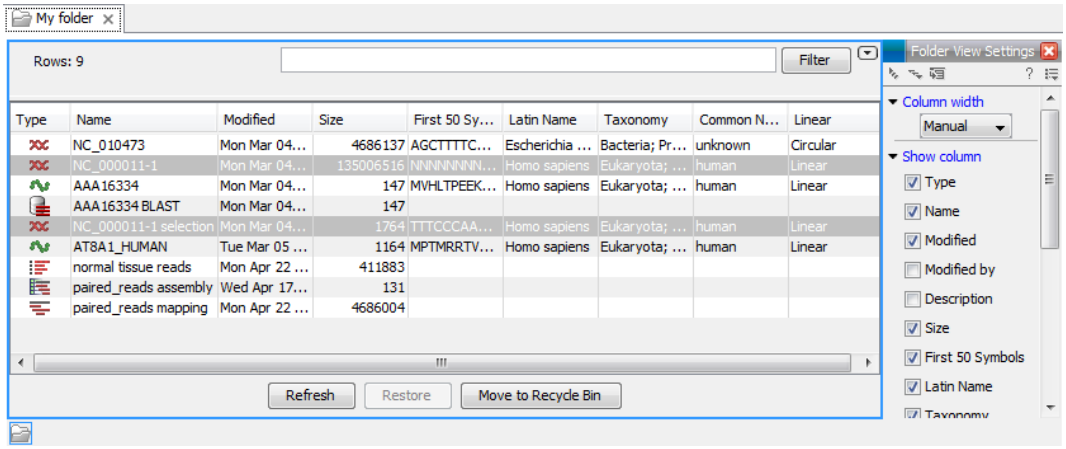


Figure 3.6: *Viewing the elements in a folder.*

When the elements are shown in the view, they can be sorted by clicking the heading of each of the columns. You can further refine the sorting by pressing Ctrl (⌘ on Mac) while clicking the heading of another column.

Sorting the elements in a view does not affect the ordering of the elements in the **Navigation Area**.

Note! The view only displays one "layer" at a time: the content of subfolders is not visible in this view. Also note that only sequences have the full span of information like organism etc.

Batch edit folder elements

You can select a number of elements in the table, right-click and choose **Edit** to batch edit the elements. In this way, you can change for example the description or name of several elements in one go.

In figure 3.7 you can see an example where the name of two sequence are renamed in one go. In this example, a dialog with a text field will be shown, letting you enter a new name for these two sequences.

Note! This information is directly saved and you cannot undo.

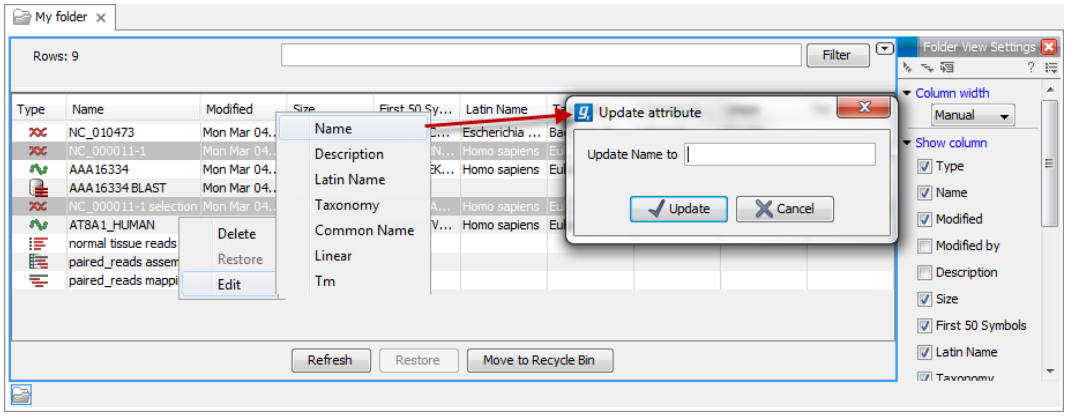


Figure 3.7: *Changing the common name of two sequences.*

Drag and drop folder elements

You can drag and drop objects from the folder editor to the **Navigation area**. This will create a copy of the objects at the selected destination. New elements can be included in the folder editor in the view area by dragging and dropping an element from a destination in the **Navigation Area** to the folder in the **Navigation Area** that you have open in the view area. It is not possible to drag elements directly from the **Navigation Area** to the folder editor in the View area.

3.2 Metadata

Metadata refers to information about data. In the context of the CLC Workbenches, this will usually mean information about samples. For example a set of reads could come from a particular specimen at a particular time point with particular characteristics. The specimen, time and characteristics would be metadata for that set of reads. The data can then be associated with its metadata in the Workbench. This can be useful for keeping track of related datasets and metadata can be used by some types of analyses in some CLC Workbenches.

Metadata can be created directly in the Workbench, but typically it will be imported from an external file (excel or text based). See section 3.2.1. It is then stored as a metadata table in the Workbench. An example of a metadata table as it might appear in the Workbench is shown in figure 3.8.

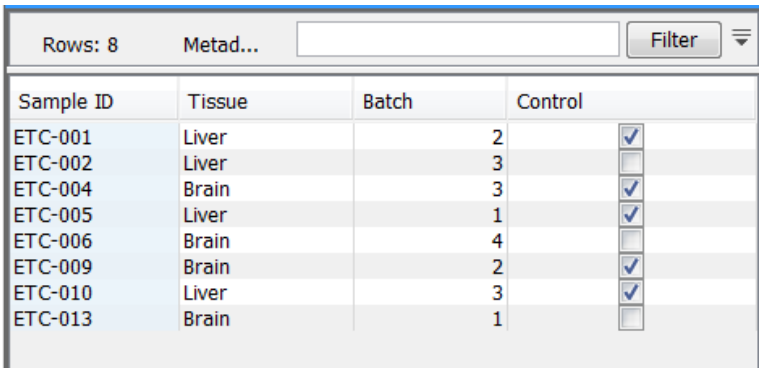


Figure 3.8: *A simple Metadata Table.*

Each column represents a property of a sample (e.g. identifier, height, age, treatment, etc.) and each row contain information relevant to a sample.

Within the CLC Workbench, one of the metadata table columns may be designated as the key column. The entries in a key column must be unique. Any column can be chosen to be the key column, but commonly it will be the first column and it would contain an identifier of some sort (e.g. a name).

There are no restrictions on the type of information that can be held in a metadata table. However, it is generally recommended that any given metadata table contains information about a related collection of entities. For example, a set of samples from the same experiment, or a set of families from the same study. Any particular data element can only be associated with *at most one* row in a given metadata table. However, that same data element can be associated with metadata in other metadata tables.

During or after metadata import, data can be associated with that metadata. Once a data element is associated with metadata, the outputs of analyses involving that data usually inherit the metadata association automatically. Inheritance like this is carried out when the metadata association for the outputs can be unambiguously identified. So, for example, if an output is derived from two inputs with different metadata associations, then neither association will be inherited by the output data elements.

Importing metadata can be done using a basic or advanced tool, and viewing and working with metadata, including data association, is done using the Metadata Table editor.

3.2.1 Importing Metadata

There are two tools that can be used to import metadata, one basic and one more advanced. A list of the benefits and limitations of each is included at the start the sections describing them.

Basic Metadata Import

The basic import tool is fast and easy, but less flexible than the advanced metadata import using the Metadata Table Editor. General features of this importer are:

- Can import from Excel (.xlsx/.xls) format files.
- The first column in the Excel file must have unique entries. It will be designated as the key column.
- Data elements in the Workbench can have metadata associated as part of the import if desired.
- Association with metadata is done by matching data element names with the entries in the first column of the spreadsheet. Name matching can be based on exact or partial matches.
- If metadata association will be done, visual feedback about the elements associated with metadata is provided in the Wizard.
- All columns will be imported as text columns.

If desired, a metadata table can be edited later from within the Metadata Table editor as described in section 3.2.2. There, you can change the column data types (e.g. to types of numbers, dates, true/false) and you can designate a new key column.

To run the basic importer, go to:

File | Import () | **Import Metadata** ()

- In the box labeled **Spreadsheet with sample information**, select the Excel file (.xlsx/.xls) to be imported.

  The rows in the spreadsheet are displayed in the Data Association table, as shown in figure 3.9.

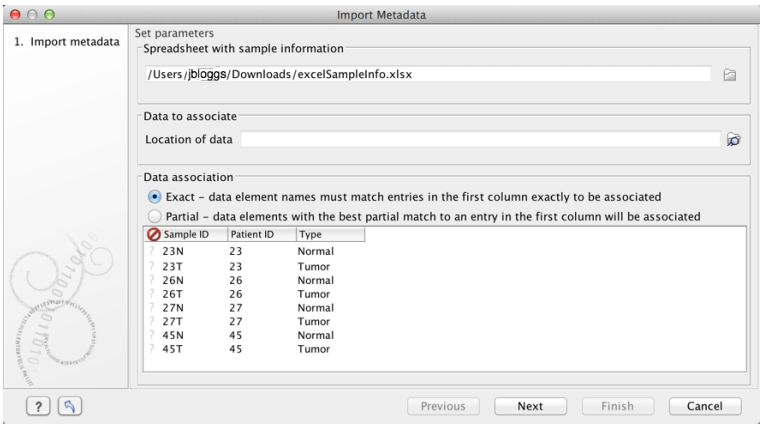


Figure 3.9: *After and excel file is selected, its rows are visible.*

  The ? symbols to the left of each row indicate that no data will be associated with that metadata row. You can choose just to proceed with the metadata import with no data association at this point by clicking on the the button labeled **Next**.

  If you wish to associate data with this metadata, then

- In the box labeled **Location of data**, select the data elements to have associations with this metadata.
- In the Data Association area, select whether data element names must exactly match the entries in the first column of the metadata to have an association created, or whether partial matches are allowed (figure 3.10). The two matching schemes are described in detail in section 3.2.2.
- Click on the button labeled **Next**.
- Select where you wish the metadata table to be saved.
- Click on the button labeled **Finish**.

The associated information can be viewed for a given data element in the Show Element Info view, as show in figure 3.11.

Advanced Metadata Import

The Metadata Table Editor can be used to import metadata from an external file. It involves more steps than the basic import tool, but is more flexible and has some basic error checking associated with data types. General features of this importer are:

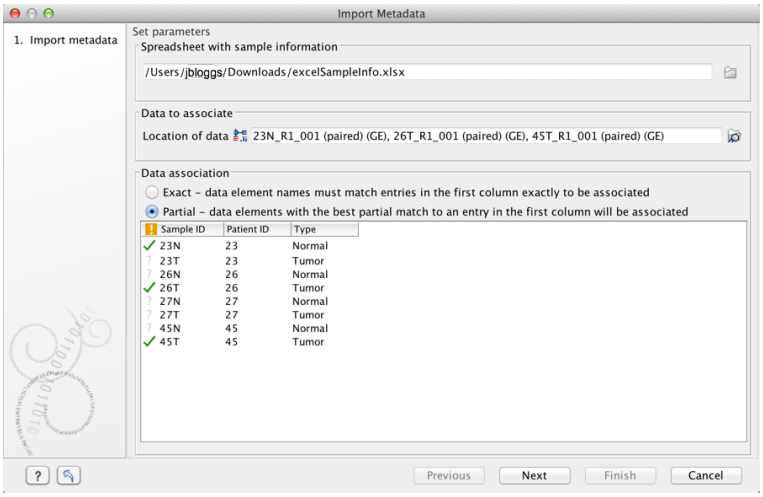


Figure 3.10: *Three data elements have been selected. Metadata rows that will have associations made to them have checkmarks beside them in the view of the table. Here, partial matching has been selected.*

Figure 3.11: *Metadata associations can be seen, edited, refreshed or deleted via the Show Element Info view.*

- Can import from Excel (.xlsx/.xls) or text files with a common delimiter can be used.
- The structure of the metadata table (the columns, their type, and the key column) must be set up before the metadata (contents) are imported.
- It is generally recommended that one column be designated as the key column. Entries in that column must have unique entries.
- The default data type for columns on creation is text, but this can be altered before import commences. When importing the metadata, an error will result if entries are found that do not match the expected data type.
- Association with metadata is done by matching data element names with the entries in the first column of the spreadsheet. Name matching can be based on exact or partial matches.
- Association of data with metadata is done as a separate step from import, providing flexibility. For example, if information in more than one column together uniquely identifies a sample, but the information within a single given column does not uniquely do so.

- Association of data with metadata can be done row by row if key column entries and the names of the relevant data elements are not related.

The Metadata Table Editor can also be used to create and populate a metadata table directly if desired.

Importing metadata using the Metadata Table Editor requires that the table structure is defined first. Information about how to do this is below. However, if the information about the data is in an excel file and the entries in the first column are unique, then the Import Metadata tool described in section 3.2.1 can be used to define the table and import the metadata in one step. That populated metadata table can then be opened in the Metadata Table Editor, where it can be refined, for example setting a new key column (or choosing not to have one), setting data types on columns, associating data as desired, and so on.

To set up a or alter a table structure using the Metadata Table Editor:

- Go to:

  File | **New** | **Metadata Table (** **)**

  This opens a new metadata table with no columns and no rows.

- Click on the button labeled **Setup Table** at the bottom of the view. A window appears like that shown in figure 3.12. The table structure will be defined using this tool.

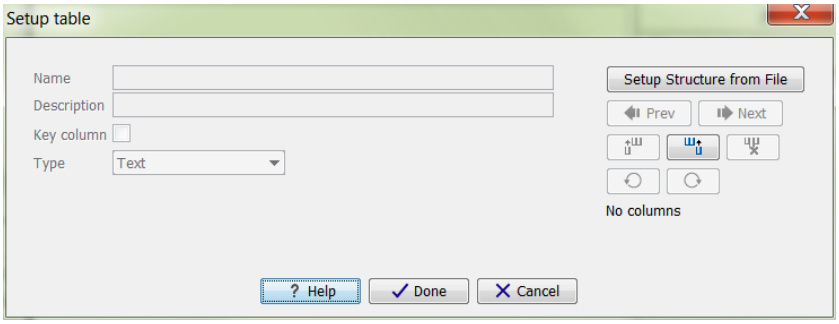


Figure 3.12: *Dialog used to add columns to an empty Metadata Table.*

- Click on the button labeled **Setup Structure from File**.

  A window will appear as shown in figure 3.13.

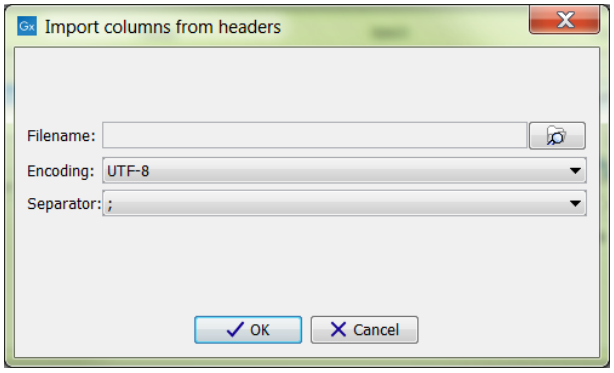


Figure 3.13: *Creating a metadata table structure based on an external file.*

You need to provide the following information:

  - **Filename** The Excel or delimited text file to import. Column names should be in the first row of this file.
  - **Encoding** For text files only: the encoding used to create the file. The default is UTF-8.
  - **Separator** For text files only: The character used to separate the columns. The default is semicolon (;).

  For each column in the external file, a column will be created in the new metadata table. By default the type of these imported columns is "Text". You will see a reminder to set the column type for each column and to designate one of the columns as the key column.

  You may modify the following information for each column:

  - **Name**. A mandatory header name or title for the column.
  - **Description**. An optional description of the information that will be held in the column. The description will appear as a tool tip, visible when you hover the mouse cursor over the column name in the metadata table.
  - **Key column**. Put a check in the box in the one column that will be the "key" column. All rows in this column must be populated and all entries in this column must be unique.
  - **Type**. The type of value allowed. The available types are:
    * **Text** Simple text.
    * **Whole number** Integer values, like `42` or `-7`.
    * **Decimal number** Decimal values, like `3.14` or `1.72e13`.
    * **Yes / No** `Yes/No` or `True/False` values are accepted. Capitalization is not necessary.
    * **Date** Local dates such as `2015-04-23` for April 23rd, 2015.
    * **Date and time** Local date and time such as `2015-04-23 13:37` for 1:37pm on April 23rd, 2015. Note the use of 24-hour clock and that no time zone information is present.

- Navigate between the columns using the (◀) Prev and (▶) Next buttons, or by using left/right arrow keys with Alt key held down.

  Modifications made to a particular column take effect as you navigate to another column, or if you close the dialog using **Done**.

  The (↺) and (↻) buttons are used undo and redo changes respectively.

- When the columns have been configured, click on the button labeled **Done**.

The table structure can also be defined manually by clicking on the (▦) button and defining each column in turn.

Columns may be deleted using the (▦✕) button. After metadata has been imported, additional columns can be added to the table structure. This can be done by importing the altered structure from an external file, where any columns not already in the metadata table will be added. Alternatively, individual columns can be added using the (▦◀) and (▦▶) buttons, which insert new columns before and after the current column respectively.

The metadata table can then be populated by importing information from an external file. The column names in the metadata table in the Workbench will be matched with those in the external

file to determine which values go into which cell. Only cell values in columns with an exact name match will be imported. If the file used contains columns not in the metadata table, the values in those columns will be ignored. Conversely, if the metadata table contains columns not present in the file, imported rows will have no values for those columns.

- Click on the button labeled **Manage Data** button at the bottom of the view. A window appears like that shown in in figure 3.14.

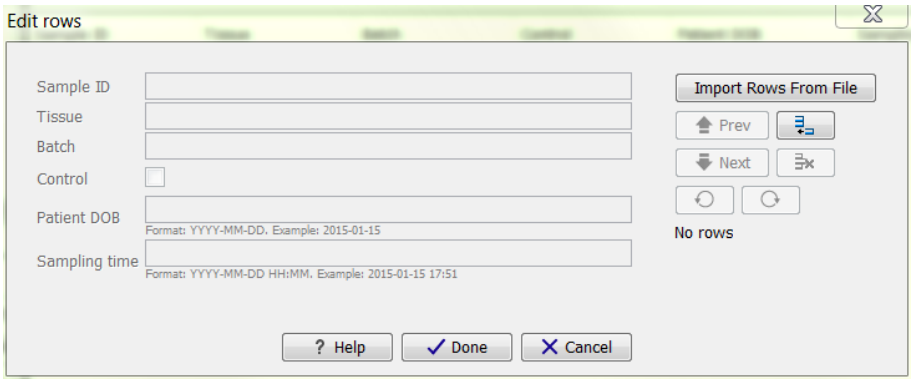


Figure 3.14: *Tool for managing the metadata itself. Notice the button labeled Import Rows from File.*

When working with an existing metadata table and adding extra rows, it is generally recommended that a key column be designated first. If a key column is not present, then all rows in the file will be imported. With no key column designated, if any rows from that file were imported into the same metadata table earlier, a duplicate row will be created. With a key column, rows with a new, unique entry for that column are added to the table and existing rows with a key entry in the file will be updated, incorporating any changes present in the file. Duplicate rows will not be created.

- Click on the button labeled **Import Rows from File** and select the external file of metadata. This brings up the window shown in figure 3.15.

  The options presented in that window are:

  - **File**. The file containing the metadata to import. This can be Excel (.xlsx/.xls) format or a delimited text file.
  - **Encoding**. For text files only: The text encoding of the seledcted file. Specifying the correct encoding is important to ensure that the file is correctly interpreted.
  - **Separator**. For text files only: the character used to separate columns in the file.
  - **Locale**. For text files only: the locale used to format numbers and dates within the file.
  - **Date format**. For text files only: the date format used in the imported file.
  - **Date-time format**. For text files only: the date-time format used in the imported file. The date and date-time templates uses the Java patterns for date and time formatting. Meaning of some of the symbols:

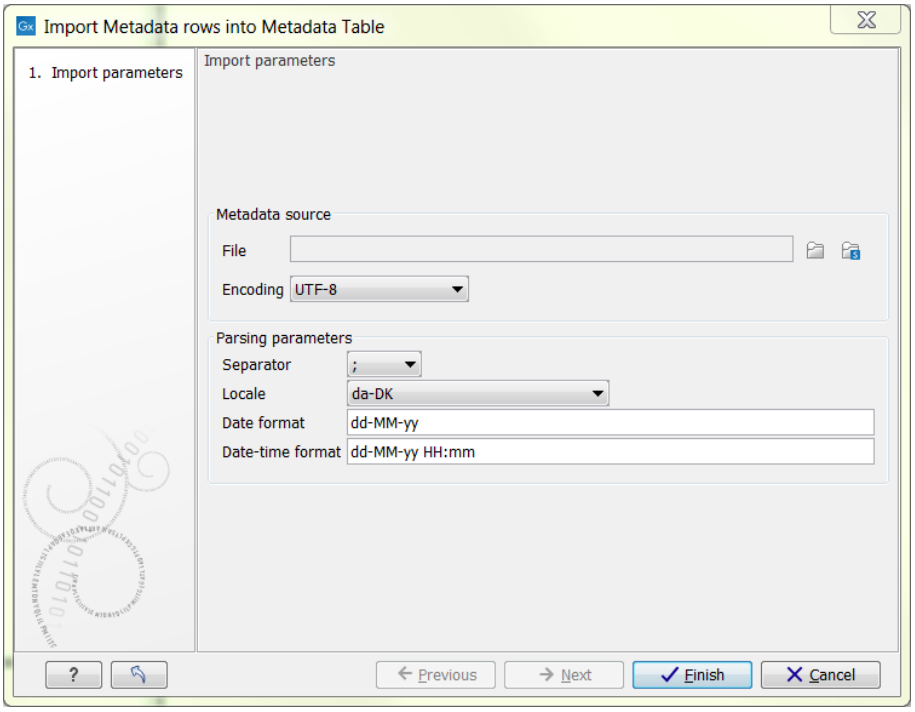


Figure 3.15: *Tool to import rows into a Metadata Table.*

Symbol	Meaning	Example
y	Year	2004; 04
d	Day	10
M/L	Month	7; 07; Jul; July; J
a	am-pm	PM
h	Hour (0-12 am pm)	12
H	Hour (0-23)	0
m	Minute	30
s	Second	55

Examples of using this:

Format	Meaning	Example
dd-MM-yy	Short date	31-12-15
yyyy-MM-dd HH:mm	Date and Time	2015-11-23 23:35
yyyy-MM-dd'T'HH:mm	ISO 8601 (standard) format	2015-11-23T23:35

With a short year format (YY), 2000 will be added when imported as, or converted to, Date or Date and time format. Thus, when working with dates before the year 2000 or after 2099, please use a four digit format for the year (YYYY).

- Click the button labeled **Finish** button when the necessary fields have been filled in.

Row information can also be added manually by clicking on the ( ) button and typing in the information for each column.

The progress and status of the row import can be seen in the Processes tab of the Toolbox. Any errors resulting from an import that failed can be reviewed here. The most frequent errors are

associated with selecting the wrong separator or encoding, or wrong date/time formats when importing rows from delimited text files.

Once the rows are imported, The metadata table can be saved.

3.2.2 Associating data elements with metadata

Typically, one would use the tools described in this section to associate data elements with metadata just after the data has been imported. Doing this at this early stage means that analysis results generated using these inputs will often inherit the metadata association. This inheritance is done when the relevant association can be determined unambiguously.

Each association between a particular data element and a row in your Metadata Table will have a "role" label that indicates what the role of the data element has. For example, a newly import sequence list could be given a role like "Sample data", or "NGS reads". Each analysis tool provides a particular role label when applying a metadata association to the outputs it generates. For example, a read mapping tool could assign the role "Un-mapped reads" to a sequence list of unmapped reads that it produces. When viewing all the data associated with a given metadata entry, these roles can help distinguish the particular data elements of interest.

The metadata table must be saved before data association options are available to use.

To associate data elements with the rows of a Metadata Table, click the **Associate Data** button at the bottom of the Metadata Table view.

If a key column has been identified for the metadata table, two options will be available:

- Association Data Automatically: The whole metadata table is used and associations between the selected data elements and the metadata are applied based on matching of the element name with the key column entries in the metadata table.
- Associate Data with Row: You select a row of the metadata and a particular data element and an association is then created. Information in the metadata table does not need to match the name of the data elements using this option. This option is also available when right-clicking a row in the table.

Each of these has benefits and restrictions. These are described at the top of each sections describing these options.

Associate Data Automatically

The main characteristics of the **Associate Data Automatically** tool are:

- Suited to associated large metadata tables or associating to many data elements.
- Well suited for use with newly imported data, where no associations already exist.
- Associations are created based on matching the information in the key colum of the metadata table with the name of the selected data elements.
- Two matching schemes are available: Exact and Partial (see section 3.2.2).

- A key column must be identified for the metadata table for this option to be available.
- Use with care with data elements that already have associations with the metadata table being worked with. As well as adding any new associations, existing associations will be *updated* to reflect the current information in the metadata table. This means associations will be *deleted* for a selected data element if there are no rows in the metadata table that match the name of that data element. See also the warning at the end of this section about this.

To run the **Associate Data Automatically** tool,

- Click the **Associate Data** button at the bottom of the Metadata Table view, and select **Associate Data Automatically**.

  Your metadata table must be saved and a key column designated for the metadata table for this option to be available.
- Select the data the tool should consider when setting up metadata associations in the window that appears. An example of this is shown in figure 3.16. You can select an item or sets of items in the navigation area on the left and move these into the selected elements list. Alternatively, you can right click on a folder and specify that all elements in the folder should be put in the selected elements list. This is illustrated in figure 3.17.

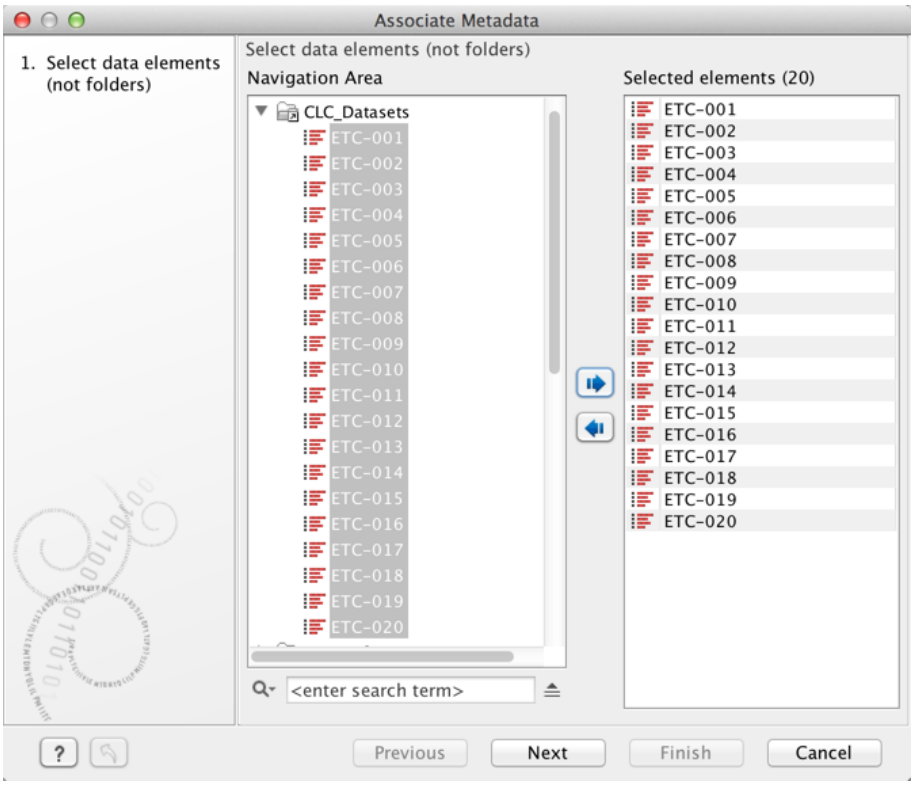


Figure 3.16: *Select data for automatic metadata association.*

.

- Click on the button labeled **Next**.

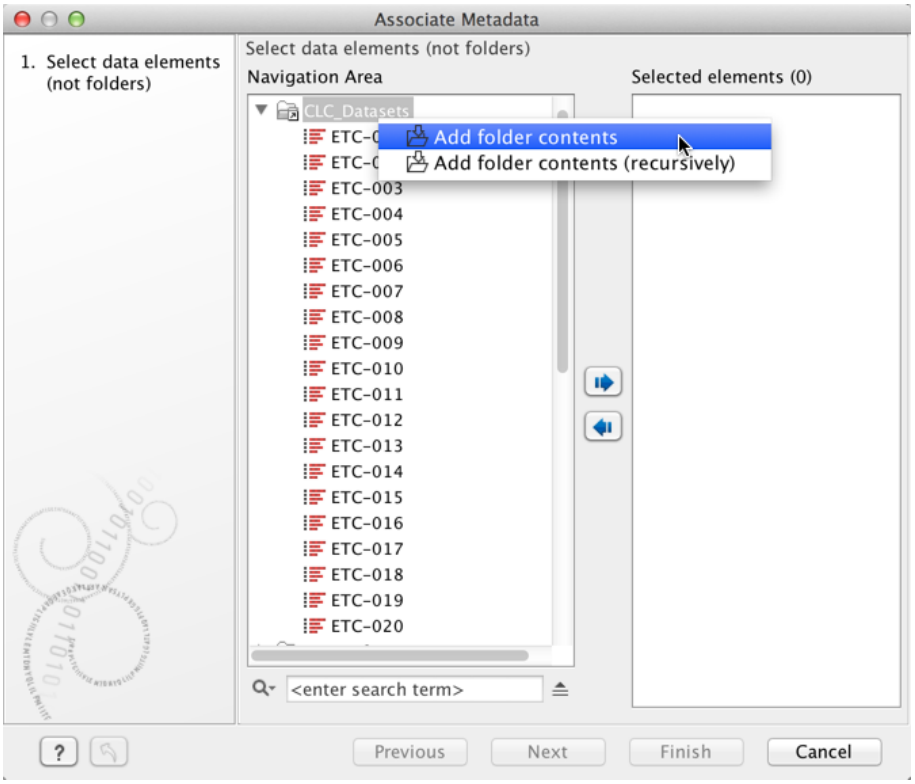


Figure 3.17: *Selecting all data elements in a folder.*

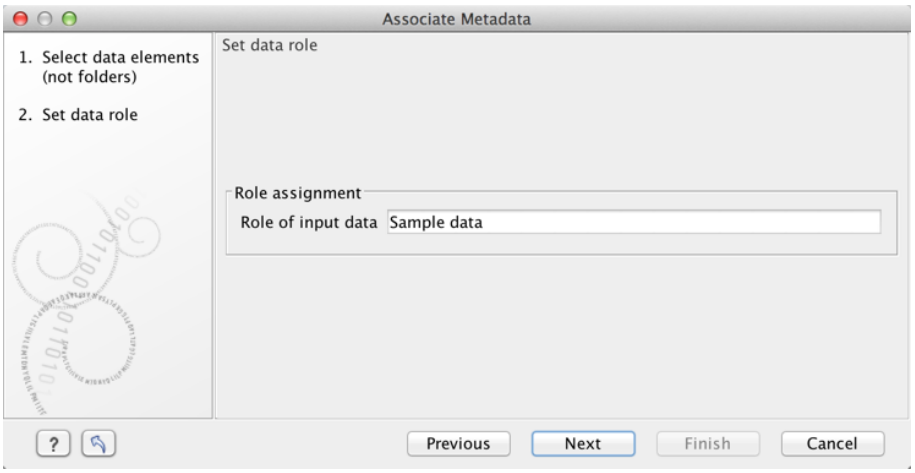


Figure 3.18: *Provide a role for the data elements. The default role provided is "Sample data".*

- Set the role that should be assigned to each data element that is associated to a metadata row (figure 3.18).
- Select whether the matching of the data element names to the entries in the key column should be based on exact or partial matching. These options are explained further below.
- Click on the button labeled **Next** and then choose to **Save** the outputs.

  Data associations and roles will be saved for data elements where the name matches a key column entry according to the selected matching scheme.

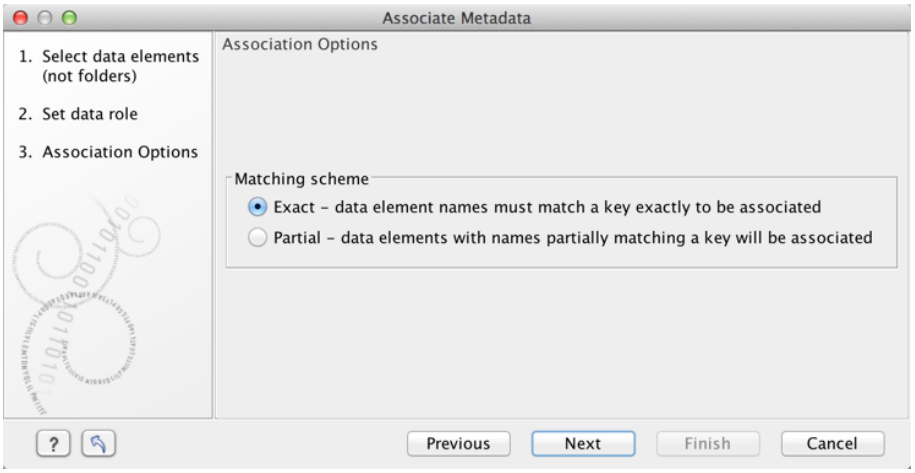


Figure 3.19: *Data element names can be matched either exactly or partially to the entries in the key column.*

Warning: It is safest only to select data elements that have no existing association to the metadata table being worked with, or carefully selecting any data elements with an existing association which you wish to update. All data selected that has an association with the metadata table being worked with will be *updated* by the automatic assocation tool. This means that any new or updated information in a metadata row can be added, but it also means that if no rows in the metadata match such a data element anymore, then the data association will be removed. This could happen if, for example, you changed the name of a data element with a metadata association, and did not change the corresponding key entry in the metadata table.

Matching schemes A data element name must match an entry in the key column of a metadata table for an association to be set up between that data element at the corresponding row of the metadata table. Two schemes are available in the **Association Data Automatically** for matching up names with key entries:

- Exact - data element names must match a key exactly to be associated. If any aspect of the key entry differs from the name of a selected data element, no association will be created.
- Partial - data elements with names partially matching a key will be associated. Here, data element names are broken into parts using common delimiters. The first whole part(s) must match a key entry in the metadata table for an assocation to be established. This option is explained in more detail below.

Partial matching rules For each data element being considered, the partial matching scheme involves breaking a data element name into components and searching for the best match from the key entries in the metadata table. In general terms, the best match means the longest key that matches entire components of the name.

The following describes the matching process in detail:

- Break the data element name into its component parts based on the presence of delimiters. It is these parts that are used for matching to the key entries of the metadata table.

Delimiters are any non-alphanumeric characters. That is, anything that is not a letter (a-z or A-Z) or number (0-9). So, for example, characters like hyphens (-), plus symbols (+), spaces, brackets, and so on, would be used as delimiters.

If partial matching was chosen with a data element called `Sample234-1 (mapped) (trimmed)` would be split into 4 parts: `Sample234`, `-1`, `(mapped)` and `(trimmed)`.

- Matches are made at the component level. A whole key entry must match perfectly to at least the first complete component of a data element name.

  For example, a key entry `Sample234` would be a match to the data element with name `Sample234-1 (mapped) (trimmed)` because the whole key entry matches the whole of the first component of the data element name. Conversely, if they key entry had been `Sample23`, no match would be identified, because they whole key entry does not match to at least the whole of the first component of the data element name.

  In cases where a data element could be matched to more than one key, the longest key matched determines the metadat row the data will be associated with.

  The table below provides examples to illustrate the partial matching system, on a table that has the keys with sample IDs like in figure 3.20) (i.e. `ETC-001`, `ETC-002`, . . . , `ETC-013`),

Data Element	Key	Reason for association
ETC-001 (Reads)	ETC-001	Key `ETC-001` matches the first part of the name
ETC-001 un-m. . . (single)	ETC-001	’’
ETC-001 un-m. . . (paired)	ETC-001	’’
ETC-002	ETC-002	Key `ETC-002` matches the whole name
ETC-003	None	No keys match this data element name
ETC-005	ETC-005	Key `ETC-005` matches the whole name
ETC-005-1	ETC-005	Key `ETC-005` matches the first part of the name
ETC-006-5	ETC-006	Key `ETC-006` matches the first part of the name
ETC-007	None	No keys match this data element name
ETC-007 (mapped)	None	’’
ETC-008	None	’’
ETC-008 (report)	None	’’
ETC-009	ETC-009	Key `ETC-009` matches the whole name

Associate Data with Row

The main characteristics of the **Associate Data with Row** tool are:

- Suited for association of a few metadata tables to a few data elements.
- Full control to select which data element should be associated with a particular metadata row.
- No requirement for a key column in the metadata table.
- No requirement for a relationship between the name of the data element and the metdata to associate it with.

To associate data elements with a particular row in the metadata table:

- Select the desired row in the metadata table by clicking on it.
- Right click and select the **Associate Data with Row** option (see figure 3.20), or click on the **Associate Data** button at the bottom of the view and choose the option **Associate Data with Row**.

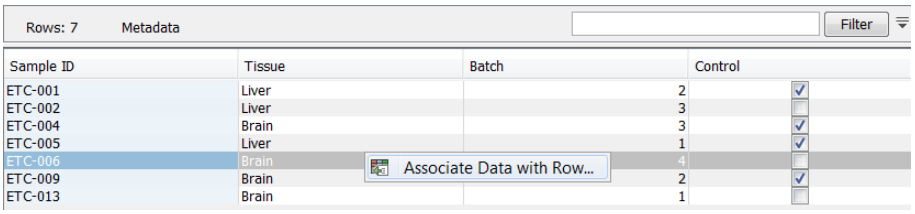


Figure 3.20: *Manual association of data elements to a metadata row.*

- A window will open within which you can select the data elements that should have an association with the metadata row.

  If a selected data element already has an association with this particular metadata table, that association will be updated. Associations with any other metadata tables will be left as they are.
- Click on the button labeled **Next**.
- Enter a role for the data elements that have been chosen.
- Click on the button labeled **Next**.
- Click on the button labeled **Next** and then choose to **Save** the outputs.

  Data associations and roles will be saved for the selected data elements.

3.2.3 Working with data and metadata

Finding data elements based on metadata

Using the Metadata Table view you can find data elements associated with rows of the metadata table. From this view, it is possible to launch analyses on selected data.

To find data elements associated with selected metadata rows:

- Select one or more rows of interest in the metadata table.
- Click on the button labeled **Find Associated Data** at the bottom of the view.

  A table with a listing of the data elements associated to the selected metadata row(s) will appear (figure 3.21).

The search results table shows the type, name, and navigation area path for each data element found. It also shows the key entry of the metadata table row with which the element is associated and the role of the data element for this metadata association. In figure 3.21, there are five data elements associated with sample ETC-009. Three are Sequence Lists, two of which have a role that tells us that they are un-mapped reads resulting from the Map Reads to Reference tool.

Clicking the **Refresh** button will re-run the search and refresh the search results table.

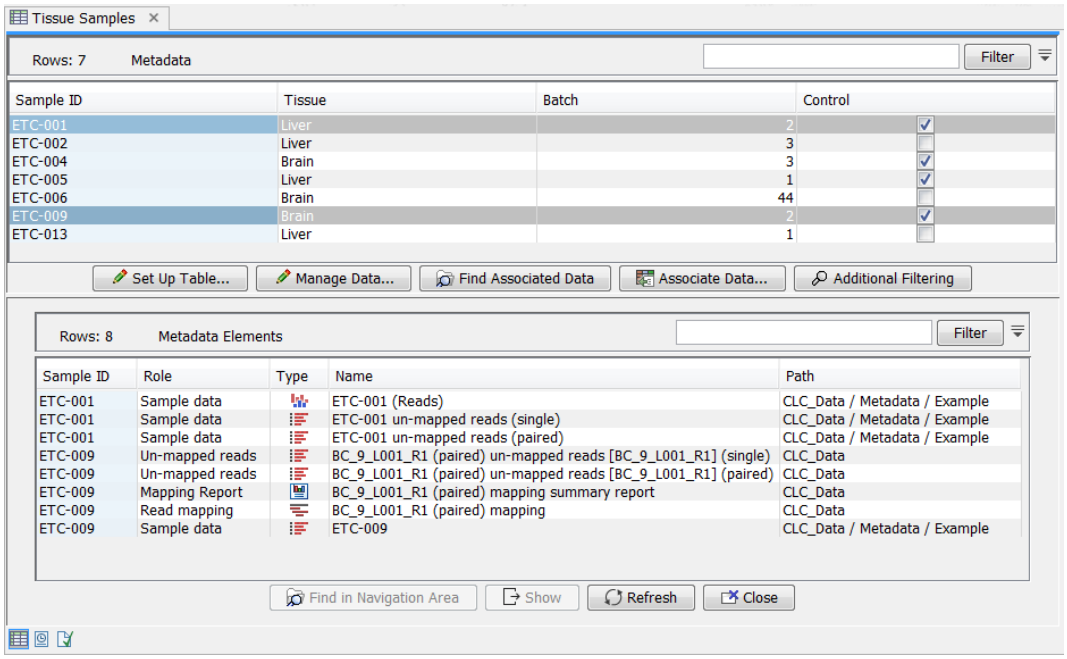


Figure 3.21: *Metadata Table with search results*

Click the button labeled **Close** to close the search table view.

Data elements listed in the search result table can be opend by clicking on the button labeled **Show** at the bottom of the view.

Alternatively, they can be highlighted in the Navigation Area by clicking the **Find in Navigation Area** button.

Analyses can be launched on the selected data elements:

- Directly. Right click on one of the selected elements and choosing the menu option Toolbox, and navigating to the tool of interest. The data selected in the search results table will be preselected in the Wizard that is launched.
- Via the Navigation area selection. Use the **Find in Navigation Area** button and then launch a tool in the Toolbox. The items that were selected in the Navigation area will be pre-selected in the Wizard that is launched.

Combining this functionality with the ability to filter for data elements in the search results box can be a powerful way to launch downstream analyses.

Identifying metadata rows without associated data

Using the Metadata Table view you can apply filters using the standard filtering tools shown at the top of the view as well as by using special metadata filtering in the **Additional Filtering** shown at the bottom. Using the special metadata filtering option **Show only Unassociated Rows**, you can filter the rows visible in the Metadata Table view so only the rows to which no data elements are associated are shown. If desired, these rows could then be used to launch one of the tools for associating data, described in section 3.2.2

This filter can be cancelled by clicking on the **Additional Filtering** button again and clicking on the **Show only Unassociated Rows** again. When the filter is active, it has a checkmark beside it. When it is inactive, it does not.

This filter can take a long time if many rows are shown in the table. When working with many rows, it can help if the full table is filtered using the general filters in advance, using the standard filters at the top of the table view. Alternatively you can pre-select some rows and filtering with the Additional filtering option **Filter to Selected Rows**. This filter can be applied multiple times. If the search takes too long, you can cancel it by unselecting the filter from the menu.

This filter can be cancelled by clicking on the **Additional Filtering** button again and clicking on the **Clear Selection Filter** option.

Viewing metadata associations

Metadata associations for a data element are shown in the Element Info view (section 10.4), see figure 3.22. To show Element Info,

right-click an element in the Navigation Area | Show | Element Info ( )

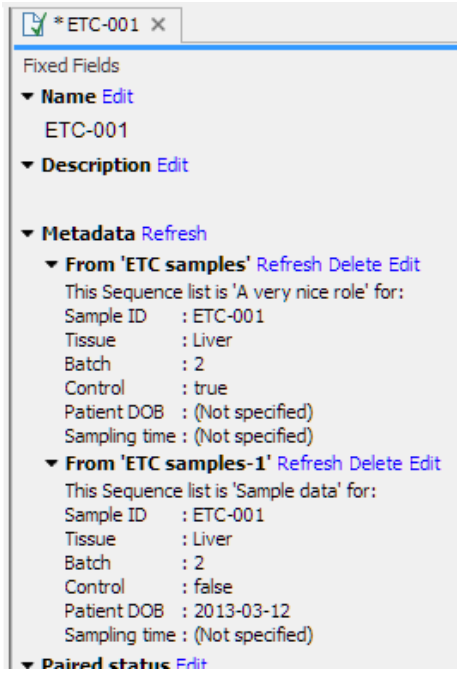


Figure 3.22: *Element Info view with a metadata association*

The Element Info view contains the details of each metadata association for the data element. The following operations are available:

- **Delete** will remove an association.
- **Edit** will allow you to change the role of the metadata association.
- **Refresh** will reload the metadata details from the Metadata Table; this functionality may be used to attempt to re-fetch metadata that was previously unavailable, e.g. due to server connectivity.

Exporting metadata

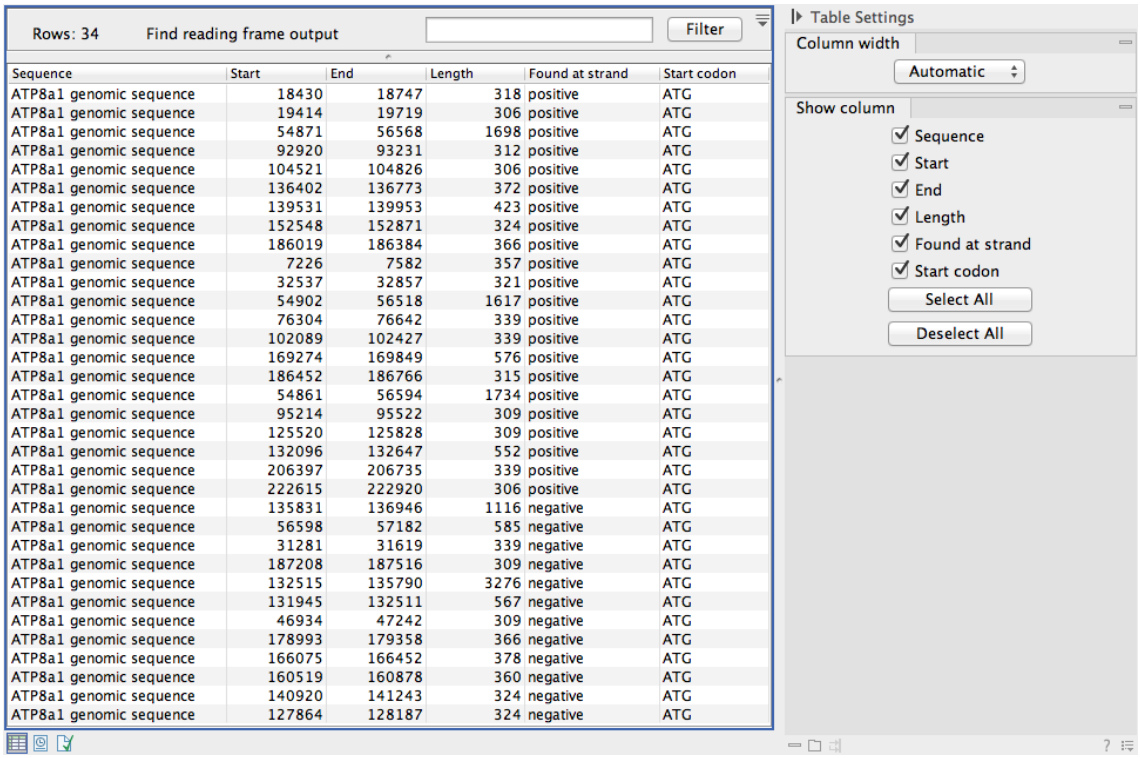
The standard Workbench export functionality can be used to export metadata tables to various formats. The system's default locale will be used for the export, which will affect the formatting of numbers and dates in the exported file.

See section 6.3 for more information.

3.3 Working with tables

Tables are used in a lot of places in the *CLC Drug Discovery Workbench*. There are some general features for all tables, irrespective of their contents, that are described here.

Figure 3.23 shows an example of a typical table. This is the table result of **Find Open Reading Frames** ( ). We use this table as an example to illustrate concepts relevant to all kinds of tables.



Sequence	Start	End	Length	Found at strand	Start codon
ATP8a1 genomic sequence	18430	18747	318	positive	ATG
ATP8a1 genomic sequence	19414	19719	306	positive	ATG
ATP8a1 genomic sequence	54871	56568	1698	positive	ATG
ATP8a1 genomic sequence	92920	93231	312	positive	ATG
ATP8a1 genomic sequence	104521	104826	306	positive	ATG
ATP8a1 genomic sequence	136402	136773	372	positive	ATG
ATP8a1 genomic sequence	139531	139953	423	positive	ATG
ATP8a1 genomic sequence	152548	152871	324	positive	ATG
ATP8a1 genomic sequence	186019	186384	366	positive	ATG
ATP8a1 genomic sequence	7226	7582	357	positive	ATG
ATP8a1 genomic sequence	32537	32857	321	positive	ATG
ATP8a1 genomic sequence	54902	56518	1617	positive	ATG
ATP8a1 genomic sequence	76304	76642	339	positive	ATG
ATP8a1 genomic sequence	102089	102427	339	positive	ATG
ATP8a1 genomic sequence	169274	169849	576	positive	ATG
ATP8a1 genomic sequence	186452	186766	315	positive	ATG
ATP8a1 genomic sequence	54861	56594	1734	positive	ATG
ATP8a1 genomic sequence	95214	95522	309	positive	ATG
ATP8a1 genomic sequence	125520	125828	309	positive	ATG
ATP8a1 genomic sequence	132096	132647	552	positive	ATG
ATP8a1 genomic sequence	206397	206735	339	positive	ATG
ATP8a1 genomic sequence	222615	222920	306	positive	ATG
ATP8a1 genomic sequence	135831	136946	1116	negative	ATG
ATP8a1 genomic sequence	56598	57182	585	negative	ATG
ATP8a1 genomic sequence	31281	31619	339	negative	ATG
ATP8a1 genomic sequence	187208	187516	309	negative	ATG
ATP8a1 genomic sequence	132515	135790	3276	negative	ATG
ATP8a1 genomic sequence	131945	132511	567	negative	ATG
ATP8a1 genomic sequence	46934	47242	309	negative	ATG
ATP8a1 genomic sequence	178993	179358	366	negative	ATG
ATP8a1 genomic sequence	166075	166452	378	negative	ATG
ATP8a1 genomic sequence	160519	160878	360	negative	ATG
ATP8a1 genomic sequence	140920	141243	324	negative	ATG
ATP8a1 genomic sequence	127864	128187	324	negative	ATG

Figure 3.23: *A table showing the results of an open reading frames analysis.*

Table viewing options

Options relevant to the view of the table can be configured in the **Side Panel** on the right.

For example, the columns that can be dispalyed in the table are listed in the section called **Show column**. The checkboxes allow you to see or hide any of the available columns for that table.

The Column width can be set to **Automatic** or **Manual**. By default, the first time you open a table, it will be set to **Automatic**. The default selected columns are hereby resized to fit the width of the viewing area. When changing to the **Manual** option, column widths will adjust to the actual header size, and each column size can subsequently by adjusted manually. When the table content exceeds the size of the viewing area, a horizontal scroll becomes available for navigation

across the columns.

Sorting tables

You can **sort** table according to the values of a particular column by clicking a column header. (Pressing Ctrl - ⌘ on Mac - while you click will refine the existing sorting).

Clicking once will sort in ascending order. A second click will change the order to descending. A third click will set the order back its original order.

3.3.1 Filtering tables

The final concept to introduce is **Filtering**. The table filter as an advanced and a simple mode. The simple mode is the default and is applied simply by typing text or numbers (see an example in figure 3.24).[1]

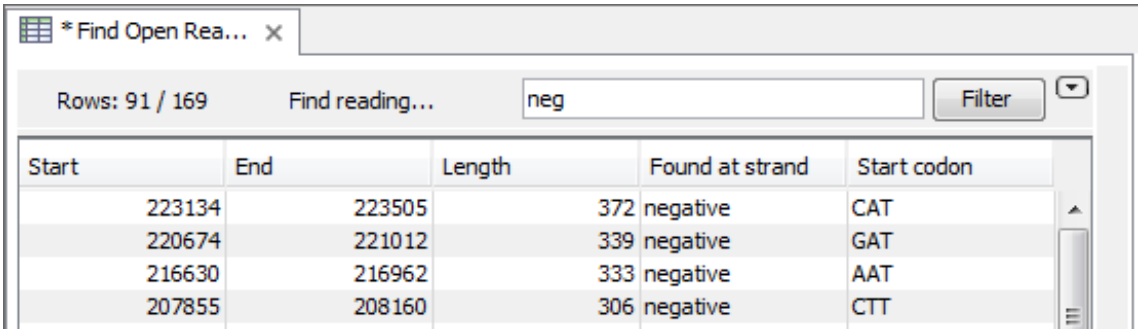


Figure 3.24: *Typing "neg" in the filter in simple mode.*

Typing "neg" in the filter will only show the rows where "neg" is part of the text in any of the columns (also the ones that are not shown). The text does not have to be in the beginning, thus "ega" would give the same result. This simple filter works fine for fast, textual and non-complicated filtering and searching.

However, if you wish to make use of numerical information or make more complex filters, you can switch to the advanced mode by clicking the **Advanced filter** ( ) button. The advanced filter is structure in a different way: First of all, you can have more than one criterion in the filter. Criteria can be added or removed by clicking the **Add** ( ) or **Remove** ( ) buttons. At the top, you can choose whether all the criteria should be fulfilled (**Match all**), or if just one of the needs to be fulfilled (**Match any**).

For each filter criterion, you first have to select which column it should apply to. Next, you choose an operator. For numbers, you can choose between:

- **=** (equal to)
- **<** (smaller than)
- **>** (greater than)
- **<>** (not equal to)

[1] Note that for tables with more than 10000 rows, you have to actually click the **Filter** button for the table to take effect.

- **abs. value <** (absolute value smaller than. This is useful if it doesn't matter whether the number is negative or positive)
- **abs. value >** (absolute value greater than. This is useful if it doesn't matter whether the number is negative or positive)

Note, that the number of digits displayed is a formatting option which can be set in the View Preferences. The true number may well be (slightly) larger. This behaviour can lead to problems when filtering on exact matches using the **=** (equal to) operator on numbers. Instead, users are advised to use two filters of inequalities (**<** (smaller than) and **>** (greater than)) delimiting a (small) interval around the target value.

For text-based columns, you can choose between:

- **starts with** (the text starts with your search term)
- **contains** (the text does not have to be in the beginning)
- **doesn't contain**
- **=** (the whole text in the table cell has to match, also lower/upper case)
- $\neq$ (the text in the table cell has to not match)
- **is in list** (The text in the table cell has to match one of the items of the list. Items are separated by comma, semicolon, or space. This filter is case-insensitive)

Once you have chosen an operator, you can enter the text or numerical value to use.

If you wish to reset the filter, simply remove () all the search criteria. Note that the last one will not disappear - it will be reset and allow you to start over.

Figure 3.25 shows an example of an advanced filter which displays the open reading frames larger than 400 that are placed on the negative strand.

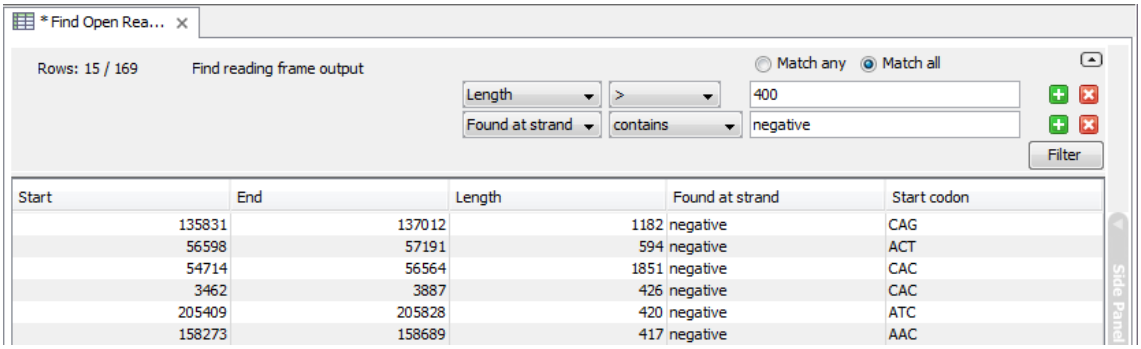


Figure 3.25: *The advanced filter showing open reading frames larger than 400 that are placed on the negative strand.*

Both for the simple and the advanced filter, there is a counter at the upper left corner which tells you the number of rows that pass the filter (91 in figure 3.24 and 15 in figure 3.25).

3.4 Customized attributes on data locations

Location-specific attributes can be set on all elements stored in a given data location. Attributes could be things like company-specific information such as LIMS id, freezer position etc. Attributes are set using a CLC Workbench acting as a client to the CLC Server.

Note that the attributes scheme belongs to a particular data location, so if there are multiple data locations, each will have its own set of attributes.

3.4.1 Configuring which fields should be available

To configure which fields that should be available[2] go to the Workbench:

right-click the data location | Location | Attribute Manager

This will display the dialog shown in figure 3.26.

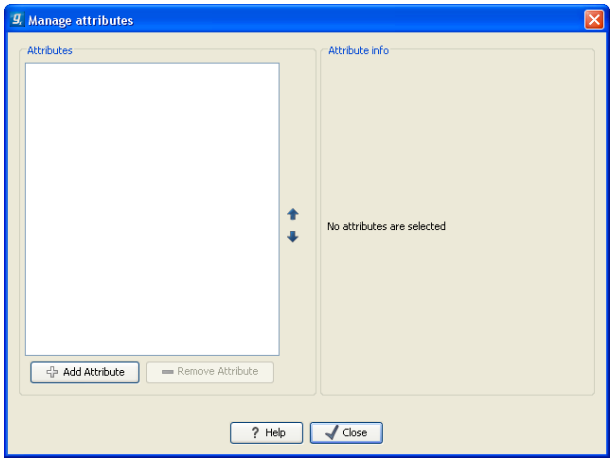


Figure 3.26: *Adding attributes.*

Click the **Add Attribute** ( ) button to create a new attribute. This will display the dialog shown in figure 3.27.

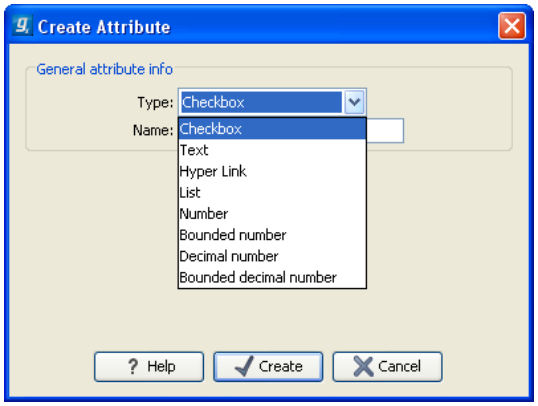


Figure 3.27: *The list of attribute types.*

First, select what kind of attribute you wish to create. This affects the type of information that can be entered by the end users, and it also affects the way the data can be searched. The following types are available:

[2] If the data location is a server location, you need to be a server administrator to do this

- **Checkbox**. This is used for attributes that are binary (e.g. true/false, checked/unchecked and yes/no).
- **Text**. For simple text with no constraints on what can be entered.
- **Hyper Link**. This can be used if the attribute is a reference to a web page. A value of this type will appear to the end user as a hyper link that can be clicked. Note that this attribute can only contain one hyper link. If you need more, you will have to create additional attributes.
- **List**. Lets you define a list of items that can be selected (explained in further detail below).
- **Number**. Any positive or negative integer.
- **Bounded number**. Same as number, but you can define the minimum and maximum values that should be accepted. If you designate some kind of ID to your sequences, you can use the bounded number to define that it should be at least 1 and max 99999 if that is the range of your IDs.
- **Decimal number**. Same as number, but it will also accept decimal numbers.
- **Bounded decimal number**. Same as bounded number, but it will also accept decimal numbers.

When you click **OK**, the attribute will appear in the list to the left. Clicking the attribute will allow you to see information on its type in the panel to the right.

3.4.2 Editing lists

Lists are a little special, since you have to define the items in the list. When you click a list in the left side of the dialog, you can define the items of the list in the panel to the right by clicking **Add Item** ( ) (see figure 3.28).

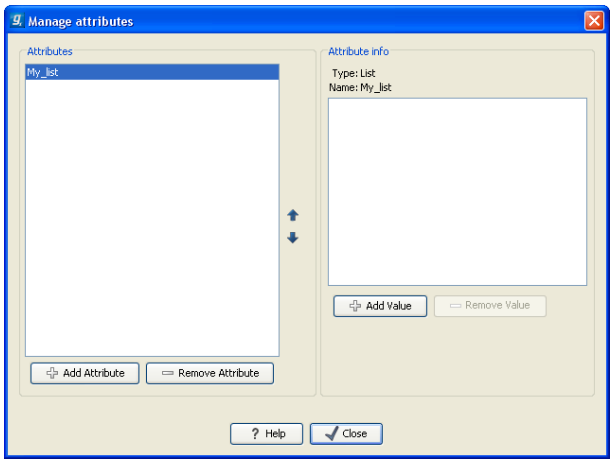


Figure 3.28: *Defining items in a list.*

Remove items in the list by pressing **Remove Item** ( ).

3.4.3 Removing attributes

To remove an attribute, select the attribute in the list and click **Remove Attribute** ( ). This can be done without any further implications if the attribute has just been created, but if you remove an attribute where values have already been given for elements in the data location, it will have implications for these elements: The values will not be removed, but they will become static, which means that they cannot be edited anymore.

If you accidentally removed an attribute and wish to restore it, this can be done by creating a new attribute of exactly the same name and type as the one you removed. All the "static" values will now become editable again.

When you remove an attribute, it will no longer be possible to search for it, even if there is "static" information on elements in the data location.

Renaming and changing the type of an attribute is not possible - you will have to create a new one.

3.4.4 Changing the order of the attributes

You can change the order of the attributes by selecting an attribute and click the **Up** and **Down** arrows in the dialog. This will affect the way the attributes are presented for the user.

3.4.5 Filling in values

When a set of attributes has been created (as shown in figure 3.29), the end users can start filling in information.

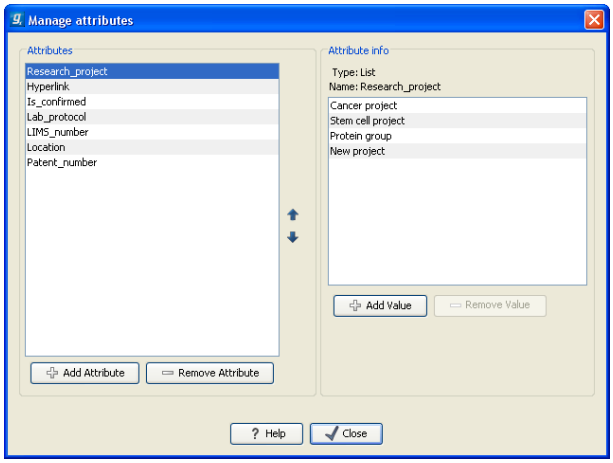


Figure 3.29: *A set of attributes defined in the attribute manager.*

This is done in the element info view:

> **right-click a sequence or another element in the Navigation Area | Show ( ) | Element info ( )**

This will open a view similar to the one shown in figure 3.30.

You can now enter the appropriate information and **Save**. When you have saved the information, you will be able to search for it (see below).

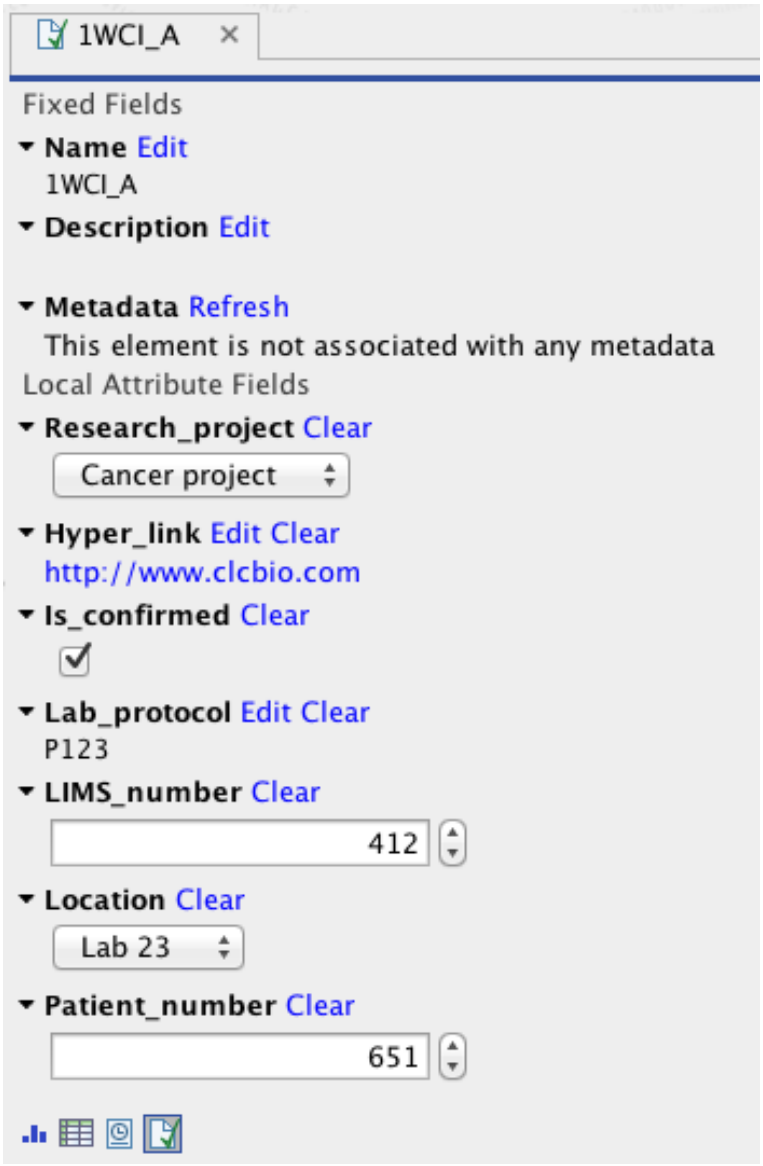


Figure 3.30: *Adding values to the attributes.*

Note that the element (e.g. sequence) needs to be saved in the data location before you can edit the attribute values.

When nobody has entered information, the attribute will have a "Not set" written in red next to the attribute (see figure 3.31).

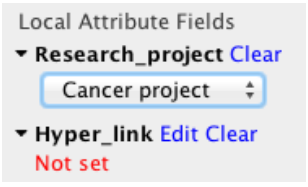


Figure 3.31: *An attribute which has not been set.*

This is particularly useful for attribute types like checkboxes and lists where you cannot tell, from the displayed value, if it has been set or not. Note that when an attribute has not been set, you cannot search for it, even if it looks like it has a value. In figure 3.31, you will *not* be able to find

this sequence if you search for research projects with the value "Cancer project", because it has not been set. To set it, simply click in the list and you will see the red "Not set" disappear.

If you wish to reset the information that has been entered for an attribute, press "Clear" (written in blue next to the attribute). This will return it to the "Not set" state.

The **Folder editor**, invoked by pressing **Show** on a given folder from the context menu, provides a quick way of changing the attributes of many elements in one go (see section 3.1.8).

3.4.6 What happens when a clc object is copied to another data location?

The user supplied information, which has been entered in the **Element info**, is attached to the attributes that have been defined in this particular data location. If you copy the sequence to another data location or to a data location containing another attribute set, the information will become fixed, meaning that it is no longer editable and cannot be searched for. Note that attributes that were "Not set" will disappear when you copy data to another location.

If the element (e.g. sequence) is moved back to the original data location, the information will again be editable and searchable.

If the e.g. Molecule Project or Molecule Table is moved back to the original data location, the information will again be editable and searchable.

3.4.7 Searching

When an attribute has been created, it will automatically be available for searching. This means that in the **Local Search** ( ), you can select the attribute in the list of search criteria (see figure 3.32).

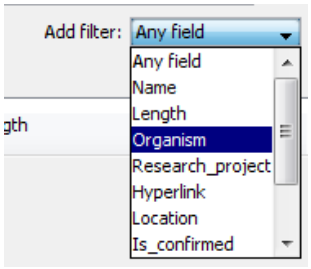


Figure 3.32: *The attributes from figure 3.29 are now listed in the search filter.*

It will also be available in the **Quick Search** below the **Navigation Area** (press Shift+F1 (Fn+Shift+F1 on Mac) and it will be listed - see figure 3.33).

3.5 Local search

There are two ways of doing text-based searches of your data, as described in this section:

- **Quick-search** directly from the search field in the **Navigation Area**.
- **Advanced search** which makes it easy to make more specific searches.

In most cases, quick-search will find what you need, but if you need to be more specific in your search criteria, the advanced search is preferable.

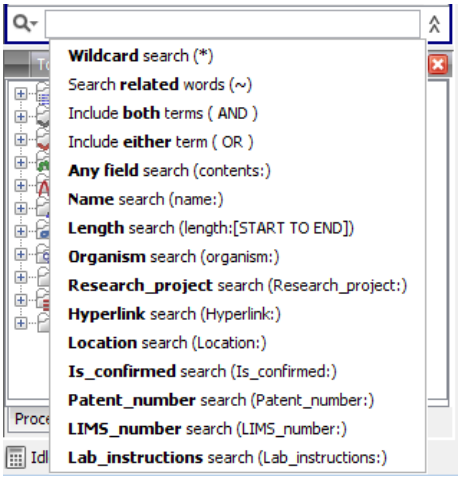


Figure 3.33: *The attributes from figure 3.29 are now available in the Quick Search as well.*

3.5.1 What kind of information can be searched?

Below is a list of the different kinds of information that you can search for (applies to both quick-search and the advanced search).

- **Name.** The name of a sequence, an alignment or any other kind of element. The name is what is displayed in the **Navigation Area** per default.
- **Length.** The length of the sequence.
- **Organism.** Sequences which contain information about organism can be searched. In this way, you could search for e.g. *Homo sapiens* sequences.
- **Custom attributes.** Read more in section 3.4

Only the first item in the list, **Name**, is available for all kinds of data. The rest is only relevant for sequences.

If you wish to perform a search for sequence similarity, use Local BLAST (see section 12.1.3) instead.

3.5.2 Quick search

At the bottom of the **Navigation Area** there is a text field as shown in figure 3.34).

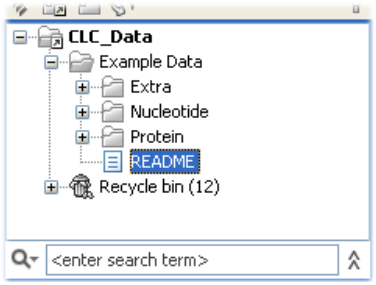


Figure 3.34: *Search simply by typing in the text field and press Enter.*

To search, simply enter a text to search for and press **Enter**.

Note that the search term supports advanced features known from web search engines, which means that the following list of characters carry special meaning: + - && || ! ( ) ^ [ ] “~* ? : \ /. To avoid this special interpretation it is suggested to put quotes around the search expression when searching for data containing the special characters, or read the section 3.5.3 on advanced search expressions.

Quick search results

To show the results, the search pane is expanded as shown in figure 3.35).

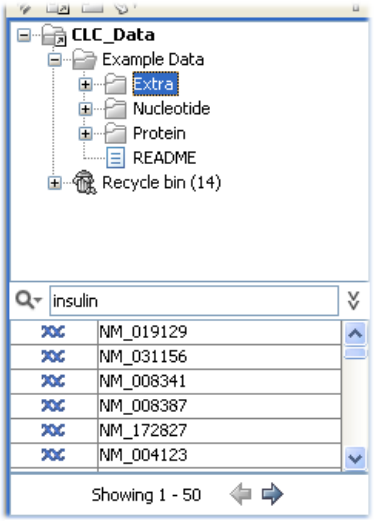


Figure 3.35: *Search results.*

If there are many hits, only the 50 first hits are immediately shown. At the bottom of the pane you can click **Next** () to see the next 50 hits (see figure 3.36).

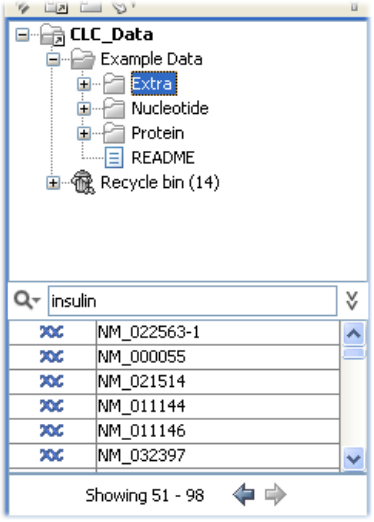


Figure 3.36: *Page two of the search results.*

If a search gives no hits, you will be asked if you wish to search for matches that start with your search term. If you accept this, an asterisk (*) will be appended to the search term.

Pressing the Alt key while you click a search result will high-light the search hit in its folder in the

Navigation Area.

In the preferences (see Chapter 4), you can specify the number of hits to be shown.

Special search expressions

When you write a search term in the search field, you can get help to write a more advanced search expression by pressing **Shift+F1**. This will reveal a list of guides as shown in figure 3.37.

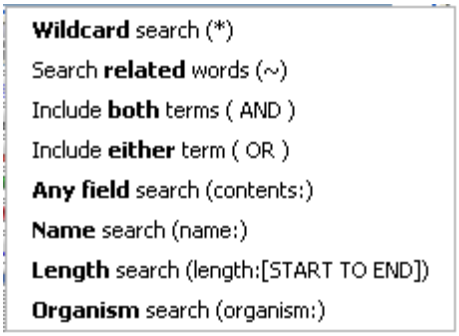


Figure 3.37: *Guides to help create advanced search expressions.*

You can select any of the guides (using mouse or keyboard arrows), and start typing. If you e.g. wish to search for sequences named BRCA1, select "Name search (name:)", and type "BRCA1". Your search expression will now look like this: "name:BRCA1".

The guides available are these:

- **Wildcard search (*)**. Appending an asterisk * to the search term will find matches starting with the term. E.g. searching for "brca*" will find both *brca1* and *brca2*.
- **Search related words (~)**. If you don't know the exact spelling of a word, you can append a tilde to the search term. E.g. "brac1~" will find sequences with a *brca1* gene.
- **Include both terms (AND)**. If you write two search terms, you can define if your results have to match both search terms by combining them with AND. E.g. search for "brca1 AND human" will find sequences where *both* terms are present.
- **Include either term (OR)**. If you write two search terms, you can define that your results have to match either of the search terms by combining them with OR. E.g. search for "brca1 OR brca2" will find sequences where *either* of the terms is present.
- **Do not include term (NOT)** If you write a term after not, then elements with these terms will not be returned.
- **Name search (name:)**. Search only the name of element.
- **Organism search (organism:)**. For sequences, you can specify the organism to search for. This will look in the "Latin name" field which is seen in the **Sequence Info** view (see section 10.4).
- **Length search (length:[START TO END])**. Search for sequences of a specific length. E.g. search for sequences between 1000 and 2000 residues: "length:1000 TO 2000".

Note! If you have added attributes (see section 3.4), these will also appear on the list when pressing **Shift+F1**.

If you do not use this special syntax, you will automatically search for both name, description, organism, etc., and search terms will be combined as if you had put OR between them.

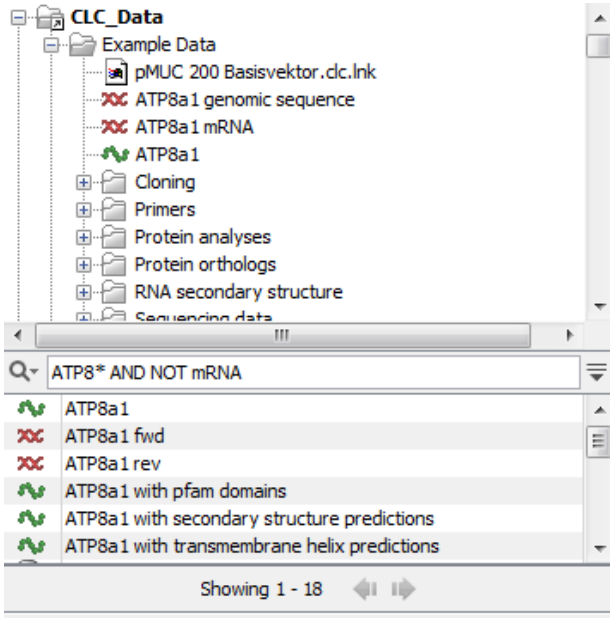


Figure 3.38: *An example of searching for elements with the name, description and organsim information that includes "ATP8" but do not include the term "mRNA".*

Search for data locations

The search function can also be used to search for a specific URL. This can be useful if you work on a server and wish to share a data location with another user. A simple example is shown in figure 3.39. Right click on the object name in the **Navigation Area** (in this case ATP8a1 genomic sequence) and select "Copy". When you use the paste function in a destination outside the Workbench (e.g. in a text editor or in an email), the data location will become visible. The URL can now be used in the search field in the Workbench to locate the object.

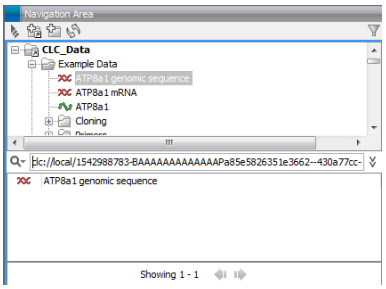


Figure 3.39: *The search field can also be used to search for data locations.*

Quick search history

You can access the 10 most recent searches by clicking the icon (🔍) next to the search field (see figure 3.40).

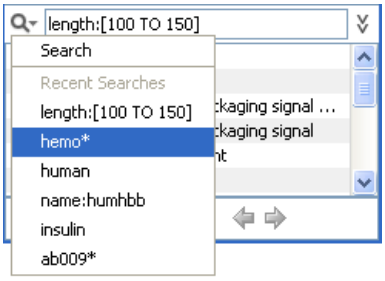


Figure 3.40: *Recent searches.*

Clicking one of the recent searches will conduct the search again.

3.5.3 Advanced search

As a supplement to the **Quick search** described in the previous section you can use the more advanced search:

Edit | Local Search ( )

or **Ctrl + Shift + F (⌘ + Shift + F on Mac)**

This will open the search view as shown in figure 3.41

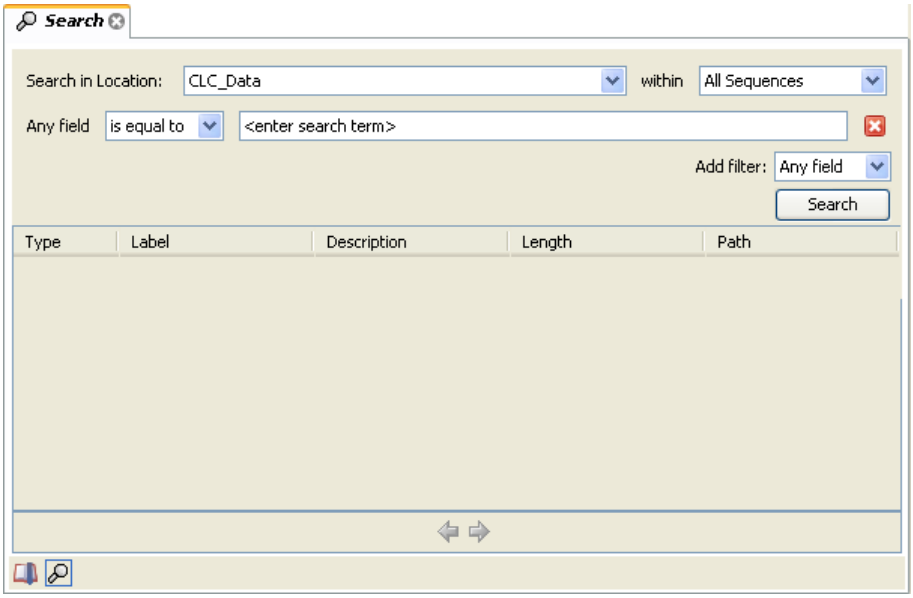


Figure 3.41: *Advanced search.*

The first thing you can choose is which location should be searched. All the active locations are shown in this list. You can also choose to search all locations. Read more about locations in section 3.1.1.

Furthermore, you can specify what kind of elements should be searched:

- All sequences
- Nucleotide sequences

- Protein sequences
- All data

When searching for sequences, you will also get alignments, sequence lists etc as result, if they contain a sequence which match the search criteria.

Below are the search criteria. First, select a relevant search filter in the **Add filter:** list. For sequences you can search for

- Name
- Length
- Organism

See section 3.5.2 for more information on individual search terms.

For all other data, you can only search for name.

If you use **Any field**, it will search all of the above plus the following:

- Description
- Keywords
- Common name
- Taxonomy name

To see this information for a sequence, switch to the **Element Info** ( ) view (see section 10.4).

For each search line, you can choose if you want the exact term by selecting "is equal to" or if you only enter the start of the term you wish to find (select "begins with").

An example is shown in figure 3.42.

This example will find human nucleotide sequences (organism is *Homo sapiens*), and it will only find sequences shorter than 10,000 nucleotides.

Note that a search can be saved ( ) for later use. You do not save the search results - only the search parameters. This means that you can easily conduct the same search later on when your data has changed.

3.5.4 Search index

This section has a technical focus and is not relevant if your search works fine.

However, if you experience problems with your search results: if you do not get the hits you expect, it might be because of an index error.

The *CLC Drug Discovery Workbench* automatically maintains an index of all data in all locations in the **Navigation Area**. If this index becomes out of sync with the data, you will experience problems with strange results. In this case, you can rebuild the index:

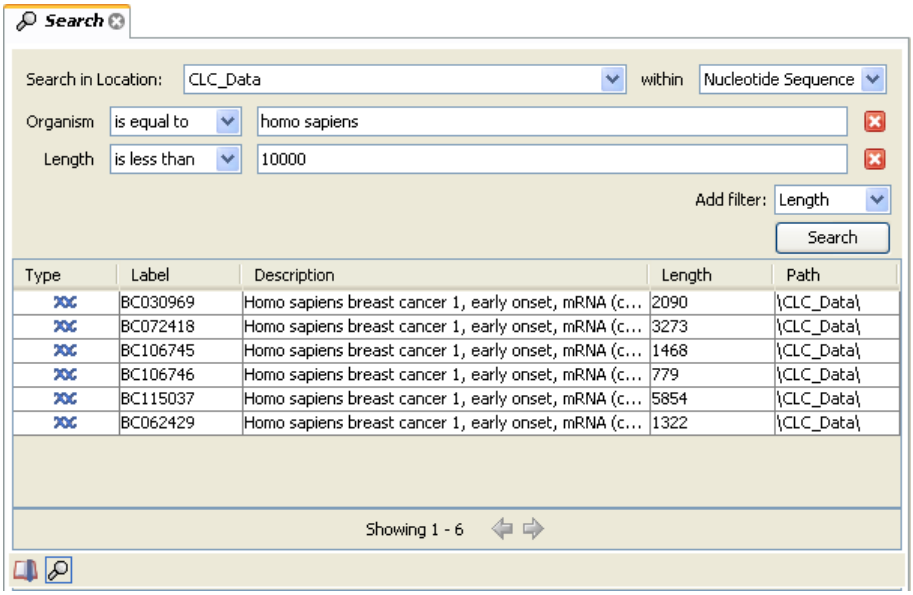


Figure 3.42: *Searching for human sequences shorter than 10,000 nucleotides.*

Right-click the relevant location | Location | Rebuild Index

This will take a while depending on the size of your data. At any time, the process can be stopped in the process area, see section 2.3.1.

Chapter 4

User preferences and settings

Contents

The first three sections in this chapter deal with the general preferences that can be set for *CLC Drug Discovery Workbench* using the **Preferences** dialog. The next section explains how the settings in the **Side Panel** can be saved and applied to other views. Finally, you can learn how to import and export the preferences.

The **Preferences** dialog offers opportunities for changing the default settings for different features of the program.

The **Preferences** dialog is opened in one of the following ways and can be seen in figure 4.1:

Edit | **Preferences (** **)**

or **Ctrl + K (⌘ + ; on Mac)**

4.1 General preferences

The **General** preferences include:

- **Undo Limit.** As default the undo limit is set to 500. By writing a higher number in this field, more actions can be undone. Undo applies to all changes made on molecules, sequences, alignments or trees. See section 2.1.5 for more on this topic.

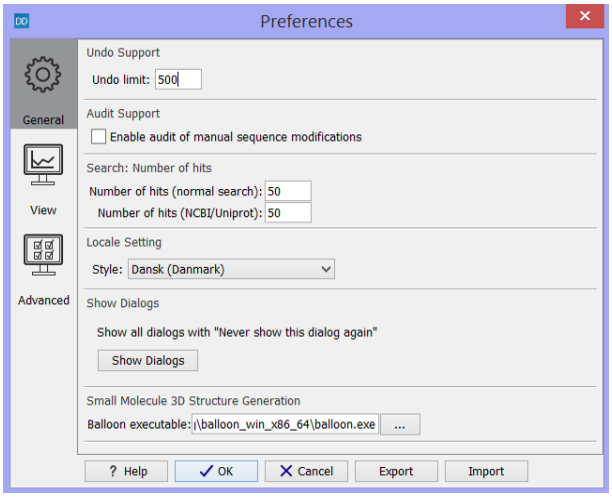


Figure 4.1: *Preferences include General preferences, View preferences, Data preferences, and Advanced settings.*

- **Audit Support.** If this option is checked, all manual editing of sequences will be marked with an annotation on the sequence (see figure 4.2). Placing the mouse on the annotation will reveal additional details about the change made to the sequence (see figure 4.3). Note that no matter whether **Audit Support** is checked or not, all changes are also recorded in the **History** ( ) (see section 2.1.2).
- **Number of hits.** The number of hits shown in *CLC Drug Discovery Workbench*, when e.g. searching NCBI. (The sequences shown in the program are not downloaded, until they are opened or dragged/saved into the Navigation Area).
- **Locale Setting.** Specify which country you are located in. This determines how punctation is used in numbers all over the program.
- **Show Dialogs.** A lot of information dialogs have a checkbox: "Never show this dialog again". When you see a dialog and check this box in the dialog, the dialog will not be shown again. If you regret and wish to have the dialog displayed again, click the button in the General Preferences: **Show Dialogs**. Then all the dialogs will be shown again.
- **Small Molecule 3D Structure Generation.** Here the location of the Balloon executable on the computer file system should be specified for the *Import Molecules from SMILES or 2D...* importer, and the paste of SMILES into a **Molecule Project** to work (see section 6.2.6 and section 6.2.7).

Figure 4.2: *Annotations added when the sequence is edited.*

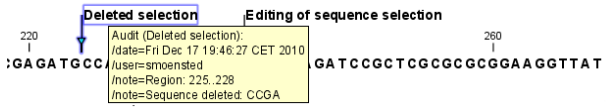


Figure 4.3: *Details of the editing.*

4.2 View preferences

There are six groups of default **View** settings:

1. **Toolbar** lets you choose the size of the toolbar icons, and whether to display names below the icons (figure 4.4).

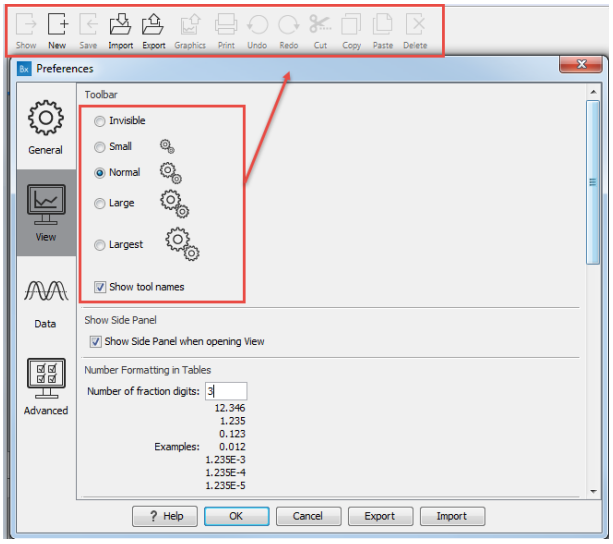


Figure 4.4: *Number formatting of tables.*

2. **Show Side Panel** allows you to choose whether to display the side panel when opening a new view. Note that for any open view, the side panel can be collapsed by clicking on the small triangle at the top left side of the settings area or by using the key combination Ctrl + U (⌘ + U on Mac).

3. **Number formatting in tables** specifies how the numbers should be formatted in tables (see figure 4.5). The examples below the text field are updated when you change the value so that you can see the effect. After you have changed the preference, you have to re-open your tables to see the effect.

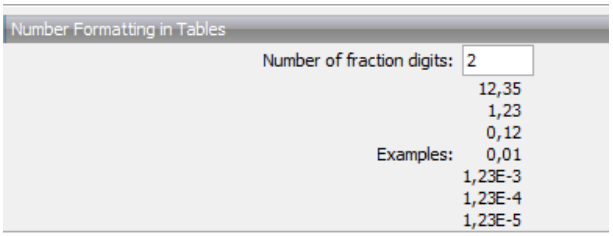


Figure 4.5: *Number formatting of tables.*

4. **Sequence Representation** allows you to change the way the elements appear in the Navigation Area. The following text can be used to describe the element:

   - Name (this is the default information to be shown).
   - Accession (sequences downloaded from databases like GenBank have an accession number).
   - Latin name.

- Latin name (accession).
- Common name.
- Common name (accession).

5. **User Defined View Settings** gives you an overview of the different Side Panel settings that are saved for each view. See section 4.5 to learn more about how to create and save style sheets. If there are other settings beside CLC Standard Settings, you can use this overview to choose which of the settings should be used per default when you open a view (see an example in figure 4.6).

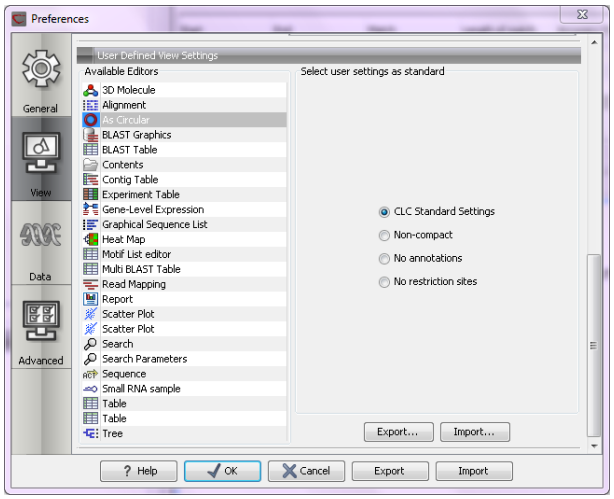


Figure 4.6: *Selecting the default view setting. In this example, the CLC Standard Settings is chosen as default.*

6. **Molecule Project 3D Editor** gives you the option to turn off the modern OpenGL rendering for **Molecule Projects** (see section 9.1.2).

4.2.1 Import and export Side Panel settings

If you have created a special set of settings in the **Side Panel** that you wish to share with other CLC users, you can export the settings in a file. The other user can then import the settings.

To export the **Side Panel** settings, first select the views that you wish to export settings for. Use Ctrl+click (⌘ + click on Mac) or Shift+click to select multiple views. Next click the **Export...**button. Note that there is also another export button at the very bottom of the dialog, but this will export the other settings of the **Preferences** dialog (see section 4.4).

A dialog will be shown (see figure 4.7) that allows you to select which of the settings you wish to export.

When multiple views are selected for export, all the view settings for the views will be shown in the dialog. Click **Export** and you will now be able to define a save folder and name for the exported file. The settings are saved in a file with a .vsf extension (View Settings File).

To import a **Side Panel** settings file, make sure you are at the bottom of the **View** panel of the **Preferences dialog**, and click the **Import...** button. Note that there is also another import button at the very bottom of the dialog, but this will import the other settings of the **Preferences** dialog (see section 4.4).

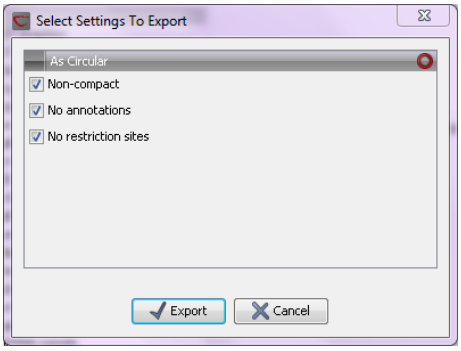


Figure 4.7: *Exporting all settings for circular views.*

The dialog asks if you wish to overwrite existing **Side Panel** settings, or if you wish to merge the imported settings into the existing ones (see figure 4.8).

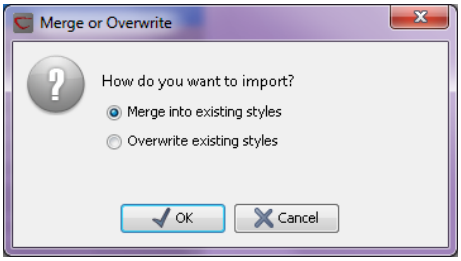


Figure 4.8: *When you import settings, you are asked if you wish to overwrite existing settings or if you wish to merge the new settings into the old ones.*

Note! If you choose to overwrite the existing settings, you will loose all the **Side Panel** settings that you have previously saved.

To avoid confusion of the different import and export options, here is an overview:

- Import and export of **bioinformatics data** such as sequences, alignments etc. (described in section 6.1).
- **Graphics** export of the views which creates image files in various formats (described in section 6.4).
- Import and export of **Side Panel Settings** as described above.
- Import and export of all the **Preferences** except the Side Panel settings. This is described in the previous section.

4.3 Advanced preferences

The **Advanced** settings include the possibility to set up a proxy server. This is described in section 1.8.

4.3.1 Default data location

The default location is used when you e.g. import a file without selecting a folder or element in the **Navigation Area** first.

The default data location for CLC Workbenches is, by default, a folder called CLC_Data in a user's home area.

This can be changed to a different location for a particular user of the Workbench by going to

Edit | **Preferences**

and then choosing the **Advanced** tab. This holds a section called **Default Data Location** and here you can choose a default from a drop down list of data locations you have already added.

Note! The default location cannot be removed. You have to select another location as default first.

If the data area you want as your default is not already available in your Workbench, you need to first add it as a new data location (see section 3.1.1).

4.3.2 NCBI BLAST

URL to use for BLAST

It is possible to specify an alternate server URL to use for BLAST searches. The standard URL for the BLAST server at NCBI is: http://blast.ncbi.nlm.nih.gov/Blast.cgi.

Note! Be careful to specify a valid URL, otherwise BLAST will not work.

4.4 Export/import of preferences

The user preferences of the *CLC Drug Discovery Workbench* can be exported to other users of the program, allowing other users to display data with the same preferences as yours. You can also use the export/import preferences function to backup your preferences.

To export preferences, open the **Preferences** dialog (Ctrl + K (⌘ + ; on Mac)) and do the following:

Export | **Select the relevant preferences** | **Export** | **Choose location for the exported file** | **Enter name of file** | **Save**

Note! The format of exported preferences is .cpf. This notation must be submitted to the name of the exported file in order for the exported file to work.

Before exporting, you are asked about which of the different settings you want to include in the exported file. One of the items in the list is "User Defined View Settings". If you export this, only the information about which of the settings is the default setting for each view is exported. If you wish to export the **Side Panel Settings** themselves, see section 4.2.1.

The process of importing preferences is similar to exporting:

Press Ctrl + K (⌘ + ; on Mac) to open Preferences | **Import** | **Browse to and select the .cpf file** | **Import and apply preferences**

4.4.1 The different options for export and import

To avoid confusion of the different import and export options, you can find an overview here:

- Import and export of **bioinformatics data** such as molecules, sequences, alignments etc.

(described in section 6.1).

- **Graphics** export of the views that create image files in various formats (described in section 6.4).
- Import and export of **Side Panel Settings** as described in the next section.
- Import and export of all the **Preferences** except the Side Panel settings. This is described above.

4.5 View settings for the Side Panel

The **Side Panel** is shown to the right of all views that are opened in *CLC Drug Discovery Workbench* and is described in further detail in section 2.1.8.

When you have adjusted a view of e.g. a sequence, your settings in the **Side Panel** can be saved. When you open other sequences, which you want to display in a similar way, the saved settings can be applied. The options for saving and applying are available at the bottom of the **Side Panel** (see figure 4.9).

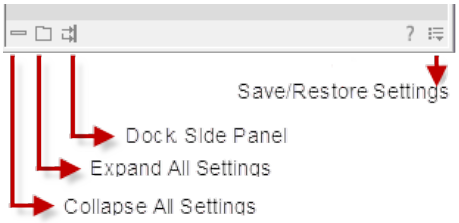


Figure 4.9: *At the bottom of the Side Panel you save the view settings*

4.5.1 Saving, removing and applying saved settings

To save and apply the saved settings, click ( ) seen in figure 4.9. This opens a menu where the following options are available (figure 4.10):

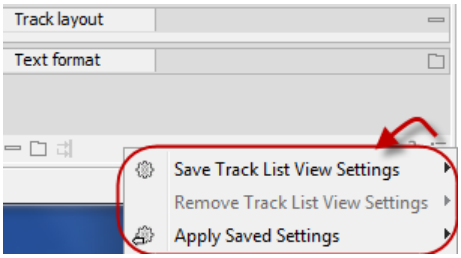


Figure 4.10: *When you have adjusted the side panel settings and would like to save these, this can be done with the "Save ... Settings" function, where "..." is the element you are working on - e.g. "Track List View", "Sequence View", "Table View", "Alignment View" etc. Saved settings can be deleted again with "Remove ... Settings" and can be applied to other elements with "Apply Saved Settings".*

- **Save ... Settings.** ( ) The settings can be saved in two different ways. When you select either way of saving settings a dialog will open (see figure 4.11) where you can enter a name for your settings.

- **For ... View in General** ( ) Will save the currently used settings with all elements of the same type as the one used for adjusting the settings. E.g. if you have selected to save settings "For Track View in General" the settings will be applied each time you open an element of the same type, which in this case means each time one of the saved tracks are opened from the **Navigation Area**. These "general" settings are user specific and will not be saved with or exported with the element.
- **On This Only** ( ) Settings can be saved with the specific element that you are working on in the View area and will not affect any other elements (neither in the View Area or in the **Navigation Area**). E.g. for a track you would get the option to save settings "On This Track Only". The settings are saved with only this element (and will be exported with the element if you later select to export the element to another destination).

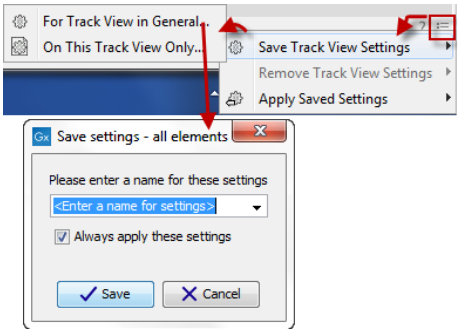


Figure 4.11: *The save settings dialog.Two options exist for saving settings. Click on the relevant option to open the dialog shown at the bottom of the figure.*

- **Remove ... Settings.** ( ) Gives you the option to remove settings specifically for the element that you are working on in the View Area, or on all elements of the same type. When you have selected the relevant option, the dialog shown in figure 4.12 opens and allows you to select which of the saved settings to remove.
  - **From ... View in General** ( ) Will remove the currently used settings on all elements of the same type as the one used for adjusting the settings. E.g. if you have selected to remove settings from all alignments using "From Alignment View in General", all alignments in your **Navigation Area** will be opened with the standard settings instead.
  - **From This ... Only** ( ) When you select this option, the selected settings will only be removed from the particular element that you are working on in the View area and will not affect any other elements (neither in the View Area or in the **Navigation Area**). The settings for this particular element will be replaced with the CLC standard settings ( ).

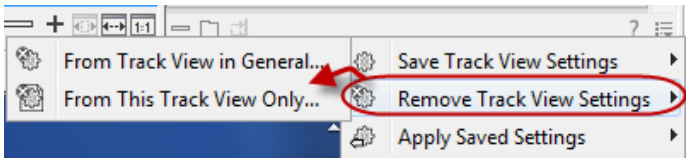


Figure 4.12: *The remove settings dialog for a track.*

- **Apply Saved Settings.** ( ) This is a submenu containing the settings that you have previously saved (figure 4.13). By clicking one of the settings, they will be applied to the current view. You will also see a number of pre-defined view settings in this submenu. They

are meant to be examples of how to use the **Side Panel** and provide quick ways of adjusting the view to common usages. At the bottom of the list of settings you will see **CLC Standard Settings** which represent the way the program was set up, when you first launched it. ( )

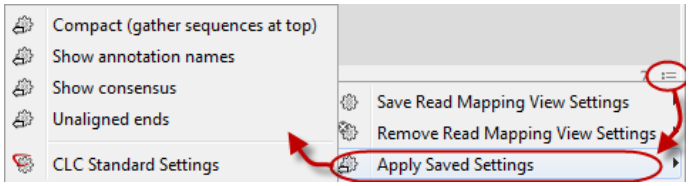


Figure 4.13: *Applying saved settings.*

The settings are specific to the type of view. Hence, when you save settings of a circular view, they will not be available if you open the sequence in a linear view.

If you wish to export the settings that you have saved, this can be done in the **Preferences** dialog under the **View** tab (see section 4.2.1).

Chapter 5

Printing

Contents

CLC Drug Discovery Workbench offers different choices of printing the result of your work.

This chapter deals with printing directly from *CLC Drug Discovery Workbench*. Another option for using the graphical output of your work, is to export graphics (see chapter 6.4) in a graphic format, and then import it into a document or a presentation.

All the kinds of data that you can view in the **View Area** can be printed. The *CLC Drug Discovery Workbench* uses a WYSIWYG principle: What You See Is What You Get. This means that you should use the options in the Side Panel to change how your data, e.g. a sequence, looks on the screen. When you print it, it will look exactly the same way on print as on the screen.

For some of the views, the layout will be slightly changed in order to be printer-friendly.

It is not possible to print elements directly from the **Navigation Area**. They must first be opened in a view in order to be printed. To print the contents of a view:

select relevant view | Print () in the toolbar

This will show a print dialog (see figure 5.1).

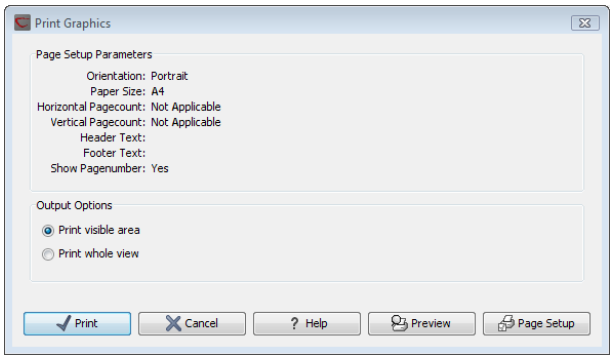


Figure 5.1: *The Print dialog.*

In this dialog, you can:

- Select which part of the view you want to print.
- Adjust **Page Setup**.
- See a print **Preview** window.

These three options are described in the three following sections.

5.1 Selecting which part of the view to print

In the print dialog you can choose to:

- **Print visible area**, or
- **Print whole view**

These options are available for all views that can be zoomed in and out. In figure 5.2 is a view of a circular sequence which is zoomed in so that you can only see a part of it.

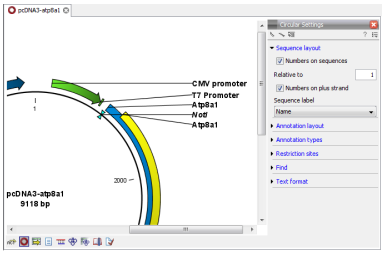


Figure 5.2: *A circular sequence as it looks on the screen.*

When selecting **Print visible area**, your print will reflect the part of the sequence that is *visible* in the view. The result from printing the view from figure 5.2 and choosing **Print visible area** can be seen in figure 5.3.

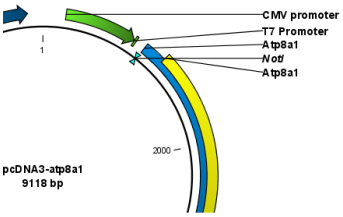


Figure 5.3: *A print of the sequence selecting Print visible area.*

On the other hand, if you select **Print whole view**, you will get a result that looks like figure 5.4. This means that you also print the part of the sequence which is not visible when you have zoomed in.

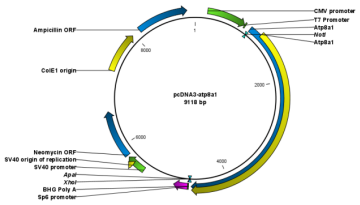


Figure 5.4: *A print of the sequence selecting Print whole view. The whole sequence is shown, even though the view is zoomed in on a part of the sequence.*

5.2 Page setup

No matter whether you have chosen to print the visible area or the whole view, you can adjust page setup of the print. An example of this can be seen in figure 5.5

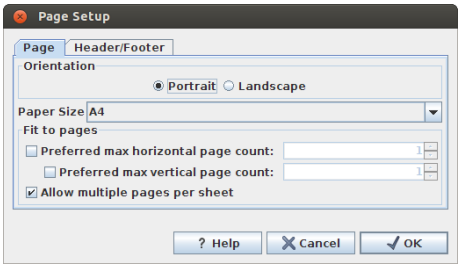


Figure 5.5: *Page Setup.*

In this dialog you can adjust both the setup of the pages and specify a header and a footer by clicking the tab at the top of the dialog.

You can modify the layout of the page using the following options:

- **Orientation**.
  - **Portrait**. Will print with the paper oriented vertically.
  - **Landscape**. Will print with the paper oriented horizontally.
- **Paper size**. Adjust the size to match the paper in your printer.
- **Fit to pages**. Can be used to control how the graphics should be split across pages (see figure 5.6 for an example).
  - **Horizontal pages**. If you set the value to e.g. 2, the printed content will be broken up horizontally and split across 2 pages. This is useful for sequences that are not wrapped
  - **Vertical pages**. If you set the value to e.g. 2, the printed content will be broken up vertically and split across 2 pages.

Note! It is a good idea to consider adjusting view settings (e.g. **Wrap** for sequences), in the **Side Panel** before printing. As explained in the beginning of this chapter, the printed material will look like the view on the screen, and therefore these settings should also be considered when adjusting **Page Setup**.

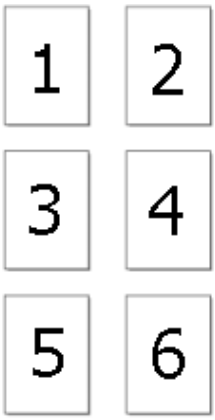


Figure 5.6: *An example where Fit to pages horizontally is set to 2, and Fit to pages vertically is set to 3.*

Header and footer Click the **Header/Footer** tab to edit the header and footer text. By clicking in the text field for either **Custom header text** or **Custom footer text** you can access the auto formats for header/footer text in **Insert a caret position**. Click either **Date**, **View name**, or **User name** to include the auto format in the header/footer text.

Click **OK** when you have adjusted the **Page Setup**. The settings are saved so that you do not have to adjust them again next time you print. You can also change the **Page Setup** from the **File** menu.

5.3 Print preview

The preview is shown in figure 5.7.

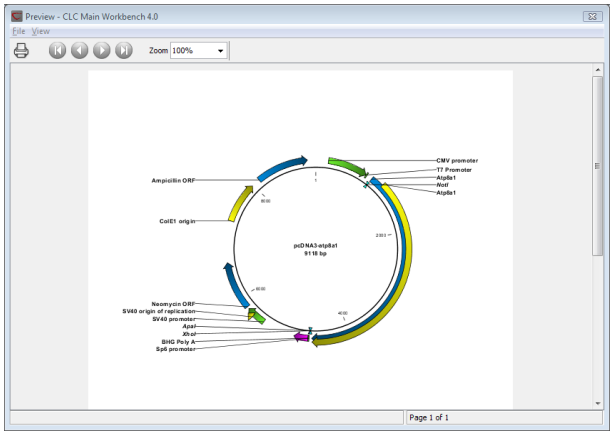


Figure 5.7: *Print preview.*

The **Print preview** window lets you see the layout of the pages that are printed. Use the arrows in the toolbar to navigate between the pages. Click Print ( ) to show the print dialog, which lets you choose e.g. which pages to print.

The **Print preview** window is for preview only - the layout of the pages must be adjusted in the **Page setup**.

Chapter 6

Import/export of data and graphics

Contents

CLC Drug Discovery Workbench handles a large number of different data formats. In order to work with data in the Workbench, it has to be imported ( ). Data types that are not recognized by the Workbench are imported as "external files" which means that when you open these, they will open in the default application for that file type on your computer (e.g. Word documents will open in Word).

This chapter first deals with importing and exporting data in bioinformatic data formats and as external files. Next comes an explanation of how to export graph data points to a file, and how to export graphics.

6.1 Standard import

CLC Drug Discovery Workbench has support for a wide range of bioinformatic data such as molecules, sequences, alignments etc. See a full list of the data formats in section F.1.

These data can be imported through the Import dialog, using drag/drop or copy/paste as explained below.

6.1.1 Import using the import dialog

To start the import using the import dialog: **click Import ( ) in the Toolbar**

click Import ( ) in the Toolbar | Standard Import

This will show a dialog similar to figure 6.1. You can change which kind of file types that should be shown by selecting a file format in the **Files of type** box.

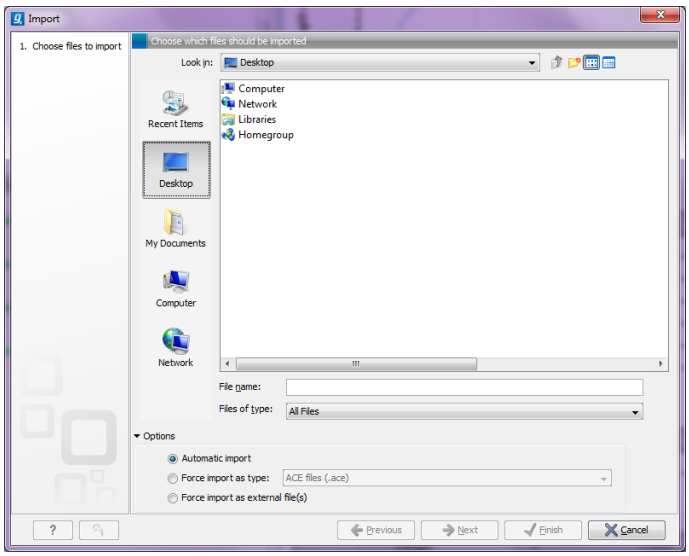


Figure 6.1: *The import dialog.*

Next, select one or more files or folders to import and click **Next**.

This allows you to select a place for saving the result files.

If you import one or more folders, the contents of the folder is automatically imported and placed in that folder in the **Navigation Area**. If the folder contains subfolders, the whole folder structure

is imported.

In the import dialog (figure 6.1), there are three import options:

Automatic import This will import the file and *CLC Drug Discovery Workbench* will try to determine the format of the file. The format is determined based on the file extension (e.g. SwissProt files have .swp at the end of the file name) in combination with a detection of elements in the file that are specific to the individual file formats. If the file type is not recognized, it will be imported as an external file. In most cases, automatic import will yield a successful result, but if the import goes wrong, the next option can be helpful:

Force import as type This option should be used if *CLC Drug Discovery Workbench* cannot successfully determine the file format. By forcing the import as a specific type, the automatic determination of the file format is bypassed, and the file is imported as the type specified.

Force import as external file This option should be used if a file is imported as a bioinformatics file when it should just have been external file. It could be an ordinary text file which is imported as a sequence.

6.1.2 Import using drag and drop

It is also possible to drag a file from e.g. the desktop into the **Navigation Area** of *CLC Drug Discovery Workbench*. This is equivalent to importing the file using the **Automatic import** option described above. If the file type is not recognized, it will be imported as an external file.

6.1.3 Import using copy/paste of text

If you have e.g. a text file or a browser displaying a sequence in one of the formats that can be imported by *CLC Drug Discovery Workbench*, there is a very easy way to get this sequence into the **Navigation Area**:

Copy the text from the text file or browser | Select a folder in the Navigation Area | Paste ()

This will create a new sequence based on the text copied. This operation is equivalent to saving the text in a text file and importing it into the *CLC Drug Discovery Workbench*.

If the sequence is not formatted, i.e. if you just have a text like this: "ATGACGAATAGGAGTTC-TAGCTA" you can also paste this into the **Navigation Area**.

Note! Make sure you copy all the relevant text - otherwise *CLC Drug Discovery Workbench* might not be able to interpret the text.

6.1.4 External files

In order to help you organize your research projects, *CLC Drug Discovery Workbench* lets you import all kinds of files. E.g. if you have Word, Excel or pdf-files related to your project, you can import them into the **Navigation Area** of *CLC Drug Discovery Workbench*. Importing an external file creates a copy of the file which is stored at the location you have chosen for import. The file can now be opened by double-clicking the file in the **Navigation Area**. The file is opened using

the default application for this file type (e.g. Microsoft Word for .doc-files and Adobe Reader for .pdf).

External files are imported and exported in the same way as bioinformatics files (see section 6.1). Bioinformatics files not recognized by *CLC Drug Discovery Workbench* are also treated as external files.

6.2 Import molecules

The supported file formats for importing molecular structures are:

- Mol2 (http://www.tripos.com/)
- Structure-data file (SDF) [Dalby et al., 1992]
- Protein Data Bank (PDB) (http://www.wwpdb.org/documentation/format33/v3.3.html)

Upon import, the imported data are converted to a **CLC Molecule Project** or **CLC Molecule Table**.

Molecule Projects ( ) are used to work with a limited number of molecules in a 3D view, and is used to setup binding sites on proteins to use for docking and visualization of molecule interactions. **Molecule Tables** ( ) are used to work with small molecules in a table view and can contain an unlimited number of molecules.

All importers assign basic chemical properties to the imported structures. This includes determining connectivity, bond order, assigning atom hybridization, and creating explicit hydrogens. For standard residues in proteins and nucleic acids, these properties are assigned based on a set of templates. Additionally, for proteins, secondary structure information can be read from PDB files if present - otherwise secondary structure is assigned using a built-in algorithm.

For small molecules, the following approach is used:

- Connectivity (covalent bonding) is based on any explicit bond information in the input file, but will also be automatically created for atoms sufficiently close to each other. The PDB importer will recognize and import covalently bound molecules as one single molecule.
- Assignment of atom hybridization is based on the geometry of the atom neighborhood (including any explicit hydrogens present). If Sybyl atom types are present in the input file, the importer will use the hybridization from these as a starting point.
- Bond order information may be present in the input file, but for file formats such as PDB, where bond order is not represented, the importers will assign bond orders based on atom hybridization, atom distances, and electronegativity. Notice, that even for file formats with explicit bond order information, the importers may change bond orders to better represent aromatic and delocalized systems.
- If no hydrogens are found on a molecule after import, explicit hydrogens will be created. Since some PDB files only contain hydrogen atoms for polar atoms, input molecules from PDB files are always checked for missing hydrogens.

- Partial atom charges are read from the input file if present. Otherwise charges are assigned according to a set of templates recognizing common chemical motifs. Notice, that charges are only used for visualization purposes - the force field used for molecular docking does not consider atom charges.

You can import molecule structures in seven different ways:

1. Using the standard importer (see section 6.2.1)
2. Using the Import Molecules with 3D Coordinates importer (see section 6.2.2)
3. Using the Add molecules to **Molecule Project** importer (see section 6.2.3)
4. From the Protein Data Bank (PDB) (see section 6.2.4)
5. Using BLAST search against the PDB database (see section 6.2.5)
6. Using the Import Molecules from SMILES or 2D importer (see section 6.2.6)
7. Copy-paste of SMILES strings (see section 6.2.7)

6.2.1 Using the standard importer

The standard importer can be used to import files from your computer file system into the workbench. The Standard Import function will import PDB files to a **Molecule Project** and Mol2 and SDF files to **Molecule Tables**. The import function is found in the toolbar

Toolbar | **Import** () | **Standard Import** ()

In the Import wizard, select the structure(s) of interest from a data location and tick "Automatic import" (figure 6.2).

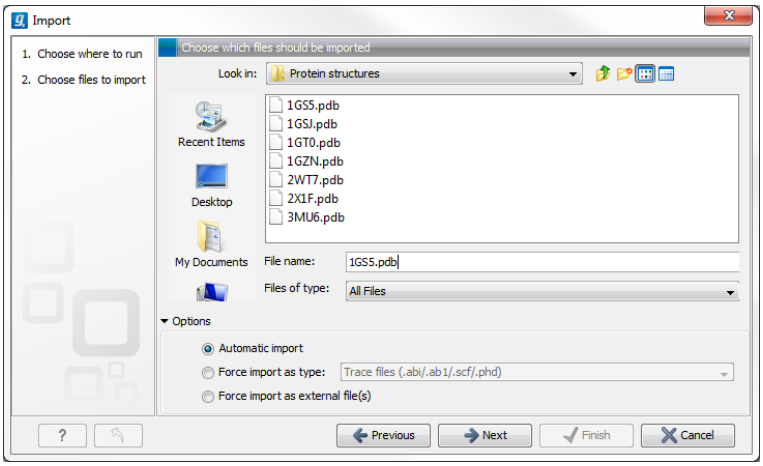


Figure 6.2: *A PDB file can be imported using the "Standard Import" function.*

Specify where to save the imported files and click Finish. Another option is to drag the files into the **Navigation Area**. This will automatically invoke the standard import of the files.

6.2.2 Using Import Molecules with 3D Coordinates

The Import Molecules with 3D Coordinates option is found in the toolbar

Toolbar | **Import** ( )| **Import Molecules with 3D Coordinates** ( )

The importer has three options (see figure 6.3).

- **Output settings:** Using this importer, you are free to choose if the files should be imported to a Molecule Project or a Molecule Table.
- **Filter settings:** If an import issue (section 6.2.9) is encountered for a molecule during import, there is an option to exclude this molecule from the imported Molecule Table or Molecule Project.
- **Subset selection:** For Mol2 and SDF files, which can contain a huge number of molecules, you can choose to only import a subset of the molecules.

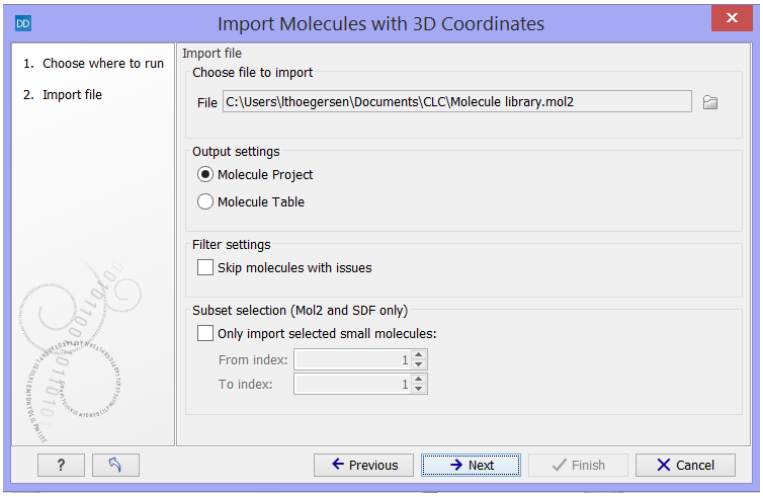


Figure 6.3: *Options when importing molecules with 3D coordinates.*

6.2.3 Add Molecules to Molecule Project

The Add Molecules to Molecule Project option is found in the toolbar

Toolbar | **Import** ( )| **Add Molecules to Molecule Project** ( )

This import option supports all three molecule structure formats and can be used to import molecules with 3D coordinates directly into an already existing **Molecule Project**.

6.2.4 From the Protein Data Bank

Molecule structures can be imported in the workbench from the Protein Data Bank using the "Download" function:

Toolbar | **Download** ( ) | **Search for PDB structures at NCBI** ( )

Type the molecule name or accession number into the search field and click on the "Start search" button (as shown in figure 6.4). The search hits will appear in the table below the search field.

Select the molecule structure of interest and click on the button labeled "Download and Open" (see figure 6.4) or double click on the relevant row in the table to open the protein structure.

Pressing the "Download and Save" button will save the molecule structure at a user defined destination in the **Navigation Area**.

The button "Open at NCBI" links directly to the structure summary page at NCBI. Clicking this button will open individual NCBI pages describing each of the selected molecule structures.

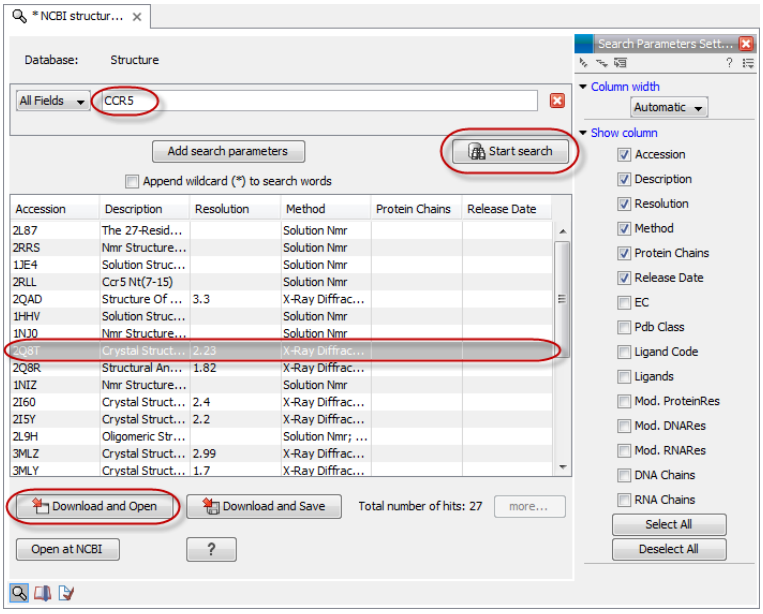


Figure 6.4: *Download protein structure from the Protein Data Bank. It is possible to open a structure file directly from the output of the search by clicking the "Download and Open" button or by double clicking directly on the relevant row.*

6.2.5 BLAST search against the PDB database

It is also possible to make a BLAST search against the PDB database, by going to:

Toolbox | **Sequence Analysis (**  **)** | **BLAST (**  **)**| **BLAST at NCBI (**  **)**

After selecting where to run the analysis, specify which input sequences to use for the BLAST search in the "BLAST at NCBI" dialog, within the box named "Select sequences of same type". More than one sequence can be selected at the same time, as long as the sequences are of the same type (figure 6.5).

Click **Next** and choose program and database (figure 6.6). When a protein sequence has been used as input, select "Program: blastp: Protein sequence and database" and "Database: Protein Data Bank proteins (pdb)".

It is also possible to use mRNA and genomic sequences as input. In such cases the program "blastx: Translated DNA sequence and protein database" should be used.

Please refer to section 12.1.1 for further description of the individual parameters in the wizard steps.

When you click on the button labeled **Finish**, a BLAST output is generated that shows local sequence alignments between your input sequence and a list of matching proteins with known

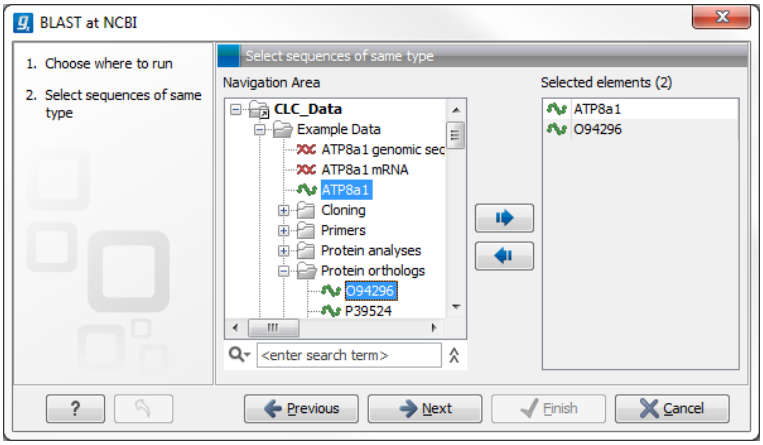


Figure 6.5: *Select the input sequence of interest. In this example a protein sequence for ATPase class I type 8A member 1 and an ATPase ortholog from S. pombe have been selected.*

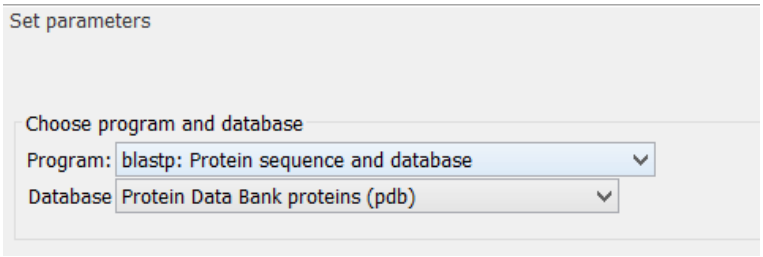


Figure 6.6: *Select database and program.*

structures available.

Note! The BLAST at NCBI search can take up to several minutes, especially when mRNA and genomic sequences are used as input.

Switch to the "BLAST Table" editor view to select the desired entry (figure 6.7). If you have performed a multi BLAST, to get access to the "BLAST Table" view, you must first double click on each row to open the entries individually.

In this view four different options are available:

- **Download and Open** The sequence that has been selected in the table is downloaded and opened in the **View Area**.
- **Download and Save** The sequence that has been selected in the table is downloaded and saved in the **Navigation Area**.
- **Open at NCBI** The protein sequence that has been selected in the table is opened at NCBI.
- **Open Structure** Opens the selected structure in a **Molecule Project** in the **View Area**.

6.2.6 Import Molecules from SMILES or 2D

This import option is found in the toolbar

Toolbar | **Import** ( ) | **Import Molecules from SMILES or 2D** ( )

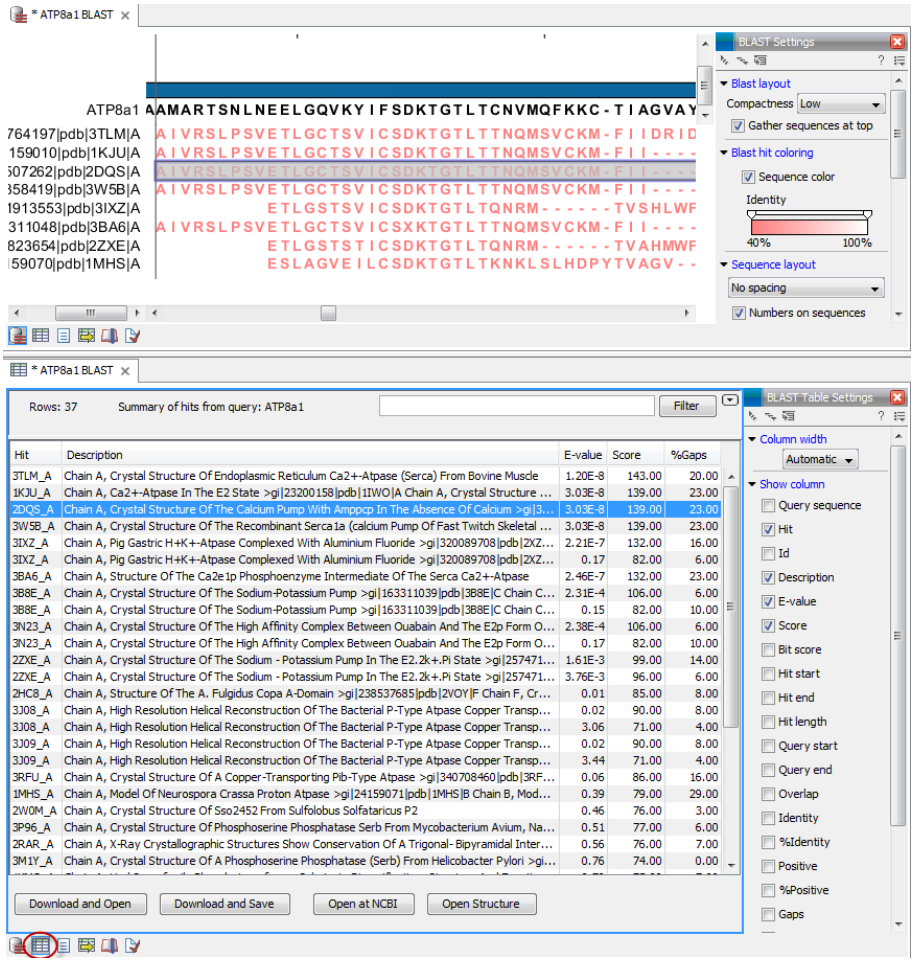


Figure 6.7: *Top: The output from "BLAST at NCBI". Bottom: The "BLAST table". One of the protein sequences has been selected. This activates the four buttons under the table. Note that the table and the BLAST Graphics are linked, this means that when a sequence is selected in the table, the same sequence will be highlighted in the BLAST Graphics view.*

Using this importer you can import a molecule library where the molecules are given as SMILES strings (.smi file) or where molecule representation is only given in 2D in SDF or Mol2 format. 3D coordinates are generated for the molecules on import. The freely available program Balloon [Vainio and Johnson, 2007] is used for 3D structure generation (see 6.2.8).

Prerequisites

The program Balloon should be downloaded from the website http://users.abo.fi/mivainio/balloon/download.php, where it is available for Windows, Mac and a variety of Linux platforms.

In the Preferences settings, found in the Edit menu, the path to the Balloon executable should be specified, for this import option to work (figure 6.8).

Note that the file MMFF94.mff, which is downloaded together with the executable, should be kept in the same folder as the executable.

The importer has three options (see figure 6.9).

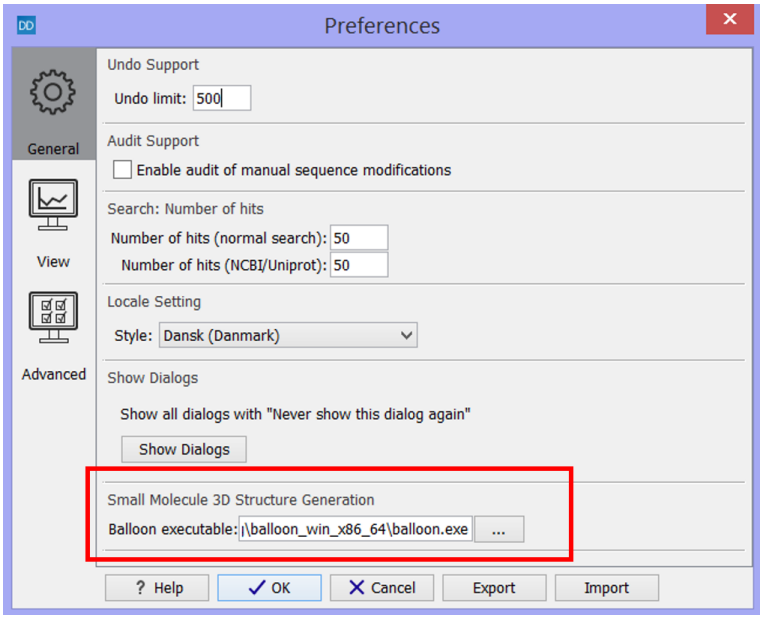


Figure 6.8: *The Preferences settings found in the Edit menu. The Balloon executable specification is seen in the bottom.*

- **3D structure generator settings:**

  Molecule Name

  In SDF files, the first line for each molecule is the molecule name. In Mol2 files, the second line for each molecule is the molecule name. If the molecules do not have a name specified in the input file, the importer will assign a user specified name. The molecule name then consists of the prefix, specified in the import wizard, and a number.

  Molecule Conformations

  The importer can return either the conformation found with lowest energy for each molecule, or an ensemble of low-energy conformations for each molecule (usually one to ten). Especially if the molecules have non-aromatic ring structures, it can be relevant to generate multiple representations, as the rings are kept rigid during docking.

- **Filter settings:** If an import issue (section 6.2.9) is encountered for a molecule during import, there is an option to exclude this molecule from the imported Molecule Table.

- **Subset selection:** You can choose only to import a specific subset of the molecules from the input file.

6.2.7 Copy-paste of SMILES strings

SMILES strings describe how atoms in molecules are connected as well as the stereochemistry of chiral centers. SMILES strings can be copied and pasted directly into a **Molecule Project** using **Edit | Paste** from the right-click context menu in the 3D view area or Project Tree, or simply using Ctrl-V (Cmd-V on Mac). A dialog box then appears where a molecule name prefix and whether to generate one or more conformations per molecule can be specified (figure 6.10). 3D coordinates are then generated for the molecules on import, and the molecules are added to the Ligands category in the Project Tree.

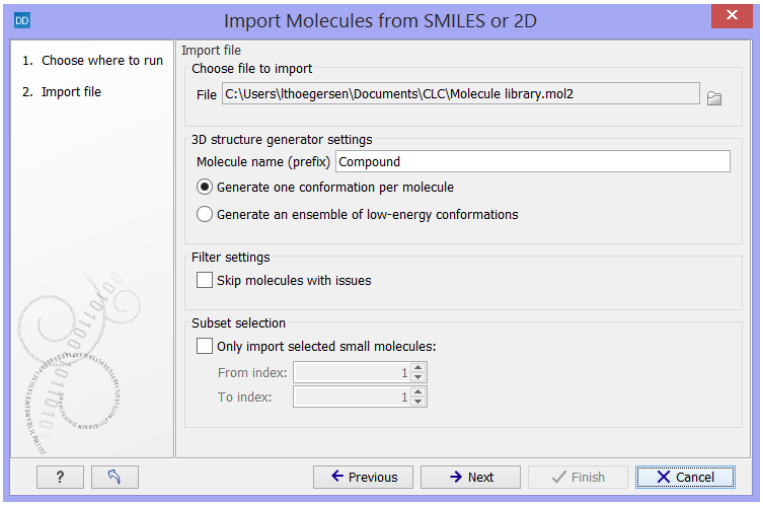


Figure 6.9: *Options when importing molecules with coordinates in 2D, or from SMILES.*

The freely available program Balloon [Vainio and Johnson, 2007] is used for the 3D structure generation (see 6.2.8).

Prerequisites

The program Balloon should be downloaded from the website http://users.abo.fi/mivainio/balloon/download.php, where it is available for Windows, Mac and a variety of Linux platforms.

In the Preferences settings, found in the Edit menu, the path to the Balloon executable should be specified, for the copy-paste option to work (figure 6.8).

Note that the file MMFF94.mff, which is downloaded together with the executable, should be kept in the same folder as the executable.

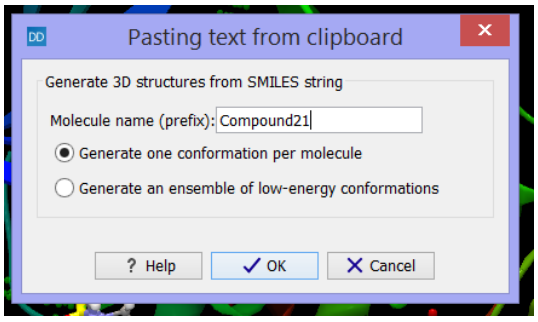


Figure 6.10: *Dialog box appearing when pasting a SMILES string into a Molecule Project.*

In 2D molecule sketching programs such as ChemDraw and MarvinSketch, a molecule sketch can be selected and the *Copy as SMILES* option used, to paste the sketched molecule into the workbench. Molecules should be copied one at a time when it is from 2D sketchers, while several SMILES strings can be copy-pasted simultaneously, if they are copied from a text file list. In the case of a list with SMILES strings, the second column can be used to specify the molecule name.

See also the tutorial *Dock Ligands from a 2D Molecule Sketch* for more details.

6.2.8 Generation of 3D structure on import

This is a description of how the Balloon algorithm is used in *CLC Drug Discovery Workbench*, to generate 3D molecule structures on import. The algorithm is described in detail in [Vainio and Johnson, 2007].

Input: A SMILES string (or coordinates in 2D) describing the topology of a molecule.

Step 1: Generation of template structure with 3D coordinates. Minimum and maximum interatomic distances are set for all atom pairs based on the input. A special type of bounds are used to specify the chirality of stereochemical centers, if they are given in the input. The sum of the violations to the bounds is minimized, to get an initial structure in 3D. The structure is then refined, making a minimization of the conformational energy as calculated by the MMFF94 force field [Halgren, 1996].

Step 2: Generation of a conformer ensemble. Different conformations are generated by rotation about rotatable bonds, changes to stereochemistry of double bonds and tetrahedral chiral centers (for those not specified on input), and changes to ring conformations. A genetic algorithm is used to generate variations to the structure. A particular molecule conformation (phenotype) is defined by the values at the 'loci' (the genotype). For example, each rotatable bond has a locus specifying the rotation value, and each chiral center, not defined on input, has a locus specifying whether or not to invert the chirality compared to the template structure.

The genetic algorithm runs in 20 generations with a population of five individuals (diverse conformations). The fitness of an individual is evaluated based on both the torsional and van der Waals terms of the MMFF94 potential energy function [Halgren, 1996].

The first generation is constructed from random 'mutations' to the template structure. The steps 1-4 below are then repeated for each generation, to search through relevant conformers, and produce a diverse set with low energy.

1. Parents are selected at random between the five individuals in the population, with a bias towards the best fit and towards promoting geometric diversity.
2. The parent's genotypes are combined (via random crossovers) to produce five offspring.
3. The offspring is then exposed to mutations, making small random changes to individual loci.
4. The five offspring together with the five individuals from the parent generation are evaluated, and five individuals are selected for the next generation based on their fitness and geometric diversity.

Step 3: Post-processing. Strain introduced into the structures is relaxes using an MMFF94 force field, where the electrostatic term and the torsional term for rotatable bonds are left out. Conformational duplicates (RMSD < 0.5 Å) are removed, and so are structures whose strain energy remains above a predefined window from the minimum energy value found in the set. For a molecule with no rotatable bonds, the energy window is 5 kcal/mol. The energy window is increased by 0.25 kcal/mol for each rotatable bond present in the molecule.

Output: The molecule conformer with the lowest energy found *or* the final post-processed ensemble of low-energy conformers.

6.2.9 Import issues

When opening an imported molecule file for the first time, a notification is briefly shown in the lower left corner of the **Molecule Project** or **Molecule Table** editor, with information of the number of issues encountered during import of the file. The issues are categorized and listed in a table view in the Issues view. The Issues list can be opened by selecting **Show | Issues** from the menu appearing when right-clicking in an empty space in the 3D view (figure 6.11) or on a row in a **Molecule Table**.

Alternatively, the issues can be accessed from the lower left corner of the view, where buttons are shown for each available view. If you hold down the Ctrl key (Cmd on Mac) while clicking on the Issues icon ( ), the list will be shown in a split view together with the 3D or table view. The issues list is linked with the molecules in the 3D or table view, such that selecting an entry in the list will select the implicated atoms in the view, and zoom to put them into the center of the 3D view.

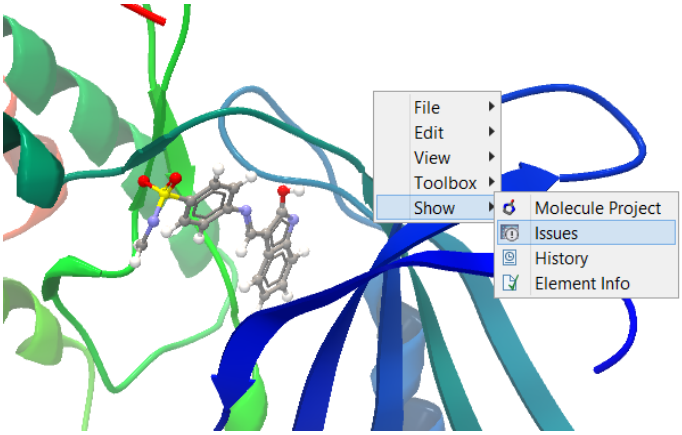


Figure 6.11: *At the bottom of the Molecule Project it is possible to switch to the "Show Issues" view by clicking on the "table-with-exclamation-mark" icon.*

6.3 Data export

The exporter can be used to:

- Export bioinformatic data in most of the formats that can be imported. There are a few exceptions (see section F.1).
- Export one or more data elements at a time to a given format. When multiple data elements are selected, each is written out to an individual file, unless compression is turned on, or "Output as single file" is selected.

The standard export functionality can be launched using the Export button on the toolbar, or by going to the menu:

File | **Export ( )**

An additional export tool is available from under the File menu:

File | **Export with Dependent Elements**

This tool is described further in section 6.3.2.

The general steps when configuring a standard export job are:

- (Optional) Select the data to export in the **Navigation Area**.
- Start up the exporter tool via the Export button in the toolbar or using the **Export** option under the File menu.
- Select the format the data should be exported to.
- Select the data to export, or confirm the data to export if it was already selected via the **Navigation Area**.
- Configure the parameters. This includes compression, multiple or single outputs, and naming of the output files, along with other format-specific settings where relevant.
- Select where the data should be exported to.
- Click on the button labeled **Finish**.

Selecting data for export - part I. You can select the data elements to export **before** you run the export tool **or after** the format to export to has been selected. If you are not certain which formats are supported for the data being exported, we recommend selecting the data in the **Navigation Area** before launching the export tool.

Selecting a format to export to. When data is pre-selected in the **Navigation Area** before launching the export tool you will see a column in the export interface called **Supported formats**. Formats that the selected data elements can be exported to are indicated by a "Yes" in this column. Supported formats will appear at the top of the list of formats (figure 6.12).

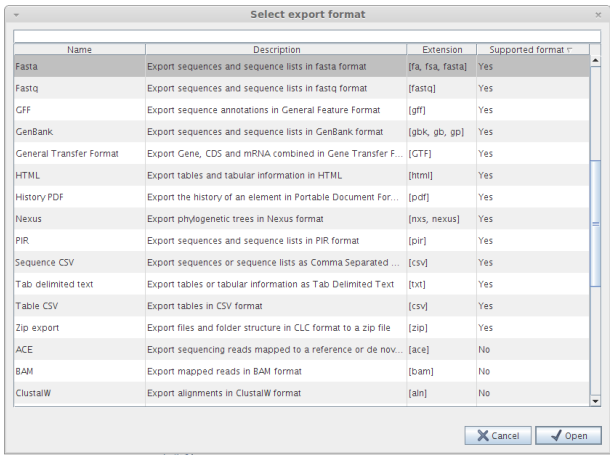


Figure 6.12: *The Select exporter dialog where sequence lists were pre-selected in the Navigation Area before launching the export tool. Here, the formats sequence lists can be exported to are listed at the top, with a Yes in the Selected formats column. Other formats are found below, with No in this column.*

Formats that cannot be used for export of the selected data have a "No" listed in the **Supported formats** column. If you have selected multiple data elements of different types, then formats which can be used for some of the selected data elements but not all of them are indicated by the text "For some elements" in this column.

Please note that the information in the **Supported formats** column only refers to the data already selected in the **Navigation Area**. If you are going to choose your data later in the export process, then the information in this column will not be pertinent.

Only one export format is available if you select a folder to be exported. This is described in more detail in section 6.3.1.

Finding a particular format in the list. You can quickly find a particular format by using the text box at the top of the exporter window as shown in figure 6.13, where formats that include the term VCF are searched for. This search term will remain in place the next time the Export tool is launched. Just delete the text from the search box if you no longer wish only the formats with that term to be listed.

When the desired export format has been identified, click on the button labeled **Open**.

Selecting data for export - part II. A dialog appears, with a name reflecting the format you have chosen. For example if the "Variant Call Format" (VCF format) was selected, the window is labeled "Export VCF".

If you are logged into a CLC Server, you will be asked whether to run the export job using the Workbench or the Server. After this, you are provided with the opportunity to select or de-select data to be exported.

In figure 6.14 we show the selection of a variant track for export to VCF format.

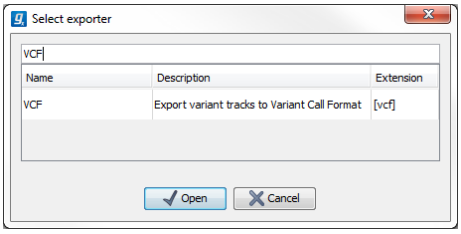


Figure 6.13: *The text field has been used to search for VCF format in the Select exporter dialog.*

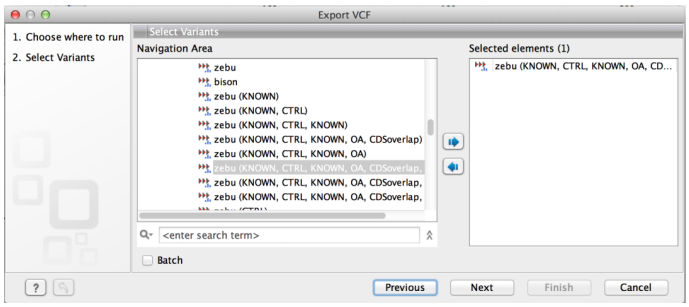


Figure 6.14: *The Select exporter dialog. Select the data element(s) to export.*

The parameters under **Basic export parameters** and **File name** are offered when exporting to any format.

There may be additional parameters for particular export formats. This is illustrated here with the VCF exporter, where a reference sequence track must be selected (see figure 6.15).

Compression options. Within the **Basic export parameters** section, you can choose to compress the exported files. The options are no compression (None), gzip or zip format. Choosing zip format results in all data files being compressed into a single file. Choosing gzip compresses the exported file for each data element individually.

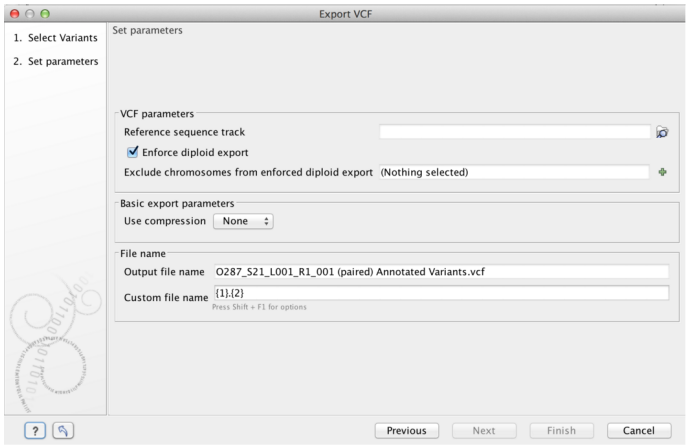


Figure 6.15: *Set the export parameters. When exporting in VCF format, a reference sequence track must be selected.*

Exporting multiple files. If you have selected multiple files of the same type, you can choose to export them in one single file (only for certain file formats) by selecting "Output as single file" in the **Basic export parameters** section. If you wish to keep the files separate after export, make sure this box is not ticked. **Note:** Exporting in zip format will export only one zipped file, but the files will be separated again when unzipped.

Choosing the exported file name(s) The default setting for the **File name** is to use the original data element name as the basename and the export format as the suffix.

When exporting just one data element, or exporting to a zip file, the desired filename could just be typed in the Custom file name box.

When working with the export of multiple files, using some combination of the terms shown by default in this field and in figure 6.17 are recommended. Clicking in the **Custome file name** field with the mouse and then simultaneously pressing the Shift + F1 keys bring up a list of the available terms that can be included in this field. You can see that "{1}" is the name of the input element and "{2}" is the file name extension. It is possible to change the input file name and the file extension name. We will look at an example to illustrate this:

In this example we would like to change the export file format to .fasta in a situation where .fa was the default format that would be used if you kept the default file extension suggestion ("{2}"). To do this replace "{2}" with ".fasta" in the "Custom file name field". You can see that when changing "{2}" to ".fasta" , the file name extension in the "Output file name" field automatically changes to the new format (see figure 6.16).

As you add or remove text and terms in the **Custome file name** field, the text in the **Output file name** field will change so you can see what the result of your naming choice will be for your data. When working with multiple files, only the name of the first one is shown. Just move the mouse cursor over the name shown in the **Output file name** field to show a listing of the all the filenames.

The last step is to specify the exported data should be saved (figure 6.18).

A note about decimals and Locale settings. When exporting to CSV and tab delimited files, decimal numbers are formatted according to the Locale setting of the Workbench (see section 4.1). If you open the CSV or tab delimited file with spreadsheet software like Excel, you should make sure that both the Workbench and the spreadsheet software are using the same Locale.

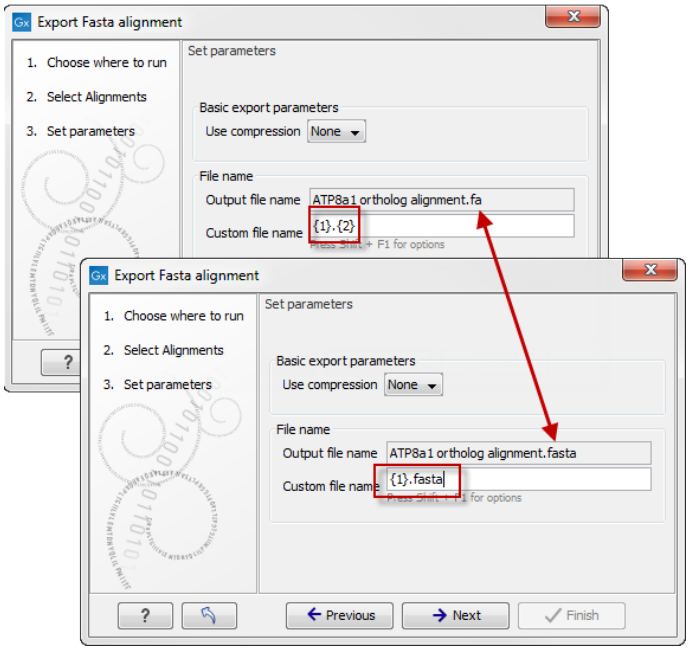


Figure 6.16: *The file name extension can be changed by typing in the preferred file name format.*

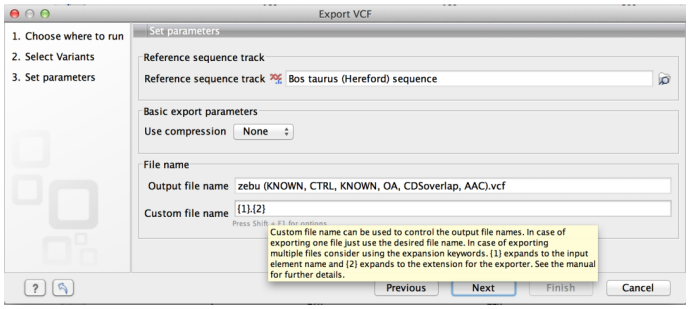


Figure 6.17: *Use the custom file name pattern text field to make custom names.*

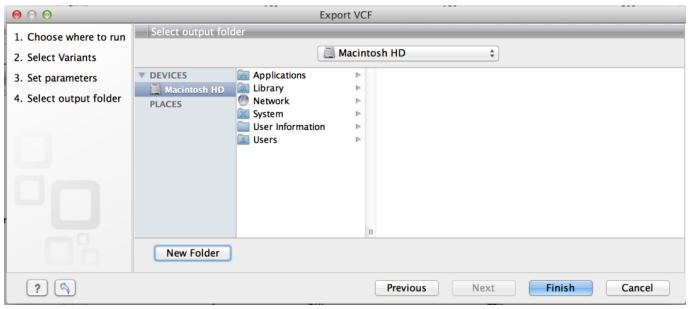


Figure 6.18: *Select where to save the exported data.*

6.3.1 Export of folders and multiple elements in CLC format

In the list of export formats presented is one called zip format. Choosing this format means that you wish to export the selected data element(s) or folders to a single, compressed CLC format file. This is useful in cases where you wish to exchange data between workbenches or as part of a simple backup procedure.

A zip file generated this way can be imported directly into a CLC Workbench using the Standard Import tool and leaving the import type as Automatic.

Note! When exporting multiple files, the names will be listed in the "Output file name" text field with only the first file name being visible and the rest being substituted by "...", but will appear in a tool tip if you hover the mouse over that field (figure 6.19).

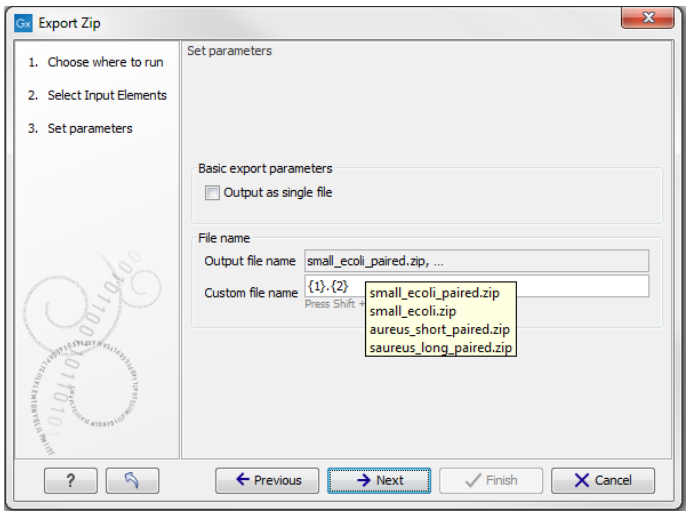


Figure 6.19: *The output file names are listed in the "Output file name" text field.*

6.3.2 Export of dependent elements

Sometimes it can be useful to export the results of an analysis and its dependent elements. That is, the results along with the data that was used in the analysis. For example, one might wish to export an alignment along with all the sequences that were used in generating that alignment.

To export a data element with its dependent elements:

- Select the parent data element (like an alignment) in the **Navigation Area**.
- Start up the exporter tool by going to **File** | **Export with Dependent Elements**.
- Edit the output name if desired and select where the resulting zip format file should be exported to.

The file you export contains compressed CLC format files containing the data element you chose and all its dependent data elements.

A zip file created this way can be imported directly into a CLC workbench by going to

File | **Import** | **Standard Import**

In this case, the import type can be left as Automatic.

6.3.3 The CLC format

The *CLC Drug Discovery Workbench* stores bioinformatic data in CLC format. The CLC format contains data, as well as information about that data like history information and comments you may have added.

A given data element in the Workbench can contain different types of data. This is reflected when exporting data, as the choice of different export formats can lead to the extraction of some

parts of that data object rather than others. The part of the data exported reflects the type of data a given format can support. As a simple example, if you export the results of an alignment to Annotation CSV format, you will get just the annotation information. If you exported to Fasta alignment format, you would get the aligned sequences in fasta format, but no annotations.

The CLC format holds all the information for a given data object. Thus if you plan to share the data with colleagues who also have a CLC Workbench or you are communicating with the CLC Support team and you wish to share the data from within the Workbench, exporting to CLC format is usually the best choice as all information associated with that data object in your Workbench will then be available to the other person who imports that data.

If you are planning to share your data with someone who does not have access to a CLC Workbench, then you will wish to export to another data format. Specifically, one they can use with the software they are working with.

6.3.4 Backing up data from the CLC Workbench

Regular backups of your data are advisable.

The data stored in your CLC Workbench is in the areas defined as CLC Data Locations. Whole data locations can be backed up directly (option 1) or, for smaller amounts of data, you could export the selected data elements to a zip file (option 2).

Option 1: Backing up each CLC Data Location

The easiest way for most people to find out where their data is stored is to put the mouse cursor over the top level directories, that is, the ones that have an icon like (), in the **Navigation Area** of the Workbench. This brings up a tool tip with the system location for that data location.

To back up all your CLC data, please ensure that all your CLC Data Locations are backed up.

Here, if you needed to recover the data later, you could put add the data folder from backup as a data location in your Workbench. If the original data location is not present, then the data should be usable directly. If the original data location is still present, the Workbench will re-index the (new) data location. For large volumes of data, re-indexing can take some time.

Information about your data locations can also be found in an xml file called model_settings_300.xml This file is located in the settings folder in the user home area. Further details about this file and how it pertains to data locations in the Workbench can be found in the Deployment Manual:

http://www.clcsupport.com/workbenchdeployment/current/index.php?manual=Changing_default_location.html

Option 2: Export a folder of data or individual data elements to a CLC zip file

This option is for backing up smaller amounts of data, for example, certain results files or a whole data location, where that location contains smaller amounts of data. For data that takes up many gigabases of space, this method can be used, but it can be very demanding on space, as well as time.

Select the data items, including any folders, in the Navigation area of your Workbench and choose to export by going to:

File | **Export (**  **)**

and choosing ZIP format.

The zip file created will contain all the data you selected. You can later re-import the zip file into the Workbench by going to:

File | **Import** ( )

6.3.5 Export of workflow output

The output from a workflow can be exported by adding one or more workflow export elements (figure 6.20). Multiple elements can be selected by holding down the Ctrl key while clicking on the desired elements.

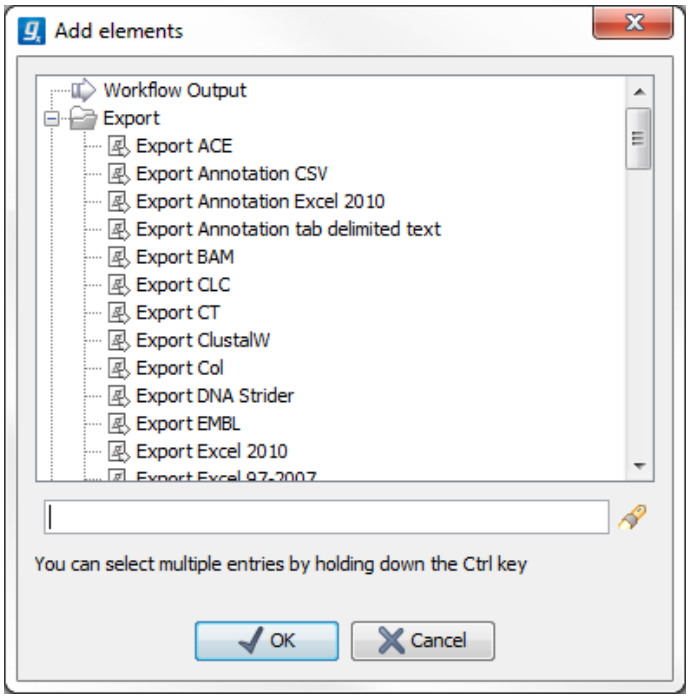


Figure 6.20: *Pressing "Add element" enables addition of workflow export elements.*

When the workflow has been created, you can set the export parameters and the location to export data to by double clicking on each export element (figure 6.21). Leave fields empty and unlocked if you wish users of the Workflow to enter this information when the Workflow is launched.

6.3.6 Export of tables

Tables can be exported in four different formats; CSV, tab-separated, Excel, or html.

When exporting a table in CSV, tab-separated, or Excel format, numbers with many decimals are printed in the exported file with 10 decimals, or in 1.123E-5 format when the number is close to zero.

Excel limits the number of hyperlinks in a worksheet to 66,530. When exporting a table of more than 66,530 rows, Excel will "repair" the file by removing all hyperlinks. If you want to keep the hyperlinks valid, you will need to export your data to several worksheets in batches smaller than 66,530 rows.

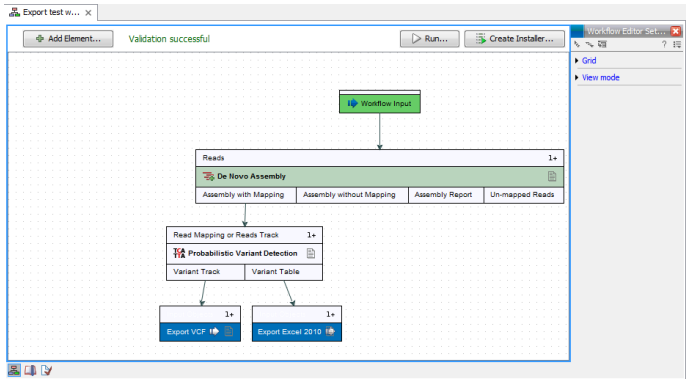


Figure 6.21: *A simple workflow with two export elements. The variant track will be exported in VCF format and the variant table in Excel format.*

When exporting a table in html format, data are exported with the number of decimals that have been defined in the workbench preference settings. When tables are exported in html format from the server or using command line tools, the default number of exported decimals is 3.

The Excel exporters, the CSV and tab delimited exporters, and the HTML exporter have been extended with the ability to export only a sub-set of columns from the object being exported. Uncheck the option "Export all columns" and click next to see a new dialog window in which columns to be exported can be selected (figure 6.22) . The user can choose them one by one or choose a predefined subset:

- All: will select all possible columns.
- None: will clear all preselected column.
- Default: will select the columns preselected by default by the software.
- Last export: will select all windows that were selected during the last export.
- Active editor (only if an active editor is currently open): the columns selected are the same than in the active editor window.

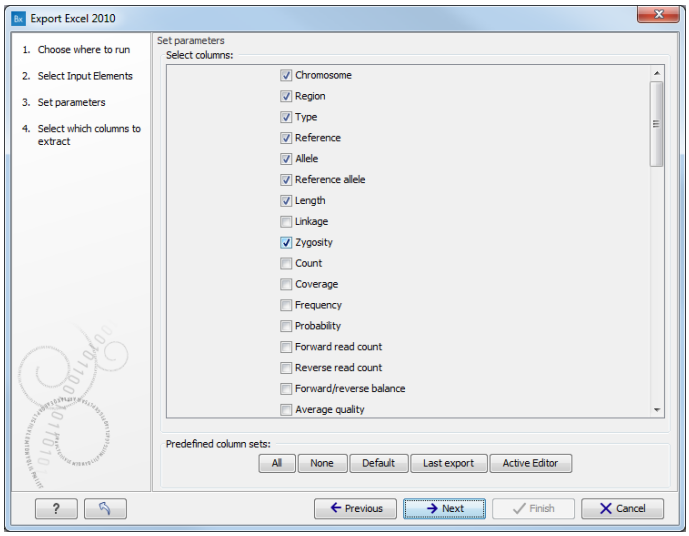


Figure 6.22: *Selecting columns to be exported.*

After selecting columns, the user will be directed to the output destination wizard page.

6.4 Export graphics to files

CLC Drug Discovery Workbench supports export of graphics into a number of formats. This way, the visible output of your work can easily be saved and used in presentations, reports etc. The **Export Graphics** function ( ) is found in the **Toolbar**.

CLC Drug Discovery Workbench uses a WYSIWYG principle for graphics export: What You See Is What You Get. This means that you should use the options in the Side Panel to change how your data, e.g. a sequence, looks in the program. When you export it, the graphics file will look exactly the same way.

It is not possible to export graphics of elements directly from the **Navigation Area**. They must first be opened in a view in order to be exported. To export graphics of the contents of a view:

select tab of View | Graphics ( ) on Toolbar

This will display the dialog shown in figure 6.23.

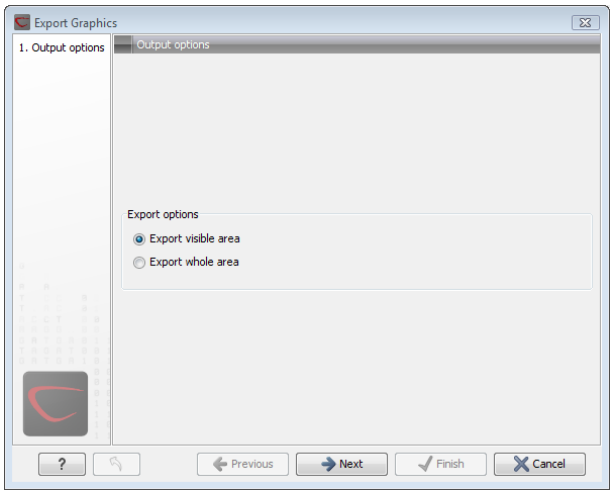


Figure 6.23: *Selecting to export whole view or to export only the visible area.*

6.4.1 Which part of the view to export

In this dialog you can choose to:

- **Export visible area**, or
- **Export whole view**

These options are available for all views that can be zoomed in and out. In figure 6.24 is a view of a circular sequence which is zoomed in so that you can only see a part of it.

When selecting **Export visible area**, the exported file will only contain the part of the sequence that is *visible* in the view. The result from exporting the view from figure 6.24 and choosing **Export visible area** can be seen in figure 6.25.

On the other hand, if you select **Export whole view**, you will get a result that looks like figure 6.26. This means that the graphics file will also include the part of the sequence which is not visible when you have zoomed in.

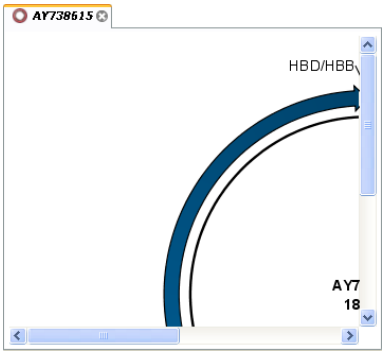


Figure 6.24: *A circular sequence as it looks on the screen when zoomed in.*

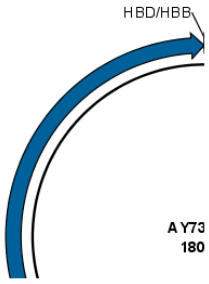


Figure 6.25: *The exported graphics file when selecting Export visible area.*

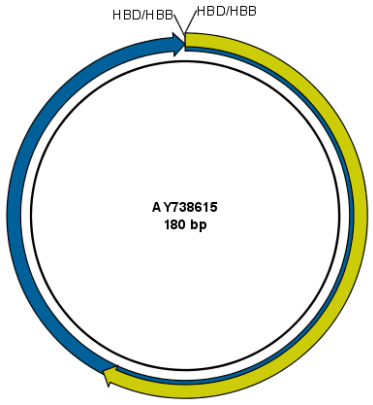


Figure 6.26: *The exported graphics file when selecting Export whole view. The whole sequence is shown, even though the view is zoomed in on a part of the sequence.*

For 3D structures, this first step is omitted and you will always export what is shown in the view (equivalent to selecting **Export visible area**).

Click **Next** when you have chosen which part of the view to export.

6.4.2 Save location and file formats

In this step, you can choose name and save location for the graphics file (see figure 6.27).

CLC Drug Discovery Workbench supports the following file formats for graphics export:

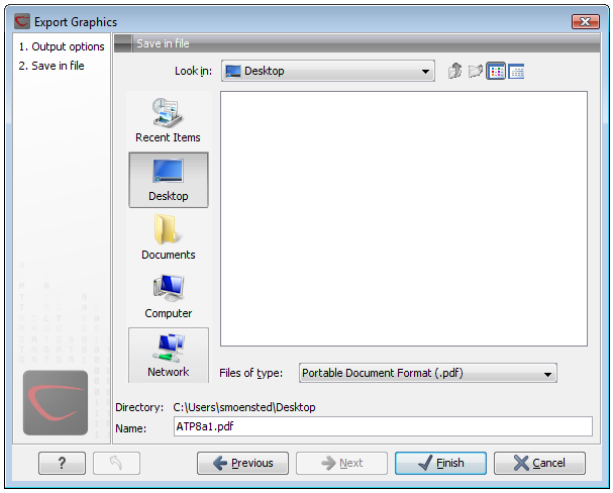


Figure 6.27: *Location and name for the graphics file.*

Format	Suffix	Type
Portable Network Graphics	.png	bitmap
JPEG	.jpg	bitmap
Tagged Image File	.tif	bitmap
PostScript	.ps	vector graphics
Encapsulated PostScript	.eps	vector graphics
Portable Document Format	.pdf	vector graphics
Scalable Vector Graphics	.svg	vector graphics

These formats can be divided into bitmap and vector graphics. The difference between these two categories is described below:

Bitmap images In a bitmap image, each dot in the image has a specified color. This implies, that if you zoom in on the image there will not be enough dots, and if you zoom out there will be too many. In these cases the image viewer has to interpolate the colors to fit what is actually looked at. A bitmap image needs to have a high resolution if you want to zoom in. This format is a good choice for storing images without large shapes (e.g. dot plots). It is also appropriate if you don't have the need for resizing and editing the image after export.

Vector graphics Vector graphic is a collection of shapes. Thus what is stored is e.g. information about where a line starts and ends, and the color of the line and its width. This enables a given viewer to decide how to draw the line, no matter what the zoom factor is, thereby always giving a correct image. This format is good for e.g. graphs and reports, but less usable for e.g. dot plots. If the image is to be resized or edited, vector graphics are by far the best format to store graphics. If you open a vector graphics file in an application like e.g. Adobe Illustrator, you will be able to manipulate the image in great detail.

Graphics files can also be imported into the **Navigation Area**. However, no kinds of graphics files can be displayed in *CLC Drug Discovery Workbench*. See section 6.1.4 for more about importing external files into *CLC Drug Discovery Workbench*.

6.4.3 Graphics export parameters

When you have specified the name and location to save the graphics file, you can either click **Next** or **Finish**. Clicking **Next** allows you to set further parameters for the graphics export, whereas clicking **Finish** will export using the parameters that you have set last time you made a graphics export in that file format (if it is the first time, it will use default parameters).

Parameters for bitmap formats For bitmap files, clicking **Next** will display the dialog shown in figure 6.28.

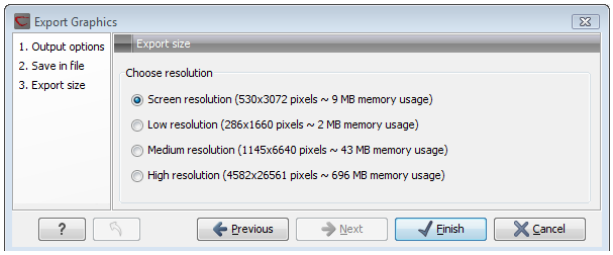


Figure 6.28: *Parameters for bitmap formats: size of the graphics file.*

You can adjust the size (the resolution) of the file to four standard sizes:

- Screen resolution
- Low resolution
- Medium resolution
- High resolution

The actual size in pixels is displayed in parentheses. An estimate of the memory usage for exporting the file is also shown. If the image is to be used on computer screens only, a low resolution is sufficient. If the image is going to be used on printed material, a higher resolution is necessary to produce a good result.

Parameters for vector formats For pdf format, clicking **Next** will display the dialog shown in figure 6.29 (this is only the case if the graphics is using more than one page).

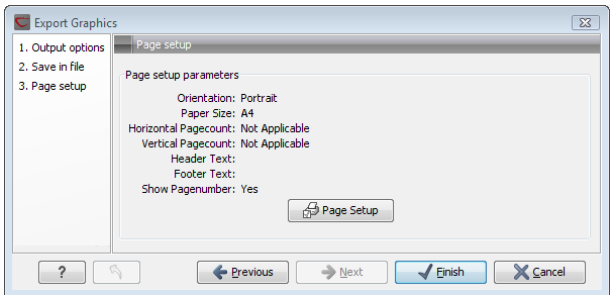


Figure 6.29: *Page setup parameters for vector formats.*

The settings for the page setup are shown, and clicking the **Page Setup** button will display a dialog where these settings can ba adjusted. This dialog is described in section 5.2.

The page setup is only available if you have selected to export the whole view - if you have chosen to export the visible area only, the graphics file will be on one page with no headers or footers.

6.4.4 Exporting protein reports

It is possible to export a protein report using the normal **Export** function () which will generate a pdf file with a table of contents:

Click the report in the Navigation Area | Export () in the Toolbar | select pdf

You can also choose to export a protein report using the **Export graphics** function (), but in this way you will not get the table of contents.

6.5 Export graph data points to a file

Data points for graphs displayed along the sequence or along an alignment or mapping can be exported to a semicolon-separated text file (csv format). An example of such a graph is shown in figure 6.30. This graph shows the coverage of reads in a read mapping.

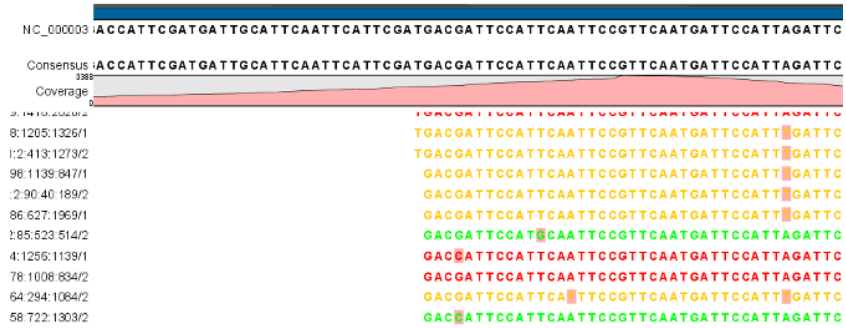


Figure 6.30: *A graph displayed along mapped reads. Right-click the graph to export the data points to a file.*

To export the data points for the graph, right-click the graph and choose **Export Graph to Comma-separated File**. Depending on what kind of graph you have selected, different options will be shown: If the graph is covering a set of aligned sequences with a main sequence, such as read mappings and BLAST results, the dialog shown in figure 6.31 will be displayed. These kinds of graphs are located under **Alignment info** in the Side Panel. In all other cases, a normal file dialog will be shown letting you specify name and location for the file.

In this dialog, select whether you wish to include positions where the main sequence (the reference sequence for read mappings and the query sequence for BLAST results) has gaps. If you are exporting e.g. coverage information from a read mapping, you would probably want to exclude gaps, if you want the positions in the exported file to match the reference (i.e. chromosome) coordinates. If you export including gaps, the data points in the file no longer corresponds to the reference coordinates, because each gap will shift the coordinates.

Clicking **Next** will present a file dialog letting you specify name and location for the file.

The output format of the file is like this:

```
"Position";"Value";
```

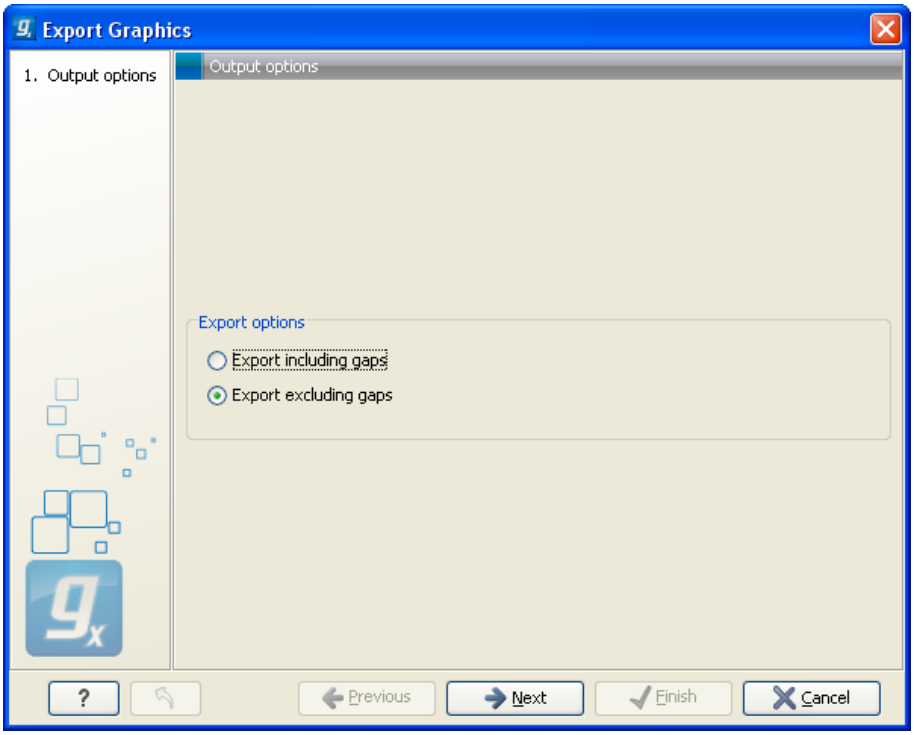


Figure 6.31: *Choosing to include data points with gaps*

```
"1";"13";
"2";"16";
"3";"23";
"4";"17";
...
```

6.6 Copy/paste view output

The content of tables, e.g. in reports, folder lists, and sequence lists can be copy/pasted into different programs, where it can be edited. *CLC Drug Discovery Workbench* pastes the data in tabulator separated format which is useful if you use programs like Microsoft Word and Excel. There is a huge number of programs in which the copy/paste can be applied. For simplicity, we include one example of the copy/paste function from a **Folder Content** view to Microsoft Excel.

First step is to select the desired elements in the view:

> **click a line in the Folder Content view | hold Shift-button | press arrow down/up key**

See figure 6.32.

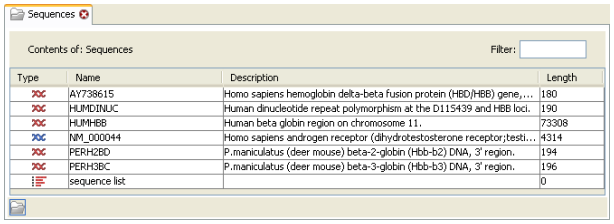


Figure 6.32: *Selected elements in a Folder Content view.*

When the elements are selected, do the following to copy the selected elements:

right-click one of the selected elements | **Edit** | **Copy (** **)**

Then:

right-click in the cell A1 | **Paste (** **)**

The outcome might appear unorganized, but with a few operations the structure of the view in *CLC Drug Discovery Workbench* can be produced. (Except the icons which are replaced by file references in Excel.)

Note that all tables can also be **Exported** ( ) directly in Excel format.

Chapter 7

Running tools, handling results and batching

Contents

This section describes how to run a tool using singles files as input, as well as how to handle and inspect results. We also review how to run tools using the batch mode when the option is enabled.

7.1 Running tools

All the analyses in the **Toolbox** are performed in a step-by-step procedure:

- Data elements to be used in the analysis are selected.
- Any configurations necessary for the tool to run are made.
- The results are opened or saved.

You can open a tool from the Toolbox by double clicking on its name in the Toolbox. In case you do not know which folder the tool you are looking for belongs to, you can use the very useful **Quick launch** function by clicking on Ctrl + Shift + T (or ⌘ + Shift + T on Mac) and typing any part of the tool name in the search field. Double click on the name of the tool in the table.

When you open a tool, a wizard pop up in the center of the View Area. Through a succession of windows you will enter the data you want to analyze, the parameters of the analysis you want

to perform and how you want to handle the results of the analysis. You can navigate between windows by clicking the buttons **Next** and **Previous** at the bottom of the window.

If you are working on a network and have access to a server, you will first be asked to "Choose where to run" the tool. This window gives the following options:

- **Workbench** to run the tool on your own computer.
- **CLC Server** to run the tool on a server.
- **Grid** to be able to choose from the drop down menu.

If you check the option "Remember setting and skip this step", you will not have to enter the information described above for this particular tool. The setting can always be changed xxxx

After having decided where to run the tool, the next window of the wizard is usually asking you to select the input file(s). This window displays on the left a replicate of your Navigation Area in which only the files that have a format adapted to the tool will be shown. The specific input file formats required by a tool are described for each tool independently in the relevant sections of the manual that you can access easily by clicking on the **Help** button((?)). For example, you can see a view of the Navigation Area in the workbench and the same Navigation Area in the wizard in figure 7.1. The Assemble Sequences tool will only accept nucleotide sequences (as such or part of a list), which explains why the file called "Read mapping", or the amino acid sequence ATP8a1 are not being displayed in the wizard Navigation Area.

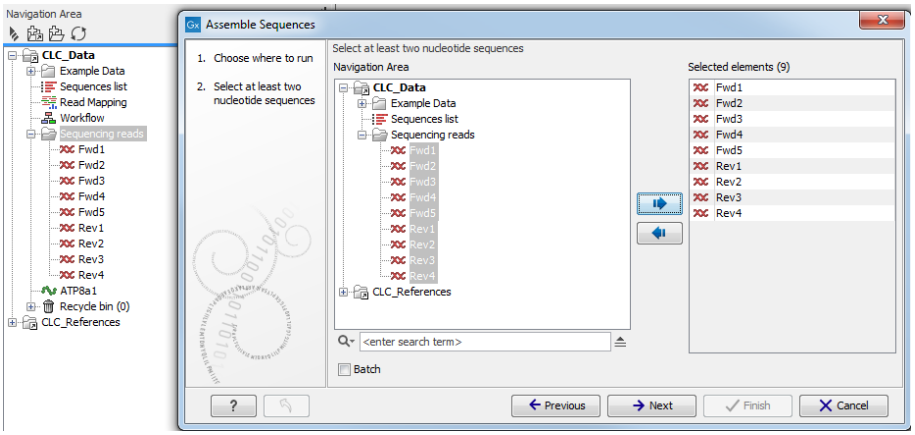


Figure 7.1: *You can select input files for the tool from the Navigation Area replicate on the left hand side of the wizard window.*

To select a file, you need to move it from the "Navigation Area" view to the "Selected elements" view. You can move a file by clicking on the arrows between the 2 views, or by double click on the file itself. Sometimes, having a file selected in the workbench Navigation Area will automatically put it in the Selected elements view. If you do not wish to work with this file as input, make sure to deselect it. You can deselect by using the arrow to send it out the selected elements view, or again by double clicking on it.

You also have the option in this window to work in batch, which means that the tool will run multiple times using each selected file as an independent dataset (as opposed to treat all selected files as a single input). To learn more about working in batch, please see section 7.3.

Once all elements are selected, you can click **Next** to proceed to the next step(s) of the wizard. During these steps you are usually required to set parameters for the tool (see figure 7.2 for

an example). You can read about specific settings for each tool in the relevant section of the manual accessible directly using the **Help** button((?)). A pop up window will open to the section of the manual describing the tool you are using. The **Restore Settings** button ( ) will reset all parameters from the pop window to their default values.

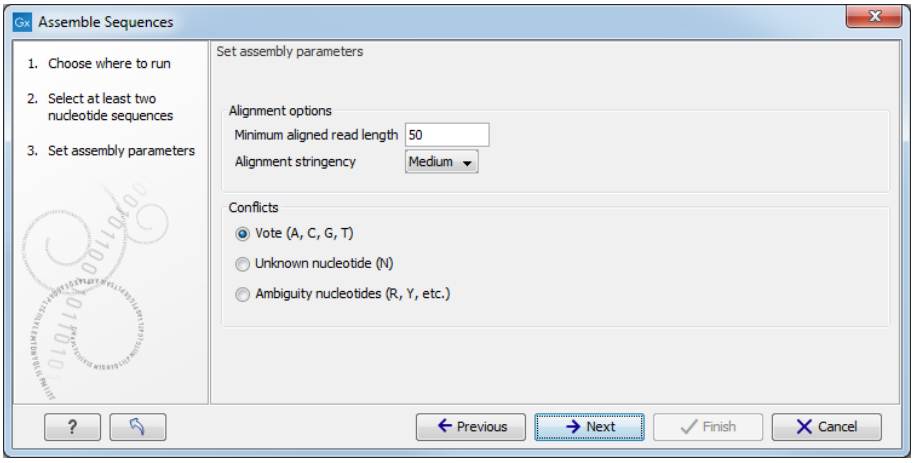


Figure 7.2: *An example of a "Set parameters" window.*

7.2 Handling results

A tool will output one or more result files, some of which are optional and can be selected - or deselected - in the last wizard window called "Result handling". The kind of output files generated by a tool as well as a description of additional files are described in the tool specific sections of the manual.

The "result handling" window also allows you to decide whether you want to open or save your results.

- **Open.** This will open the result of the analysis in a view. This is the default setting.
- **Save** The results will be saved rather than opened. You will be prompted for where you wish the results to be saved. See figure 7.3. You can save to an existing area or create a new folder to save the results into.

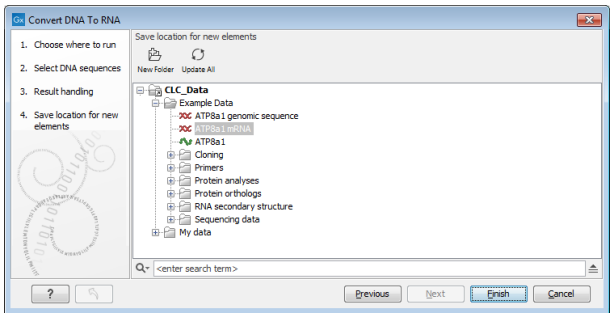


Figure 7.3: *Specify where to save the results of an analysis.*

You may also have the option to "Open log". A log is a textual view of the progress of the job.

Click on the button labeled **Finish** to start the tool.

If you chose "Open" results, they will open automatically open one or several Views in the View Area. Each View is described by a file name appended with an asterisk to indicate that this View has not been saved yet. To save it, drag the View tab to the relevant location in the Navigation Area, or simply use the usual Ctrl + S (or ⌘ + S). You can also right click on the tab and choose "Save" or use the "Save" button above the Navigation Area.

If you chose "Save" results, they will not open automatically, but they saved in the location you can specify in an extra wizard window. You can open the results by finding the file name in the Navigation Area after the tool is done processing, or by using the little arrow to the right of the analysis name in the Processes tab and choosing the option "Show results"(see figure 7.4).

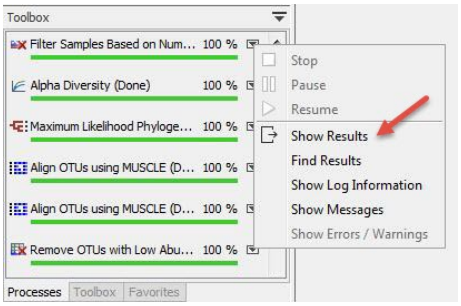


Figure 7.4: *Find your results using the little arrow to the right of the analysis name in the Processes tab.*

7.3 Batch processing

Batch processing refers to running an analysis multiple times, using different inputs for each analysis run. For example, if you have 10 sequence lists and wish to run 10 mapping analyses, one per sequence list, then these 10 analyses could be launched by setting up one batch job. When a job is run in batch mode, parameter settings stay the same for each run. It is just the inputs that are changed.

This section describes batch processing as it applies to most workbench tools and to workflows with a single input element.

Batching installed workflows with multiple input elements, where **all** input elements will be changed per batch, is done differently (see section 7.3.5).

7.3.1 Standard batch processing

Standard batch mode is activated by clicking the **Batch** checkbox in the dialog where the input data is selected (figure 7.5).

Unlike launching a single task, you can select a folder as well as, or instead of, individual data elements for the analysis.

Identifying batch units A batch unit is the set of data that will be used as a single input set for a given run of an analysis. A given batch unit can consist of one or more data elements.

If a folder is selected as input to a batch analysis, each folder or data element directly under that folder will be considered a batch unit. This means:

- Each individual data element contained directly within the folder is a batch unit.
- Each subfolder directly within this folder is a batch unit, so all elements within a given subfolder will be considered as single input for the purposes of the analysis.
- Elements in any more deeply nested subfolders (e.g. subfolders of subfolders of the originally selected folder) will not be considered for the analysis.

An example of a batch run is shown in figure 7.5.

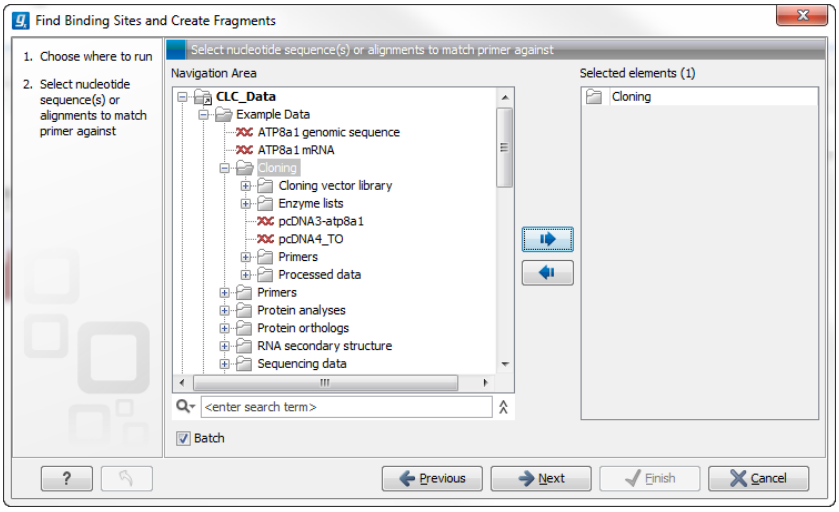


Figure 7.5: *The Cloning folder includes both folders and sequences.*

7.3.2 Batch overview

The next Wizard step is the batch overview where you have the opportunity to refine the list of data that will be in each batch unit. For example, you could use this step to ensure that only trimmed sequence lists and not all sequence lists, should be used for the analysis that is being setup.

The `Cloning` folder that is found in the example data (see section 1.6.2) contains two sequences ( ) and four folders ( ). If the Batch checkbox is checked and the Cloning folder is selected for an analysis, then after clicking on the button labeld **Next** an overview of the batch units like that in figure 7.6 is shown.

The batch overview lists the batch units on the left and the contents of the selected batch unit on the right.

In this example, the two sequences are defined as separate batch units because they are located at the top level of the Cloning folder. Of the four subfolders of the Cloning folder (see figure 7.5), three are listed in this view. In each of these subfolders, any data elements that the analysis could use as input will be used unless action is taken at this point to exclude some of these. Clicking on the name of each batch unit on the left hand side will initially cause the display of all the data elements that can be used by the analysis tool. These elements will be used as a single set of input for a particular analysis run. So, for example, in figure 7.6, all the elements in the subfolder `Cloning vector library` that will be included as part of a single analysis run shown on the right-hand side of the dialog.

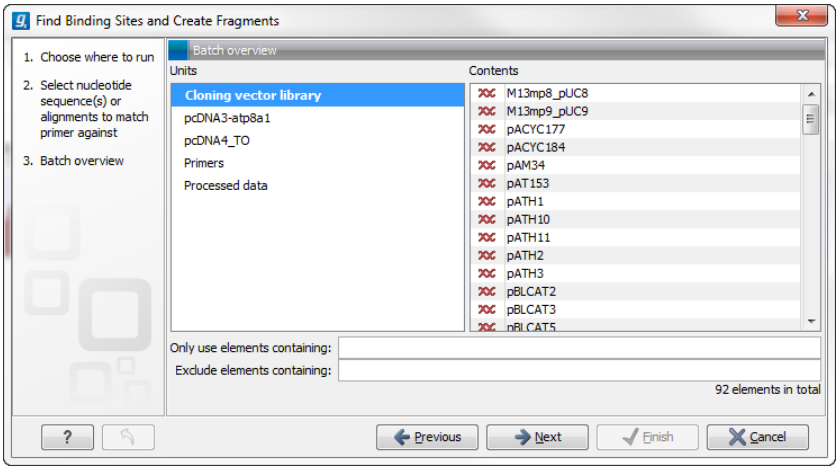


Figure 7.6: *Overview of the batch run. At the bottom right, the number of files to be analysed, summed across all batch units, is shown, 92 in this case.*

Note that the fourth subfolder of the Cloning folder, the `Enzyme lists` folder, is not listed as a batch unit. It is because it does not contain any data that can be used by the tool being launched.

Including and excluding data elements in batch units There are three ways to refine the data elements that should be included in a batch unit, and thereby get taken forward into the analysis.

1. **Use the fields labeled Only use elements containing and Exclude elements containing at the bottom of the batch overview** This refinement is done based on data element names. for example, only paired reads might be desired for the analysis, in which case, putting the text "paired" into the **Only use elements containing** field might be useful.

2. **Remove a whole batch unti** Right-click on the batch unit to be removed and choose the option **Remove Batch Unit**.

3. **Remove a particular data element from a batch unit** Right click on the element of a batch unit to be removed and choose the option **Remove Element**. This can be useful when filtering based on name, described in the first option, cannot be used to refine the batch units specifically enough.

7.3.3 Parameters for batch runs

The subsequent dialogs depend on the analysis being run and the data being input. Generally, one of the batch units will be specified as the parameter prototype and will then be used to guide the choices in the dialogs. By default, the first batch unit (marked in bold) is used for this purpose. This can be changed by right-clicking another batch unit and choosing the option **Set as Parameter Prototype**.

When launching tools normally (non-batch runs), the Workbench does much validation of inputs and parameters. When running in batch, this validation is not performed. This means that some analyses will fail if combinations of input data and parameters are not right. Therefore we recommend that batching is used when the batch units are quite homogenous in terms of the type and size of data.

7.3.4 Running the analysis and organizing the results

The last step in setting up a batch analysis is to choose where to save the outputs (figure 7.7).

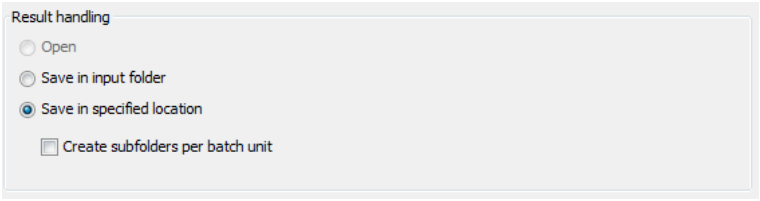


Figure 7.7: *Options for saving results when the tool was run in batch.*

The options available are:

- **Save in input folder** Save all outputs into the same folder as the input data. If the batch units consisted of folders, then the results of each analysis would be saved into the folder with the data it was generated using. If the batch units were individual data elements, then all the results will be placed into the same folder as those input data elements.
- **Save in specified location** Choose the folder where the outputs should be saved to, where when:
  - **Create subfolders per batch unit** is **unchecked**, all results for all batch units will be written to the specified folder.
  - **Create subfolders per batch unit** is **checked**, results for each batch unit will be written to a newly created subfolder of the selected folder. One subfolder is created per batch unit.

When the batch run is started, there will be one "master" process representing the overall batch job, and there will then be a separate process for each batch unit. The behavior this is different for Workbenches and Servers:

- On a Workbench, only one batch unit is run at a time. So when the first batch unit is done, the second will be started and so on. This avoids many parallel analyses that would draw on the same compute resources and slow down the computer.
- On a CLC Server (see `http://clcbio.com/server`), all the processes are placed in the queue, and the queue takes care of distributing the jobs. This means that if the server set-up includes multiple nodes, different batch unit analyses may be run in parallel.

To stop the whole batch run, stop the "master" process. From the Workbench, this can be done by finding the master process in the Processes tab in the bottom left hand corner. Click on the little triangle on the right hand side of the master process and choose the option **Stop**.

For some analyses, there is an extra option in the final step to create a log of the batch process. This log will be created in the beginning of the process and continually updated with information about the results. The log will either be saved with the results of the analysis or opened in a view with the results, depending on how you chose to handle the results.

7.3.5 Batch launching workflows with multiple inputs

This section describes the launching of workflows with multiple inputs, where **all** input elements will be changed per batch. This launch mechanism is not intended for workflows with multiple input elements where one of the input elements remains the same in all batches, such as workflows meant to compare several tissues to a unique control tissue. At the moment, batch launching of such workflows is not possible, unless the common item is saved under different names as many times as there should be batches.

For workflows with multiple inputs where the inputs all need to change for each batch run, information specifying the grouping of the data elements and what role each element plays in a given analysis needs to be imported into the system from an Excel spreadsheet.

The requirements for launching such workflows in batch mode are:

- The workflow must be installed on the Workbench, meaning that the workflow is accessible from the Toolbox (as opposed to workflows accessible from the Navigation Area). See section 8.2 to learn how to install a workflow.
- The workflow is characterized by more than one input file, and all input elements are unique per batch. You cannot reuse a common input element (such as control reads for example), unless it has been saved under different names in the Navigation Area.
- An Excel format file (.xlsx/.xls) must be provided, with at least 3 different columns:
  - **Unique ID** The first column must contain either the exact name of the data elements to be used as inputs, or partial name information such that data elements being entered into the analysis can be uniquely identified and matched with the information contained in the spreadsheet (see section 3.2.2 to learn more about matching partial names).
  - **grouping** A second column must specify which data elements should be analyzed together in a given batch unit: this would be the ID of a single individual when comparing different tissues from the same individual (one individual per batch); or a family name when identifying variants existing within one family (one family per batch).
  - **Type** The third column must specify the type for each data element: the values in this column distinguish tissue samples from controls, or inform about the disease status of a family member (affected/non-affected/proband) when identifying disease causing variants.

(Figure 7.15) shows an example of a spreadsheet used in the case of tissue comparison. Note that the "grouping" and "type" are context specific, and will depend on the analysis performed, i.e., on the tools that constitute the workflow.

To launch a workflow with multiple input elements in batch mode, right click on the name of the workflow in the Toolbox and select the option "Run in Batch Mode..." (figure 7.9).

A wizard opens and in the first window, you need to specify:

- An Excel file containing the information about the data to be analyzed (figure 7.10). Note that this file dos not need to be saved in your Navigation Area. When it has been selected, the table found in the lower part of the wizard will show recapitulate the content of the Excel

Sample ID = exact of partial name of the reads file to ensure a unique match between the reads and the metadata.	Patient ID = grouping values. Identical values will be analyzed together in one batch unit.	Type = value that defines which tissue is the control tissue and which is the sample tissue to be compared to the control.
23N	23	Normal
23T	23	Tumor
26N	26	Normal
26T	26	Tumor
27N	27	Normal
27T	27	Tumor
45N	45	Normal
45T	45	Tumor

Figure 7.8: *Example of a spreadhseet necessary to run a workflow in batch, where the workflow intend to compare two tissue samples.*

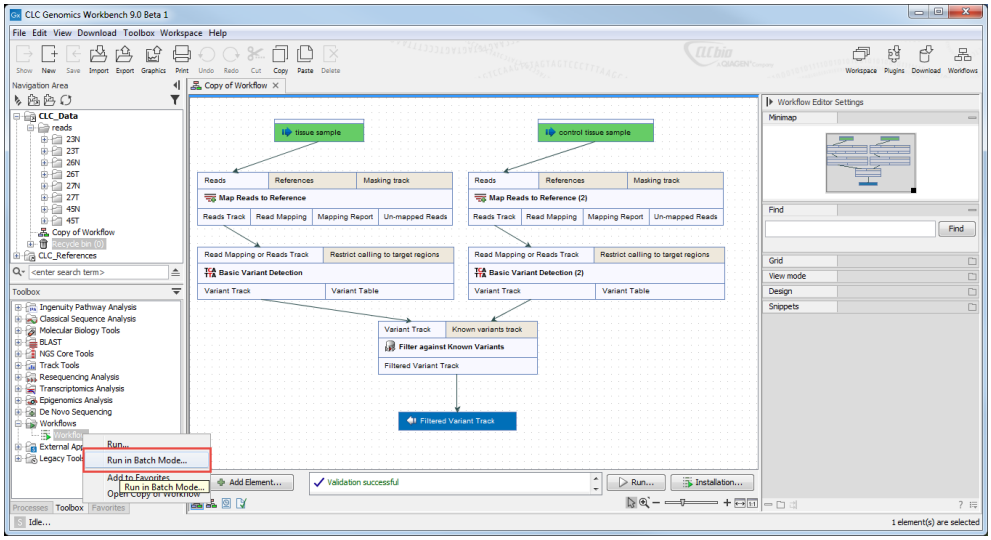


Figure 7.9: *The option to "Run in Batch Mode..." appears in the context menu when you right click on the name of an installed workflow that has multiple input elements in the Toolbox panel.*

sheet. The location of the data for this analysis is not yet specified, so a red, no-entry sign is visible in the header of the first column.

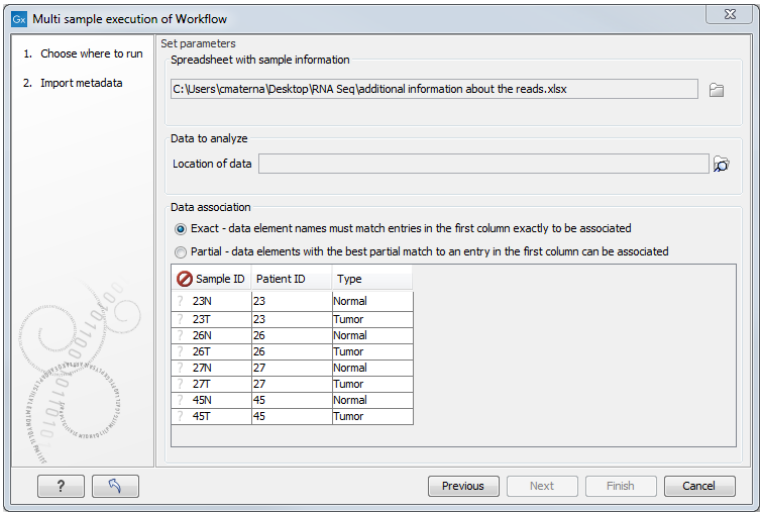


Figure 7.10: *Select the information about the data to be analyzed and the folder holding the data to analyze. An example of an Excel sheet with the relevant information is shown.*

- The location of the reads: click on the Navigation button next to the "Location of data" field and specify the folder(s) that contain(s) the data, as shown in figure 7.11.

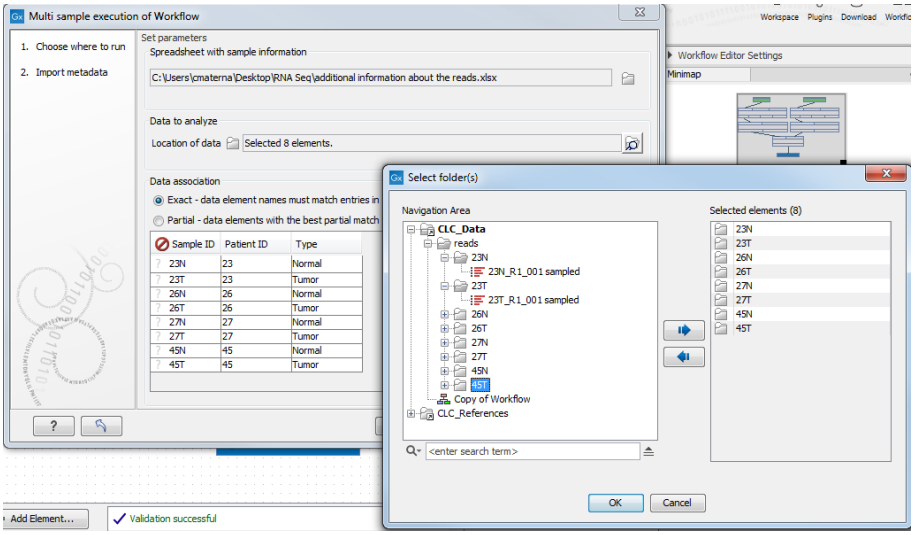


Figure 7.11: *Select the folder(s) that contain(s) the data to be analyzed.*

Data elements within the selected folders are considered for the analysis. Subfolders and their contents are not considered unless the subfolder is also selected. Individual data elements cannot be selected.

- Select the appropriate matching scheme - exact or partial. The matching rules applied are the same as those used for metadata association: "Exact" means that data element names must exactly match an entry in the first column of the Excel file; "Partial" matching allows for data elements names partially matching an entry in the first column. "Exact" is selected by default. Partial matching rules are described in detail in section 3.2.2.

An icon with a green check mark (✓) appears in the table preview next to rows where a data element corresponding to a row of the Excel sheet was uniquely identified. If no match can be made to a given row of the Excel sheet, a question mark (?) is displayed.

Graphical symbols are also presented in the header of the first column of the preview pane to give information about the overall status of the matching of rows in the Excel sheet with data elements in the Workbench:

- When no data elements match information in the Excel sheet, a red, no entry symbol (⊘) is displayed. In this situation, the button labeled **Next** is not enabled. This is the expected state before any data elements have been selected.
- A yellow exclamation mark (!) indicates that some, but not all rows in the Excel sheet have been matched to a data element in the selected folder(s).
- A green checkmark (✓) indicates that all rows in the Excel sheet have been matched to a data element in the selected folder(s).

In figure 7.12, the green check mark symbol in the header of the first column in the preview pane indicates that data elements were identified for each of the rows in the Excel sheet. You can click on the button labeled "Next".

The next wizard window is called "Select grouping parameters and analysis inputs".

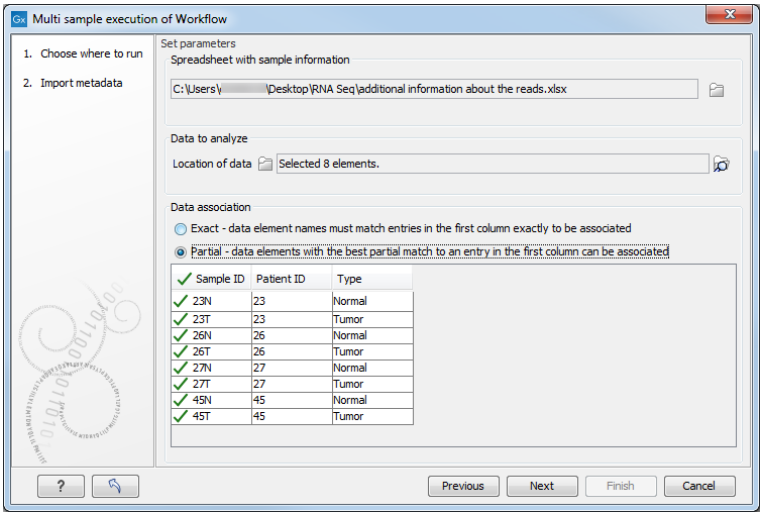


Figure 7.12: *View of the Data Association table after all samples were successfully associated.*

- In the **Group by** drop down menu, select the name of the column containing information that specifies which samples should be analyzed together.
- In the **Type** drop down menu, select the name of the column containing information that can be mapped to the workflow input type of each data element.

In the same window you will need to further specify the inputs of the workflow. What needs to be specified here is dependant on the workflow itself.

An example is shown in figure 7.13. **Group by** is set to a column specifying "Patient ID", because each workflow run will analyze a sample pair. **Type** is set to the "Type" column, because the workflow inputs are either tumor or normal tissues. The sample columns section maps data elements to the different workflow inputs, in this case "Tissue sample" is set to "Tumor", and "Control tissue sample" to "Normal".

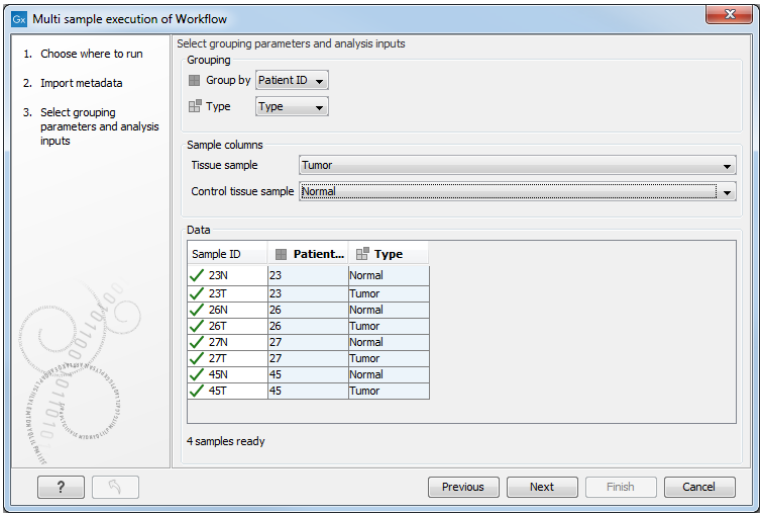


Figure 7.13: *Grouping samples.*

The rest of the wizard is dependant of the tools included in the workflow. Fill in the appropriate information and save the results of your workflow in a folder you can create in the Navigation Area.

As in a regular batching mode, you can use the progress bar to see how the job is progressing (figure 7.14): a process called "Batch Process" indicates how many batches have been completed, while the ones situated above show the analysis progress of a particular batch unit.

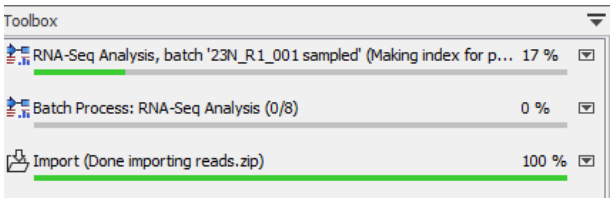


Figure 7.14: *Check on the progress of your workflow being run in batch mode using the Processes tab below the Toolbox.*

Ready-to-use workflows with more than one input in the *Biomedical Genomics Workbench* fall within two categories; 1) the Somatic Cancer workflow that compares tumor and normal samples, and 2) the Hereditary Disease workflows where a trio or a family of four is analyzed in one workflow.

(Figure 7.15) shows an example of the spreadsheet used in the Somatic Cancer workflows.

Sample ID = exact of partial name of the reads file to ensure a unique match between the reads and the metadata.	Patient ID = grouping values. Identical values will be analyzed together in one batch unit.	Type = value that defines which tissue is the control tissue and which is the sample tissue to be compared to the control.
23N	23	Normal
23T	23	Tumor
26N	26	Normal
26T	26	Tumor
27N	27	Normal
27T	27	Tumor
45N	45	Normal
45T	45	Tumor

Figure 7.15: *Example of a spreadhseet necessary to run a workflow in batch, where the workflow intend to compare two samples.*

To launch a workflow with multiple input elements in batch mode:

- Right click on the name of the workflow in the Toolbox panel in the bottom left hand side of the Workbench and select the option "Run in Batch Mode..." (figure 7.16).

  A Wizard like that in figure 7.17 should appear.

- Select the Excel file containing the information about the data to be analyzed (figure 7.18).

- Specify the folder with the data, as shown in figure 7.19.

  Data elements within the selected folders are considered for the analysis. Subfolders and their contents are not considered unless the subfolder is also selected. Individual data elements cannot be selected.

- Select the appropriate matching scheme - exact or partial. The matching rules applied are the same as those used for metadata association. Exact means that data element names must exactly match an entry in the first column of the Excel file. Partial matching allows for data elements names partially matching an entry in the first column. Partial matching rules are described in detail in section 3.2.2.

  An icon with a green check mark (✓) appears in the table preview next to rows where a data element corresponding to a row of the Excel sheet was uniquely identified. If no match can be made to a given row of the Excel sheet, a question mark (?) is displayed.

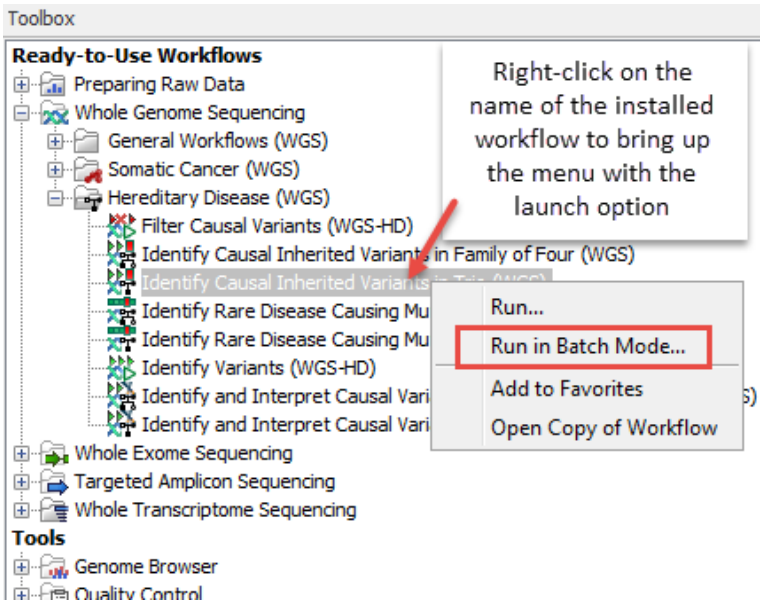


Figure 7.16: *The option to "Run in Batch Mode..." appears in the context menu when you right click on the name of an installed workflow that has multiple input elements in the Toolbox panel.*

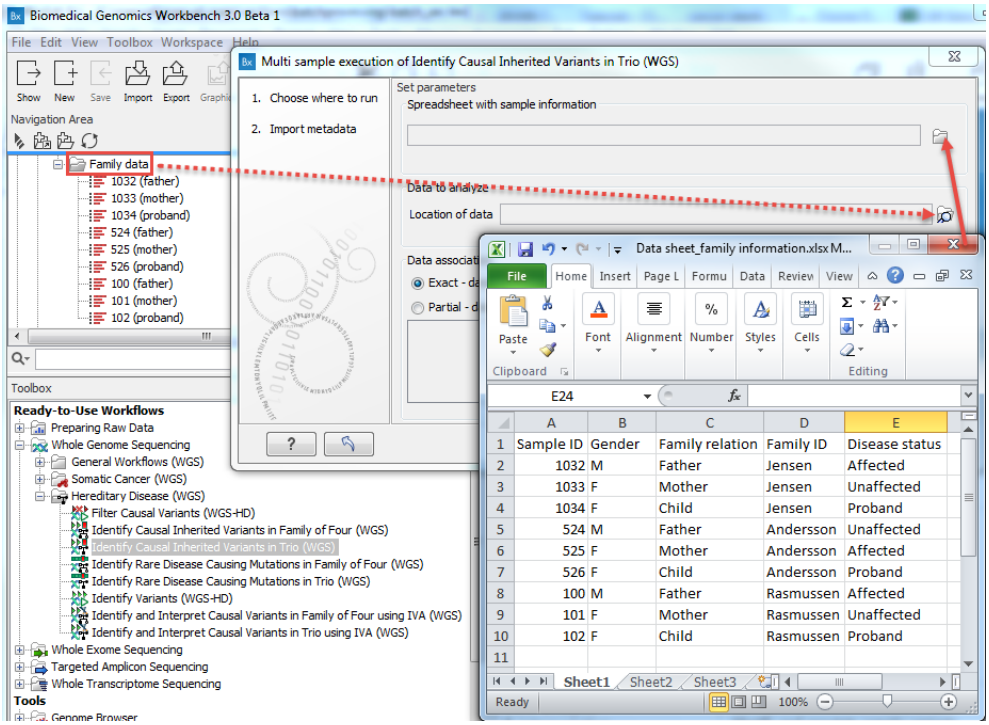


Figure 7.17: *Select the information about the data to be analyzed and the folder holding the data to analyze. An example of an Excel sheet with the relevant information is shown.*

Graphical symbols are also presented in the header of the first column of the preview pane to give information about the overall status of the matching of rows in the Excel sheet with data elements in the Workbench:

- When no data elements match information in the Excel sheet, a red, no entry symbol (⃠) is displayed. In this situation, the button labeled **Next** is not enabled. This is the expected state before any data elements have been selected.

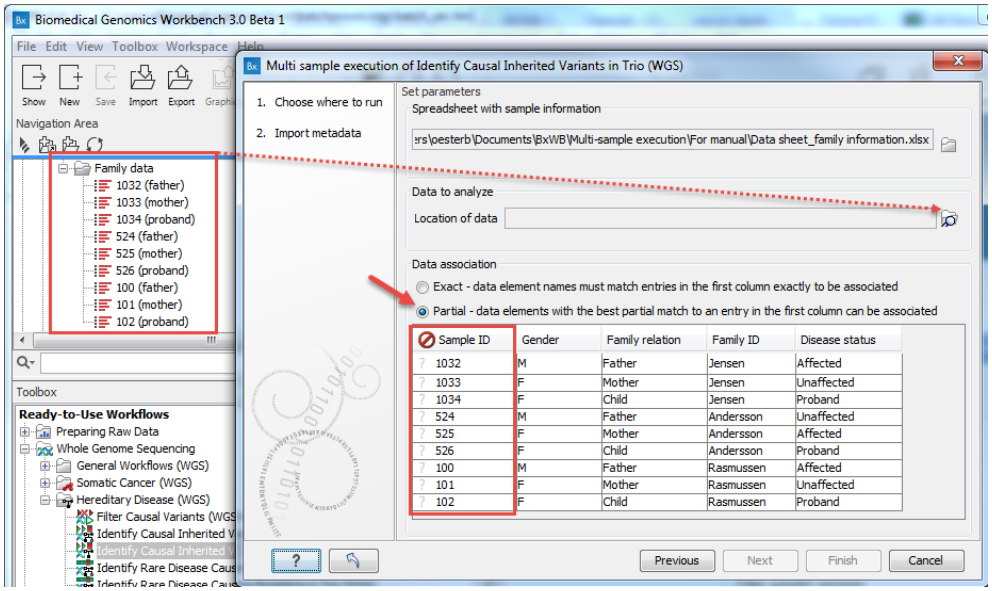


Figure 7.18: *When the Excel sheet has been selected, the table found in the lower part of the wizard will show the content of the Excel sheet. The location of the data for this analysis is not yet specified, so a red, no-entry sign is visible in the header of the first column.*

- A yellow exclamation mark () indicates that some, but not all rows in the Excel sheet have been matched to a data element in the selected folder(s).
- A green checkmark () indicates that all rows in the Excel sheet have been matched to a data element in the selected folder(s).

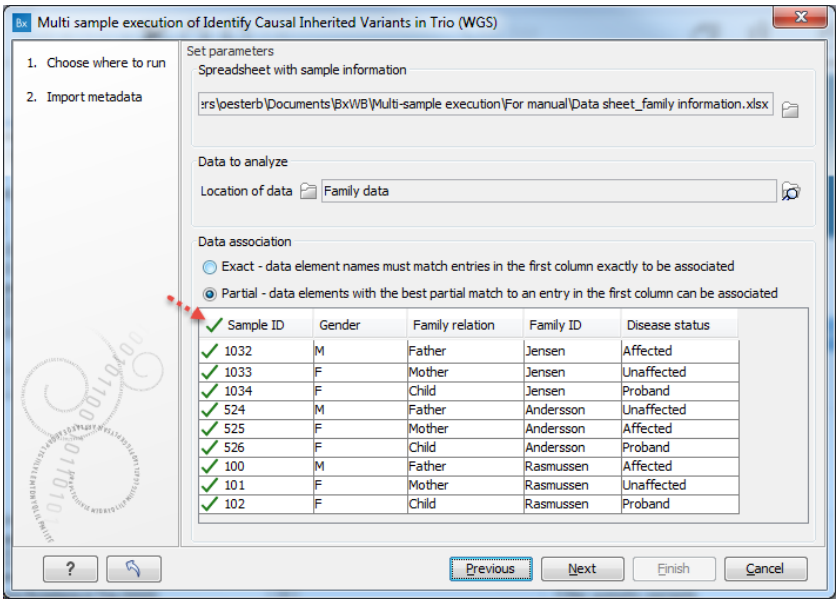


Figure 7.19: *Click on a folder or folders that contain the data to be analyzed. Here, the green check mark symbol in the header of the first column in the preview pane indicates that data elements were identified for each of the rows in the Excel sheet.*

- Click on the button labeled **Next**.

  In the Grouping area of the Wizard shown in figure 7.20:

- In the **Group by** drop down menu, select the name of the column containing information that specifies which samples should be analyzed together.
- In the **Type** drop down menu, select the name of the column containing information that can be mapped to the workflow input type of each data element.

  An example is shown in figure 7.20, where a hereditary workflow is being launched in batch mode. **Group by** is set to a column specifying family names, because each workflow run will analyze a particular family. **Type** is set to the Disease status column, because the workflow inputs are an unaffected parent, an affected parent and a proband, and the Disease status column holds entries that can be mapped to these input types.

  The same Excel sheet shown in figure 7.20 could also be used where the workflow input types were instead mother, father and child. In that case, the column called Family relation would be set as the **Type**, since that is the column with entries that can be mapped to those particular workflow input types.

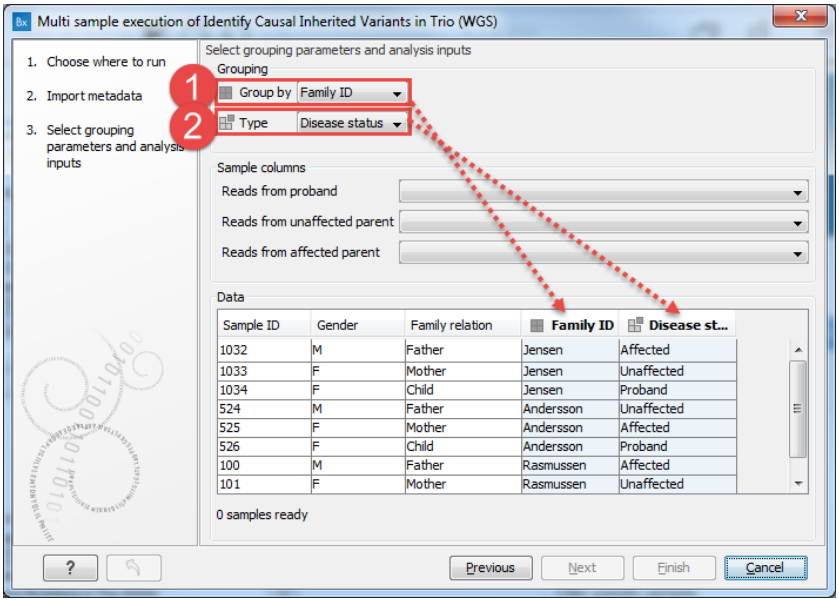


Figure 7.20: *A hereditary workflow is being lauched in batch mode. A given workflow run should analyze a family group, so the Group by entry is set to the column Family ID, where family groupings are specified. The workflow input types here are an unaffected parent, an affected parent and a proband. Information that can be mapped to these input types is held in the Disease status column, so this is selected in the Type drop down menu.*

Further details about the information in the Type column is now entered in the Sample columns area of the Wizard. For each input type for the workflow being launched, a drop down menu is provided containing the column entries from the column specified as containing the **Type** information.

- For each workflow input type listed, click on the drop down menu and select the term used to identify that particular input type. See figure 7.21.
- Click on the button labeled **Next**.
- Work through any remaining Wizard steps where analysis details are presented and configure any unlocked parameters.

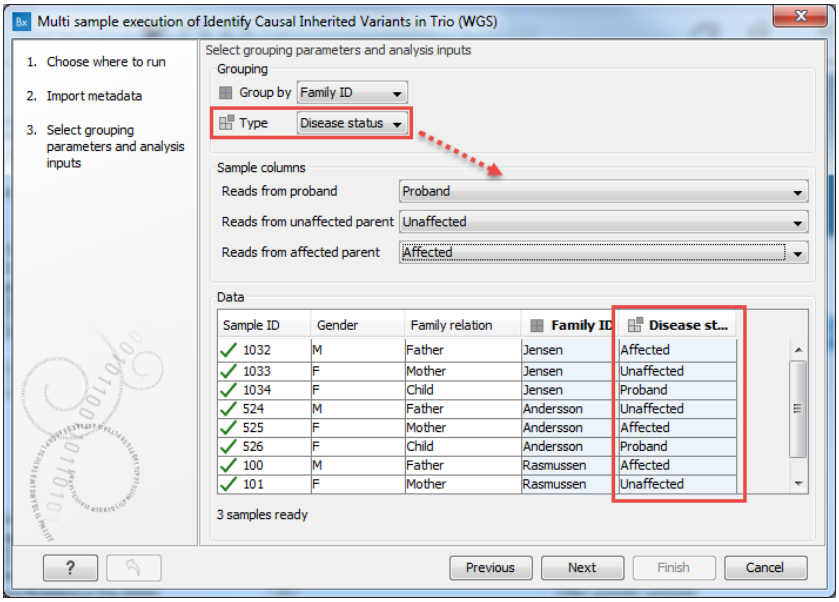


Figure 7.21: *The selections shown here indicate that data elements identified as matching rows from the Excel sheet containing "Proband" in the Disease status column should be used as the workflow input type "proband", data elements identified as matching rows containing "Unaffected" should be used as the workflow input type "unaffected parent", and data elements identified as matching rows containing "Affected" should be used as the workflow input type "affected parent".*

- Choose where to save the outputs of the analysis.
- Click on the button labeled Finish.

Important note: When running the Identify Rare Disease Causing Mutations ready-to-use workflows in batch mode, the gender of all proband samples in a given batch run must be the same. In other words, if multiple families are analyzed in a batch run, **the probands must all be female or they must all be male**. This is because proband gender is specified as a parameter, and the parameter values provided when setting up a workflow are then used for each analysis in the batch. The same condition applies when running a workflow in batch mode that includes a Trio Analysis. The gender of all child samples being analyzed in a given batch run must be the same.

Chapter 8

Workflows

Contents

The *CLC Drug Discovery Workbench* provides a framework for creating, distributing, installing and running workflows. A workflow consists of a series of connected tools where the output of one tool is used as input for another tool. Once the workflow is set up, it can be installed (either in your own Workbench or it can be shared with colleagues and installed on a server). In that way it becomes possible to analyze lots of samples using the same standard pipeline, the same reference data and the same parameters.

This chapter will first explain how to create a new workflow, and next go into details about the installation and execution of a workflow.

Note that the examples below are using tools from the *CLC Genomics Workbench* that are not necessarily available in the *CLC Drug Discovery Workbench*. But the principles and workflow framework can be used in the same way with tools from *CLC Drug Discovery Workbench*.

8.1 Creating a workflow

A workflow can be created by pressing the "Workflows" button ( ) in the toolbar and then selecting "New Workflow..." ( ).

Alternatively, a workflow can be created via the menu bar:

File | **New** | **Workflow ( )**

This will open a new view with a blank screen where a new workflow can be created.

8.1.1 Adding workflow elements

First, click the **Add Element** ( ) button at the bottom (or use the shortcut Shift + Alt + E). This will bring up a dialog that lists the elements and tools, which can be added to a workflow (see figure 8.1).

Alternatively elements can be dragged directly from the **Toolbox** into the workflow. Only workflow enabled elements can be dropped in the workflow.

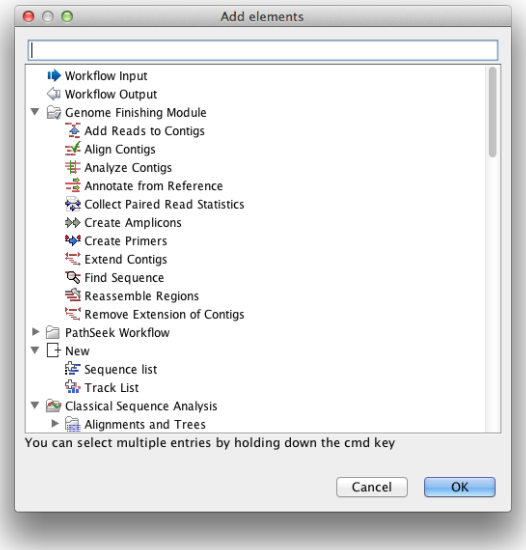


Figure 8.1: *Adding elements in the workflow.*

Elements that can be selected in the dialog are mostly tools from the Toolbox. However, there are two special elements on the list: the elements that are used for input and output. These are explained in section 8.1.5.

You can select more than one element in the dialog by pressing Ctrl (⌘ on Mac) while selecting. Click OK when you have selected the relevant tools. You can add more later on if you wish.

You will now see the selected elements in the editor (see figure 8.2).

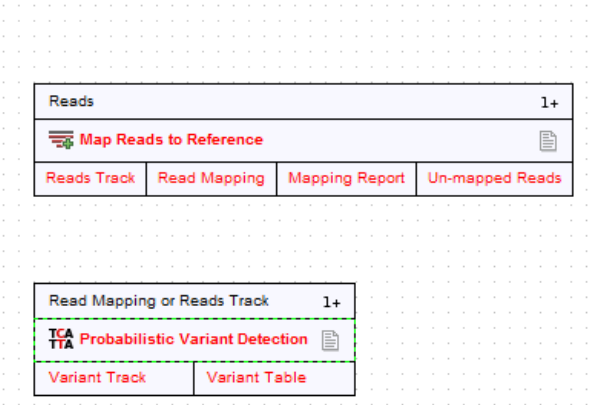


Figure 8.2: *Read mapping and variant calling added to the workflow.*

Once added, you can move and re-arrange the elements by dragging them with the mouse. To do this, click on part of the box containing the name of the element and then, keeping the mouse button depressed, drag the element to the desired position.

8.1.2 Configuring workflow elements

Each of the tools can be configured by right-clicking the name of the tool as shown in figure 8.3.

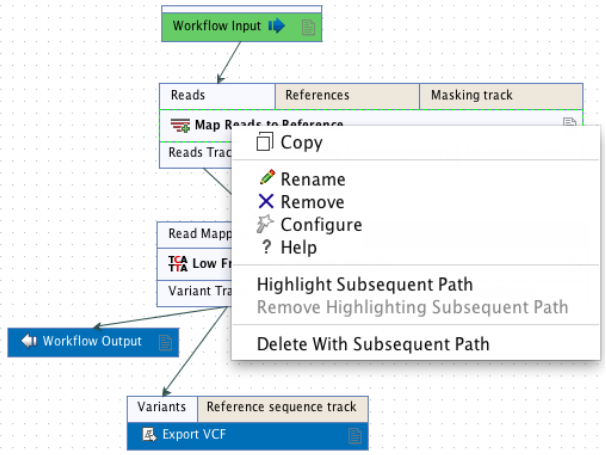


Figure 8.3: *Configuring a tool.*

The first option you are presented with is the option to **Rename** the element. This can be useful when you wish to discriminate between several copies of the same tool in a workflow. The name of the element is also visible as part of the process description when the workflow is run. To rename the element, right click on the tool in the workflow and select the "Rename" option, or click on the tool in the workflow and then press the F2 key.

The **Remove** option is used to remove elements from the workflow. The shortcut Alt + Shift + R removes all elements from the workflow.

You can also configure a given element using the **Configure** option from the right click menu or by double-clicking on the element. This opens a dialog with options for setting parameters, selecting reference data, selecting the export destination of specified columns, etc. An example

is shown in figure 8.4.

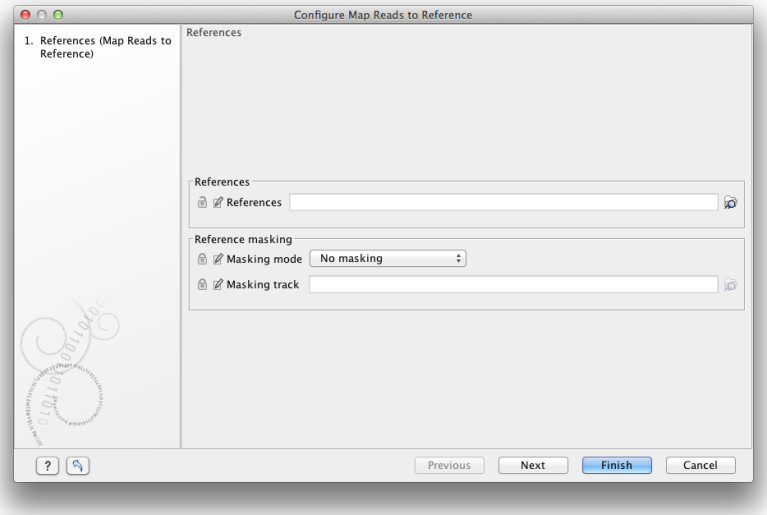


Figure 8.4: *Configuring read mapper parameters.*

Click through the dialogs using **Next** and press **Finish** when you are done. This saves the parameter settings for that tool. These are then applied when the workflow is executed.

You can also change the name of a parameter if you so wish, for example, to help with usability for the intended users of a workflow. To do this, click on the edit icon ( ) and enter a new name.

Special consideration should be given when configuring reference data in a workflow. For example, when configuring a read mapping tool, such as shown in figure 8.3, you have to define a reference genome that sequences will be mapped to. You configure this by selecting data in the **Navigation Area**. If you distribute the workflow and it is installed on a different system, where that data is not accessible in same relative location, the workflow installation procedure will involve defining new reference data to use. This is explained in more detail in section 8.2.

The lock icons in the dialog are used for specifying whether the parameter should be locked and unlocked as described in the next section. Locking parameters means that the workflow will be run with the same parameters every time; the user will not be prompted to supply values for locked parameters when they launch the workflow.

Once an element has been configured, the workflow element will be shaded with a darker color to help in distinguishing which elements have been configured.

The **Highlight Subsequent Path** option causes the element that was clicked on, and all the elements downstream of that one, to be highlighted. Other elements will be grayed out (figure 8.5). The **Remove Highlighting Subsequent Path** option reverts the highlighting, returning to the normal workflow layout.

In some workflows, many elements use the same reference data. There is a quick way of configuring these: right-click on the empty space and choose **Configure All References**. A dialog then appears listing all the reference data needed by the workflow. When you click on the button labeled Finish, only the elements where the 'active' column is checked will be configured.

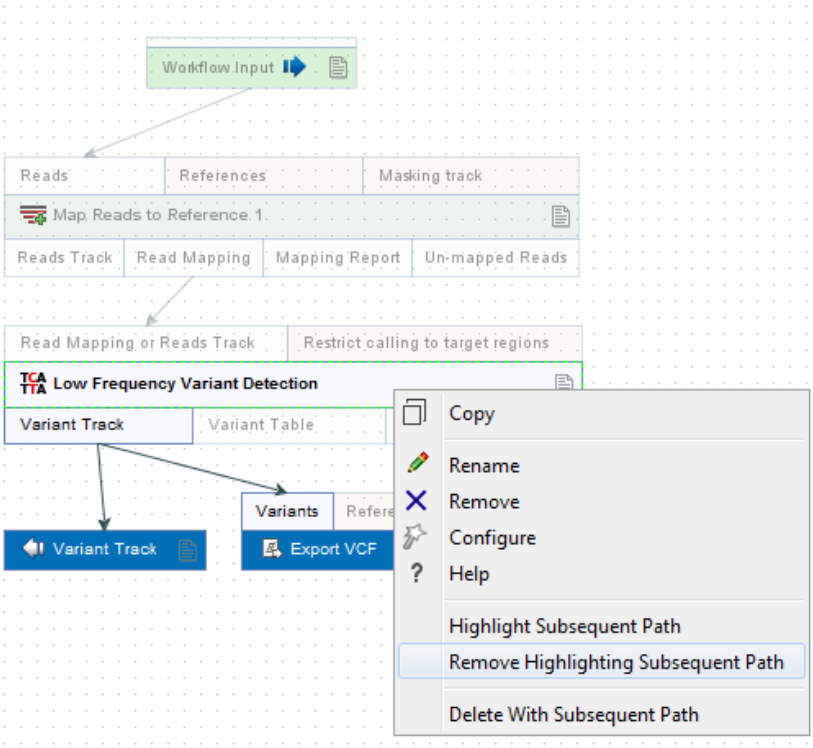


Figure 8.5: *Highlight path from the selected tool and downstream.*

Similarly, instead of configuring the various tools individually, the **Configuration Editor** enables the specification of all settings, references, masking parameters etc. through a single wizard window (figure 8.6). This editor is accessed through the ( ) icon located in the lower left corner.

8.1.3 Locking and unlocking parameters

Figure 8.7 illustrates the different stages in the lifecycle of a workflow.

The first stage, workflow creation, was explained in the section above. The next stage, the installation of a workflow on a Workbench or Server is explained in section 8.2). The final stage is the executation of the workflow via the **Toolbox**, just like other tools.

During the creation step, the workflow author can specify which parameters should be locked or unlocked. If a parameter is locked, it cannot be changed in the installation or the execution step. Conversely, if it is left open, that parameter can be changed, either when running the Workflow or when installing it. See section 8.2). The lock icons shown in figure 8.4 specify whether the parameter is left open or whether it is locked.

By default, data parameters are unlocked. When installing the workflow on a different system to the one where it was created, the connection to the data needs to be re-established. This is only possible when the parameter is unlocked. Data parameters should only be locked if they should not be set, or if the workflow will only be installed in a setting where there is access to the same data in the same location as the system where the Workflow was created.

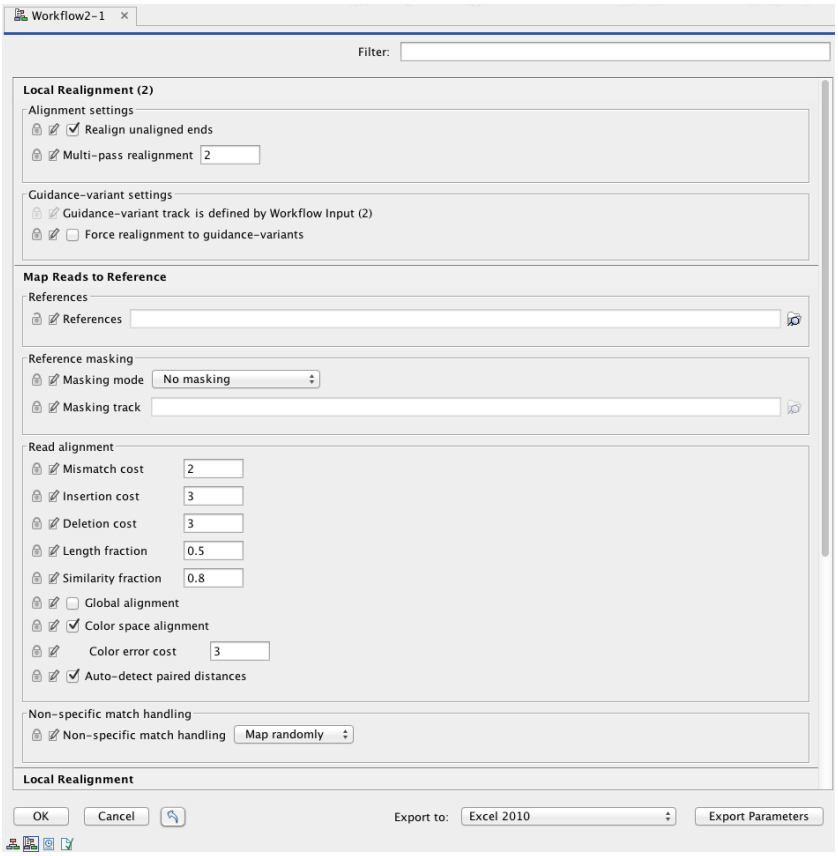


Figure 8.6: *The Configuration Editor can be used configure all the tools that can be configured in a given Workflow.*

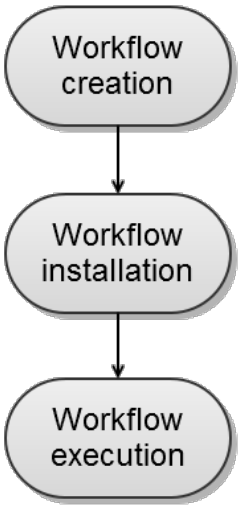


Figure 8.7: *The life cycle of a workflow.*

8.1.4 Connecting workflow elements

Figure 8.8 explains the different parts of a workflow element.

At the top of each element a description of the required type of input is found. In the right-hand side, a symbol specifies whether the element accepts multiple incoming connections, e.g. `+1`

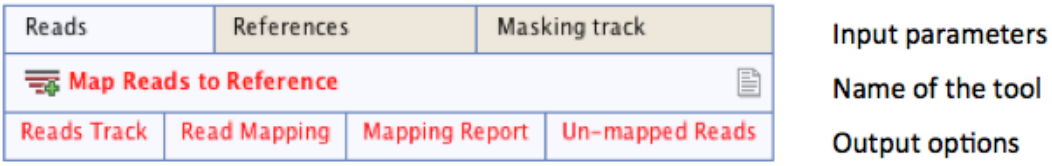


Figure 8.8: *A workflow element consists of three parts: input, name of the tool, and output.*

means that more than one output can be connected, and no symbol means that only one can be connected. At the bottom of each element there are a number of small boxes that represent the different kinds of output that is produced. In the example with the read mapper shown in figure 8.2, the read mapper is able to produce a reads track, a report etc.

Each of the output boxes can be connected to further analysis in three ways:

- By dragging with the mouse from the output into the input box of the next element. This is shown in figure 8.9. A green border around the box will tell you when the mouse button can be released, and an arrow will connect the two elements (see figure 8.10).
- Right-clicking the output box will display a list of the possible elements that this output could be connected to. You can also right-click the input box of an element and connect this to a matching output of another element.
- Alternatively, if the element to connect to is not already added, you can right-click the output and choose **Add Element to be Connected**. This will bring up the dialog from figure 8.1, but only showing the tools that accepts this particular output. Selecting a tool will both add it to the workflow and connect with the output you selected. You can also add an upstream element of workflow in the same way by right-clicking the input box.

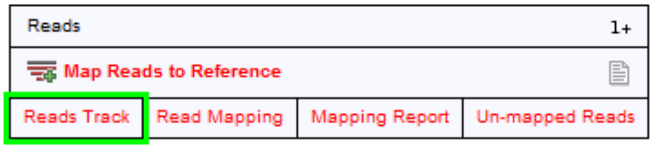


Figure 8.9: *Dragging the reads track output with the mouse.*

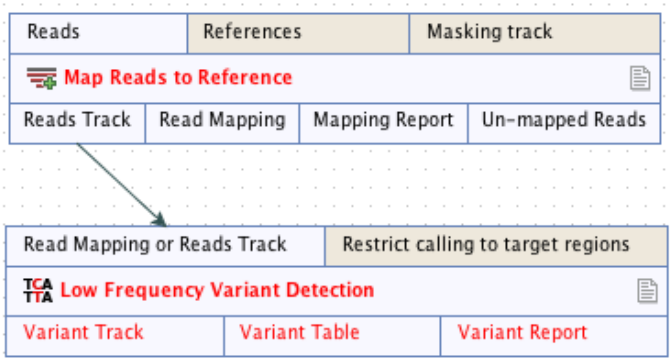


Figure 8.10: *The reads track is now used for variant calling.*

All the logic of combining output and input is based on matching the type of input. So the read mapper creates a reads track and a report as output. The variant caller accepts reads tracks

as input but not mapping reports. This means that you will not be able to connect the mapping report to the variant caller.

Figure 8.11 demonstrates how one tool can receive input from two different sources; 1) a reads track that is the input that hold the data that is to be analyzed (in this case reads that is to be locally realigned), and 2) a parameter that can have different functions depending on the tool that it is connected to (in this case the InDel track is used as a guidance track for the local realignment. In other situations the parameter track could be used for e.g. annotation or could provide a reference sequence).

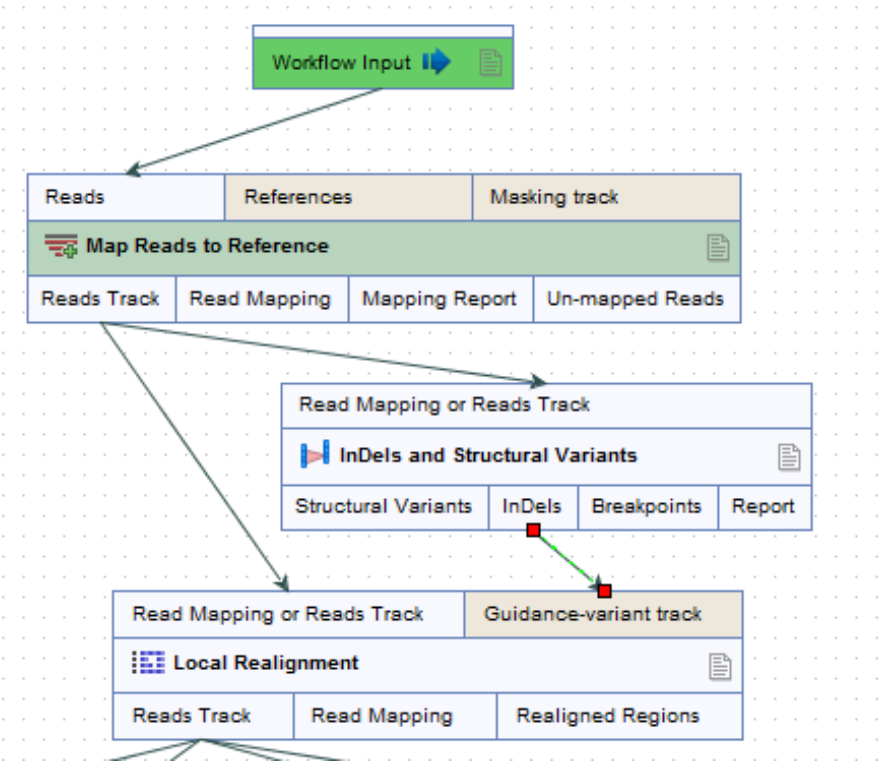


Figure 8.11: *A tool can receive input from both the generated output from another tool (in this example a reads track) and from a parameter (in this case indels detected with the InDels and Structural Variants tool).*

8.1.5 Input and output

Besides connecting the elements together, you have to decide what the input and the output of the workflow should be. We will first look at specification of the output, which is done by right-clicking the output box of any tool and selecting **Use as Workflow Output** as shown in figure 8.12.

You can mark several outputs this way throughout the workflow. Note that no intermediate results are saved unless they are marked as workflow output[1].

By double-clicking the output box, you can specify how the result should be named as shown in figure 8.13.

[1]When the workflow is executed, all the intermediate results are indeed saved temporarily but they are automatically deleted when the workflow is completed. If a part of the workflow fails, the intermediate results are not deleted.

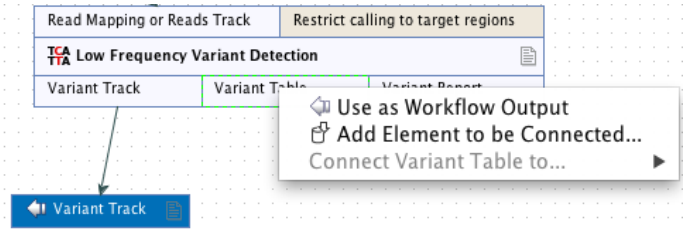


Figure 8.12: *Selecting a workflow output.*

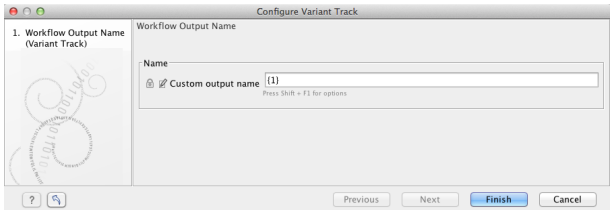


Figure 8.13: *Specifying naming of a workflow output.*

In this dialog you can enter a name for the output result, and you can make use of two dynamic placeholders for creating this name (press Shift + F1 to get assistance):

- {1} Represents the default name of the result. When running the tool outside of a workflow, this is the name given to the result.
- {2} Represents the name of the workflow input (not the input to this particular tool but the input to the entire workflow).

An example of a meaningful name to a variant track could be `{2} variant track` as shown in figure 8.14. If your workflow input is named `Sample 1`, the result would be `Sample 1 variant track`.

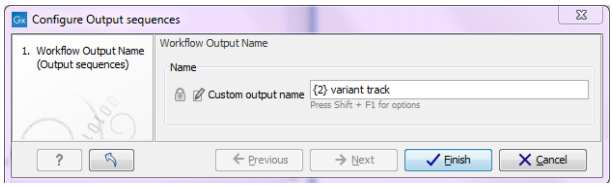


Figure 8.14: *Providing a custom name for the result.*

In addition to output, you also have to specify where the data should go into the workflow by adding an element called **Workflow Input**. This can be done by:

- Right-clicking the input box of the first tool and choosing **Connect to Workflow Input**. By dragging from the workflow input box to other input boxes several tools can use the input data directly.
- Pressing the button labeled **Add Element** (or right-click somewhere in the workflow background area and select **Add Element** from the menu that appears). The input box must then be connected to the relevant tool(s) in the workflow by dragging from the Workflow Input box to the "input description" part of the relevant tool(s) in the workflow.

At this point you have only prepared the workflow for receiving input data, but not specified which data to use as input. To be able to do this you must first save the workflow. When this has been done, the button labeled **Run** is enabled which allows you to start executing the workflow. When you click on the button labeled **Run** you will be asked to provide the input data.

Multiple input files can be used when:

- Data is generated within the Workflow
- Data is held within the Workbench
- Data is a combination of the two situations above

It can be useful to rename input elements when working with multiple input files, so that it is easy to discriminate between them when they are shown during workflow execution.

Note: Once the multiple input feature is used in a workflow, it is not possible to run the workflow in batch mode.

You can choose the order in which inputs will be processed by an element by right clicking on the input parameter box at the top of the element and choosing the option **Order Inputs**. This is most relevant for elements involved in data visualization. The feature **Order Inputs** is enabled when there are at least two inputs connected to the element (see figure 8.15). A small window will open, in which you can indicate the preferred order of the inputs to that element by moving them up and down in the list (figure 8.16). From this point forward, the order of the inputs is displayed on the branches connecting the inputs to elements.

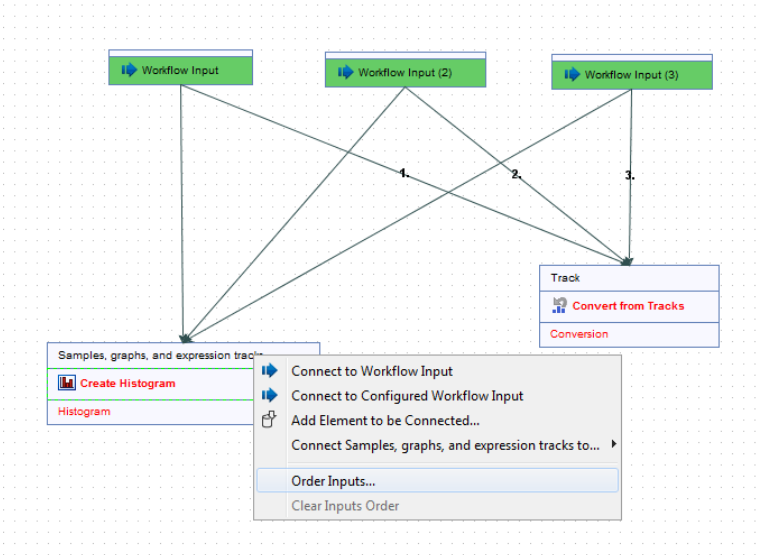


Figure 8.15: *Right-click on the input parameter box to see the Order Inputs function.*

The feature **Order Workflow Inputs** allows you to set the order that a user will be asked for each input when they run the workflow. This option is enabled as soon as the workflow has two or more inputs (figure 8.17). Right click on empty space in the Workflow editor to start this tool. A small window will open in which the different inputs can be moved up and down to indicate the desired order (figure 8.18).

The example in figure 8.19 shows how to generate a track list in a workflow. Any track based on a compatible genome can be added to the same track list. This includes reference tracks as well

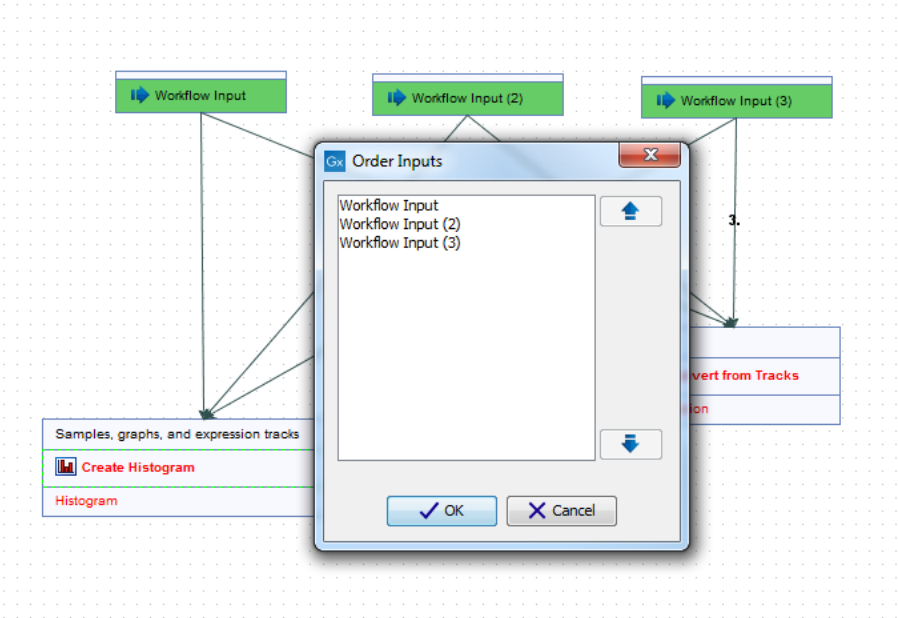


Figure 8.16: *Define the inputs order for the element.*

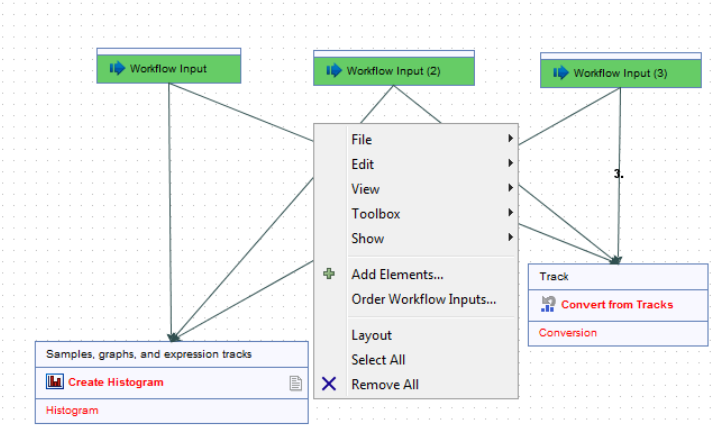


Figure 8.17: *Right click on empty space in the Workflow editor to open the Order Workflow Inputs tool.*

as track results generated by elements of that workflow. In the latter case, only those for which a workflow output element has been configured can be included in a track list.

8.1.6 Layout

The workflow layout can be adjusted automatically. Right clicking in the workflow editor will bring up a pop-up menu with the option "Layout". Click on "Layout" to adjust the layout of the selected elements (Figure 8.20). Only elements that have been connected will be adjusted.

Note! The layout can also be adjusted with the quick command Shift + Alt + L.

Note! It is very easy to make an image of the workflow. Simply select the elements in the workflow (this can be done pressing Ctrl + A, by dragging the mouse around the workflow while holding down the left mouse button, or by right clicking in the editor and then selecting "Select

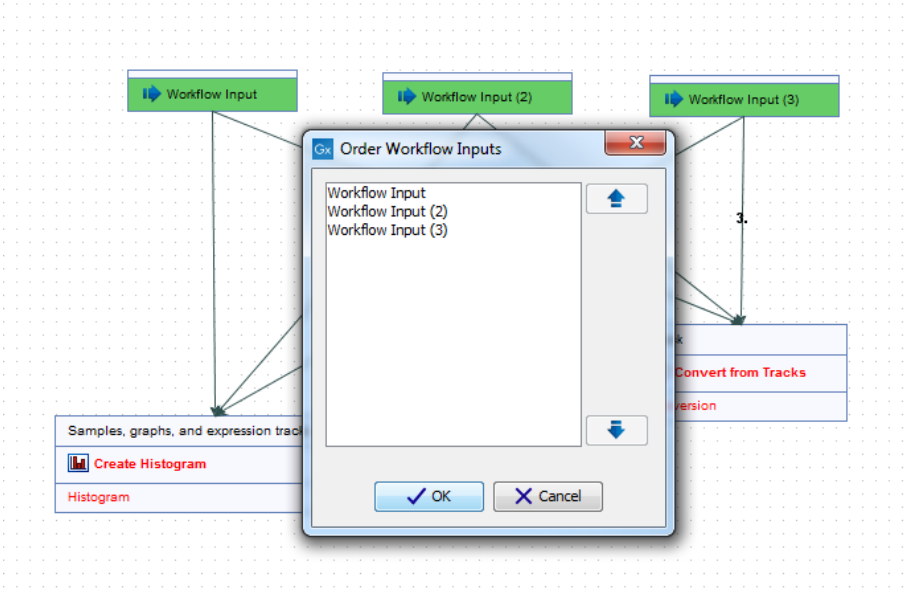


Figure 8.18: *Define the order of the inputs for the workflow.*

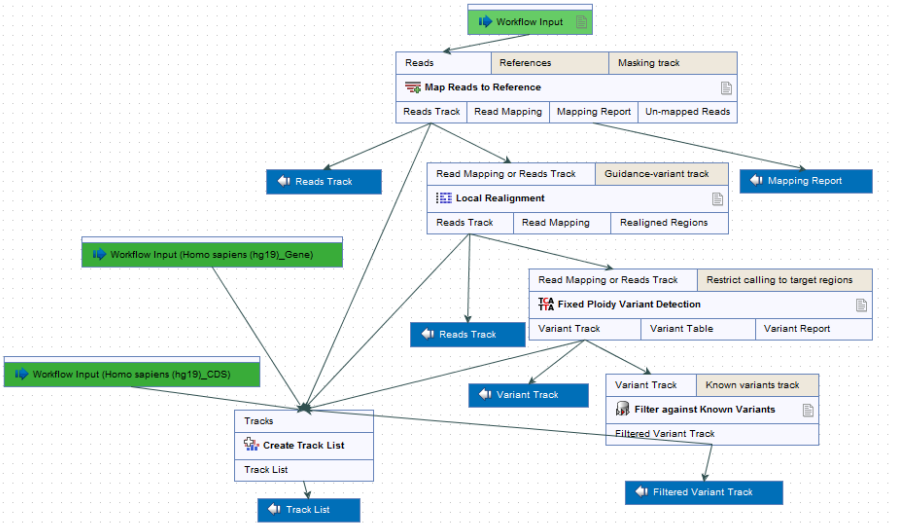


Figure 8.19: *Generation of a track list including data generated within the Workflow, as well as data held in the Workbench.*

All"), then press the Copy button in the toolbar ( ) or CTRL + C. Press Ctrl + V to paste the image into the wanted destination e.g. an email or a text or presentation program.

8.1.7 Input modifying tools

An input modifying tool is a tool that manipulates its input objects (e.g. adds annotations) without producing a new object. This behavior differs from the rest of the tools and requires special handling in the workflow.

In the workflow an input modifying tool is marked with the symbol (Ⓜ) (figure 8.21).

Restrictions apply to workflows that contain input modifying tools. For example, branches are not allowed where one of the elements is a modifying tool (see figure 8.22), as it cannot be

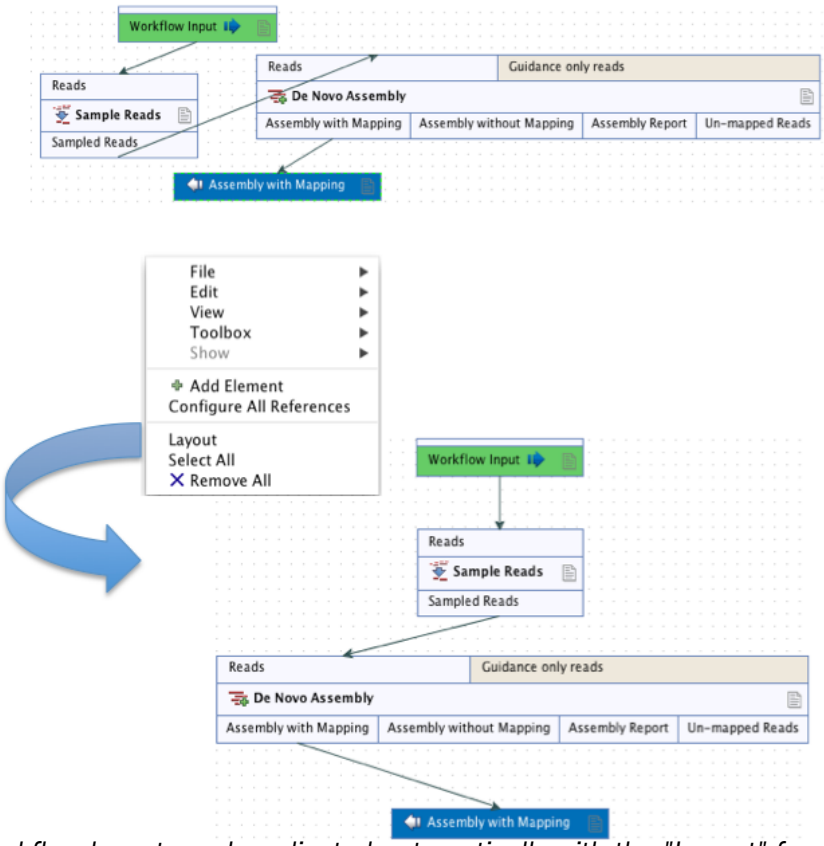


Figure 8.20: *A workflow layout can be adjusted automatically with the "Layout" function.*

Figure 8.21: *Input modifying tools are marked with the letter M.*

guaranteed which workflow branch will be executed first, which in turn means that different runs can result in production of different objects. Hence, if a workflow is constructed with a branch where one of the succeeding elements is a modifying tool, a message in red letters will appear saying "Branching before a modifying tool can lead to non-deterministic behavior". In such a situation the "Run" and "Create Installer" buttons will be disabled (figure 8.22).

The problem can be solved by resolving the branch by putting the elements in the right order (with respect to order of execution). This is shown in figure 8.23 that also shows that the "Run" and "Create Installer" buttons are now enabled. In addition, a message in green letters has appeared saying "Validation successful".

As input modifying tools only modify existing objects without producing a new object, it is not possible to add a workflow output element directly after an input modifying tool (figure 8.24).

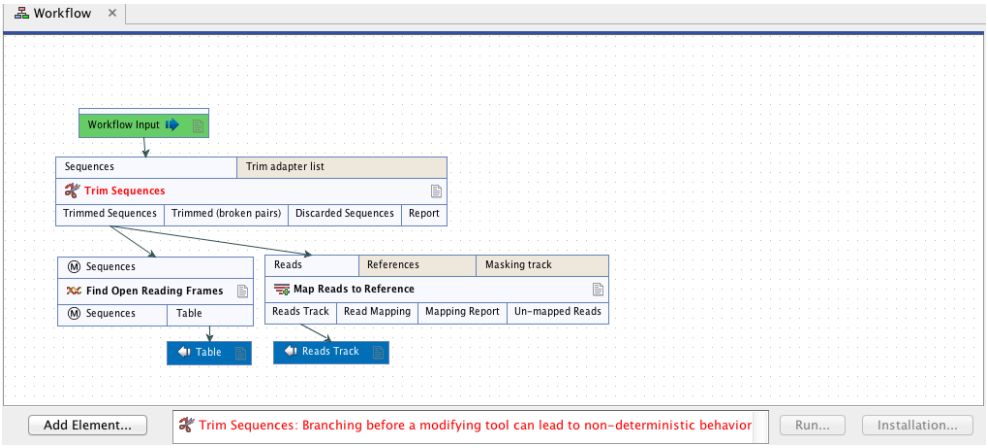


Figure 8.22: *A branch containing an input modifying tool is not allowed in a workflow.*

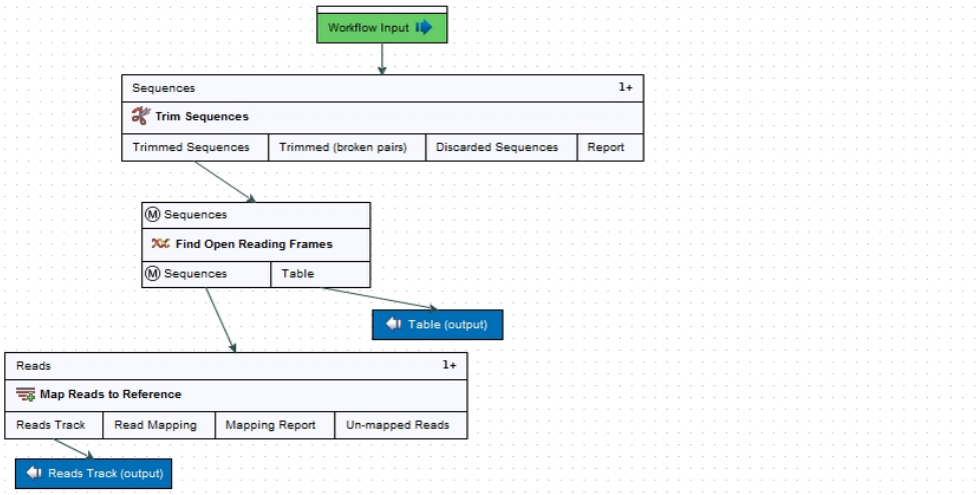


Figure 8.23: *A branch containing an input modifying tool has been resolved and the workflow can now be run or installed.*

A workflow output element can only be added when other tools than input modifying tools are included in the workflow.

If the situation occur where more input modifying tools are used succeedingly, a copy of the object will be created in addition to using the modified object as input at the next step of the chain (see figure 8.25). In order to see this output you must right click on the output option (marked with a red arrow in figure 8.25) and select "Use as Workflow Output".

When running a workflow where a workflow output has been added after the first input modifying tool in the chain (see figure 8.26) the output arrow is marked with "copy" to indicate that this is a copy of the result that is used as input at the next level in the chain. When running this workflow you will be able to see the copy of the output from the first input modifying tool in the **Navigation Area** (at the destination that you selected when running the workflow).

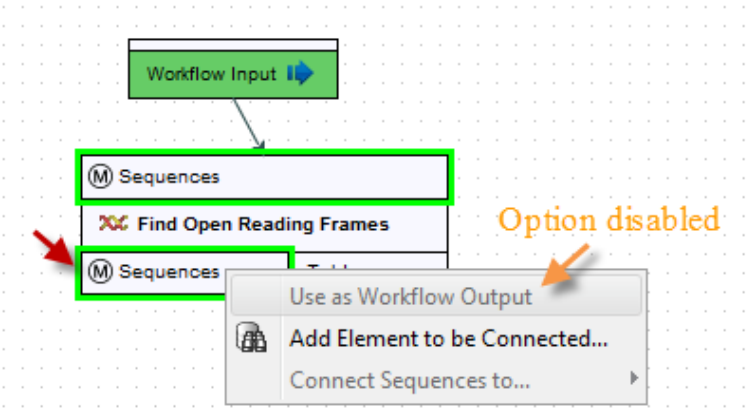


Figure 8.24: *A workflow output element cannot be added if the workflow only contains an input modifying tool.*

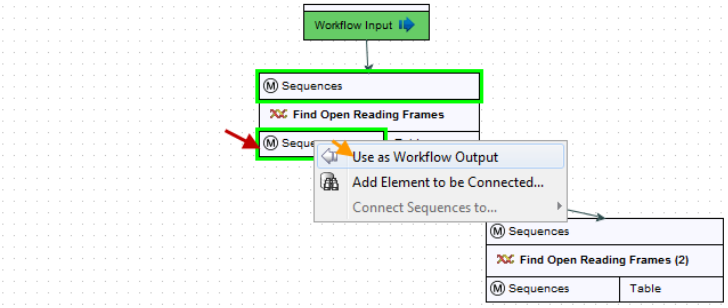


Figure 8.25: *A workflow output element can be added when more than one input modifying tool is used succeedingly (despite that the workflow only contains input modifying tools). Select "Use as Workflow Output" to make a copy of the output.*

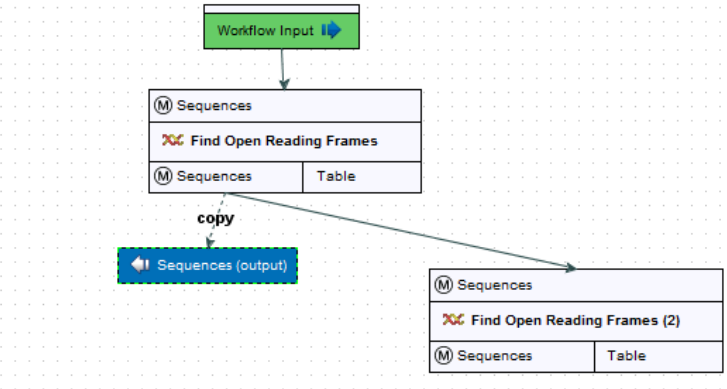


Figure 8.26: *A workflow output element can be added when more than one input modifying tool is used succeedingly (despite that the workflow only contains input modifying tools). Note that this output is marked with "copy" to indicate that this is a copy of the result that is used as input at the next level in the chain.*

8.1.8 Workflow validation

At the bottom of the view, there is a text with a status of the workflow (see figure 8.27). It will inform about the actions you need to take to finalize the workflow.

The validation may contain several lines of text. Scroll the list to see more lines. If one of the

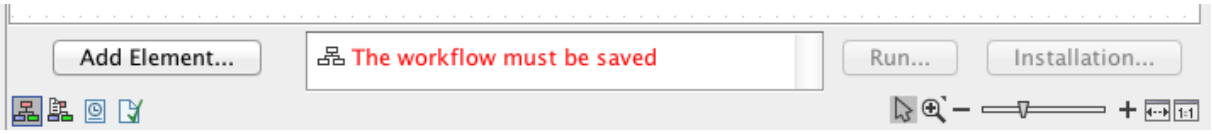


Figure 8.27: *A workflow is constantly validated at the bottom of the view.*

errors pertain to a specific element in the workflow, clicking the error will highlight this element.

The following needs to be in place before a workflow can be executed:

- All input boxes need to be connected either to the workflow input or to the output of other tools.
- At least one output box from each tool needs to be connected to either a workflow output or to the input box of another tool.
- Additional checks that the workflow is consistent.

Once these conditions are fulfilled, the status will be "Validation successful", the **Run** button is enabled. Clicking this button will enable you to try running a data set through the workflow to test that it produces the expected results. If reference data has not been configured (see section 8.1.2), there will be a dialog asking for this as part of the test run.

8.1.9 Workflow creation helper tools

In the workflow editor **Side Panel**, you will find the following workflow display settings that can be useful to know (figure 8.28):

Grid

- Enable grid You can display a grid and control the spacing and color of the grid. Per default, the grid is shown, and the workflow elements snap to the grid when they are moved around.

View mode

- Collapsed The elements of the workflow can be collapsed to allow a cleaner view and especially for large workflows this can be useful.
- Highlight used elements Ticking **Highlight used elements** (or using the shortcut Alt + Shift + U) will show all elements that are used in the workflow whereas unused elements are grayed out.
- Rulers Vertical and horizontal rules can be visualised
- Auto Layout Ticking **Auto Layout** will ensure rearrangement of elements once new elements are added.
- Connections to background Connecting arrows are shown behind elements. This may easy reading of element names and accessible parameters.

Design

- Round elements Enable rounding of the element boxes.

- Show shadow Shadows of element boxes can be added.
- Configured elements Background color can be customized.
- Inpupt elements Background color can be customized.
- Edges Color of connecting arrows can be customized.

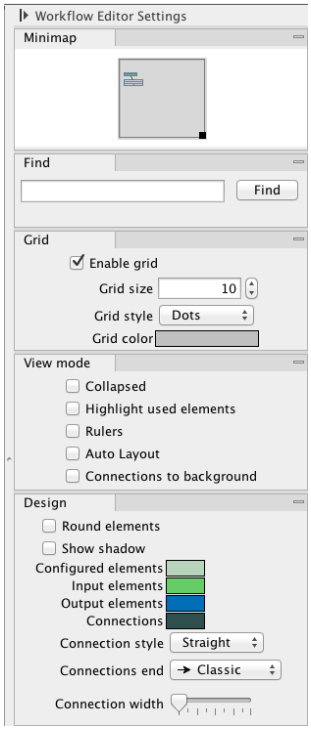


Figure 8.28: *The Side Panel of the workflow editor.*

8.1.10 Adding to workflows

Additional elements can be added to an already existing workflow by dragging it from the navigation area into the workflow editor and joining more elements as necessary. The new workflow must be saved and validated before it can be executed. Two or more workflows can be joined by dragging and dropping one from the Navigation Area, into another that is already open in the main viewing area. The output of one must be connected to the input of the next to allow the whole workflow to run in one go.

Workflows do not need to be valid to be dragged in to the workflow editor, but they must have been migrated to the current version of the workbench.

8.1.11 Snippets in workflows

When creating a new workflow, you will often have a number of connected elements that are shared between workflows. Instead of building workflows from scratch it is possible to reuse components of an existing workflow. These components are called snippets and can exist of e.g. a read mapper and a variant caller.

Snippets can be created from an existing workflow by selecting the elements and the arrows connecting the selected elements. Next, you must right-click in the center of one of the selected elements. This will bring up the menu shown in figure 8.29.

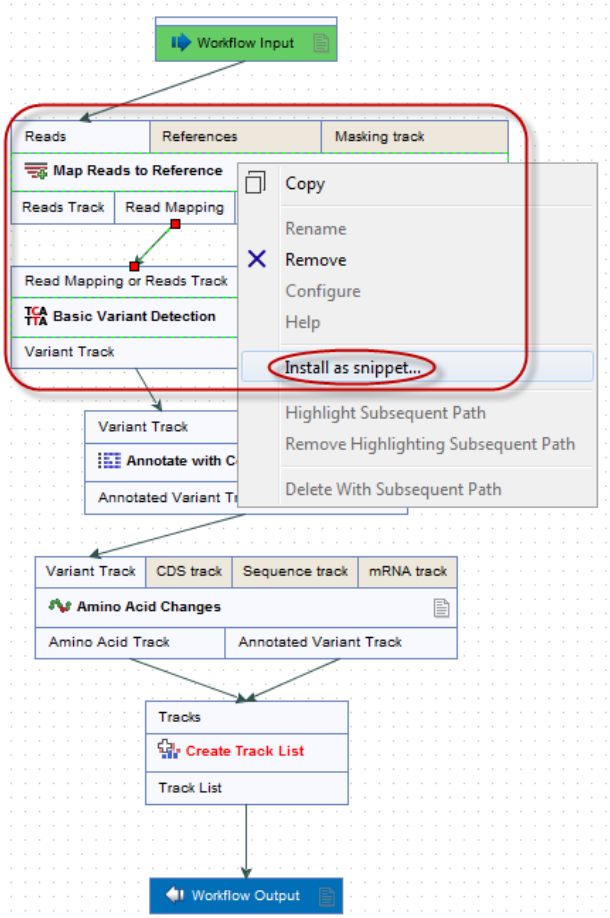


Figure 8.29: *The selected elements are highlighted with a red box in this figure. Select "Install as snippet".*

When you have clicked on "Install as snippet" the dialog shown in figure 8.30 will appear. The dialog allows you to name the snippet and view the selected elements that are included in the snippet. You are also asked to specify whether or not you want to include the configuration of the selected elements and save it in the snippet or to only save the elements in their default configuration.

Click on the button labeled **OK**. This will install your snippet and the installed snippet will now appear in the **Side Panel** under the "Snippets" tab (see figure 8.31)

Right-clicking on the installed snippet in the **Side Panel** will bring up the following options (figure 8.32):

- **Add** Adds the snippet to the current open workflow
- **View** Opens a dialog showing the snippet, which allows you to see the structure
- **Rename** Allows renaming of the snippet.
- **Configure** Allows to change the configuration of the installed snippet.
- **Uninstall** Removes the snippet.
- **Export** Exports the snippet to ones computer, allowing to share it.

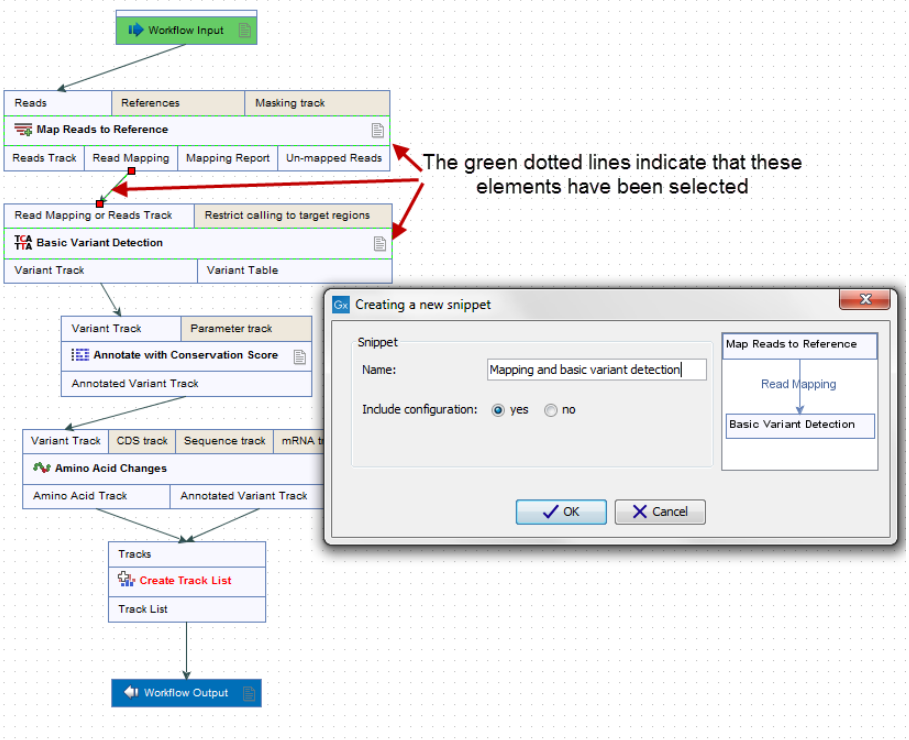


Figure 8.30: *In the "Create a new snippet" dialog you can name the snippet and select whether or not you would like to include the configuration. In the right-hand side of the dialog you can see the elements that are included in the snippet.*

- **Migrate** Migrates the snippet (if migration is required).

If you right-click on the top-level folder you get the options shown in figure 8.33:

- **Create new group** Creates a new folder under the selected folder.
- **Remove group** Removes the selected group (not available for the top-level folder)
- **Rename group** Renames the selected group (not available for the top-level folder)

In the **Side Panel** it is possible to drag and drop a snippet between groups to be able to rearrange and order the snippets as desired. An exported snippet can either be installed by clicking on the 'Install from file' button or by dragging and dropping the exported file directly into the folder where it should be installed.

Add a snippet to a workflow Snippets can be added to a workflow in two different ways; It can either be added by dragging and dropping the snippet from the **Side Panel** into the workflow editor, or it can be added by using the "Add elements" option that is shown in figure 8.34.

8.1.12 Change the order of tracks in the Genome Browser View

When modifying an existing workflow or creating a custom workflow that include the tool "**Create New Genome Browser View**" you may want to be able to adjust the order in which the tracks are

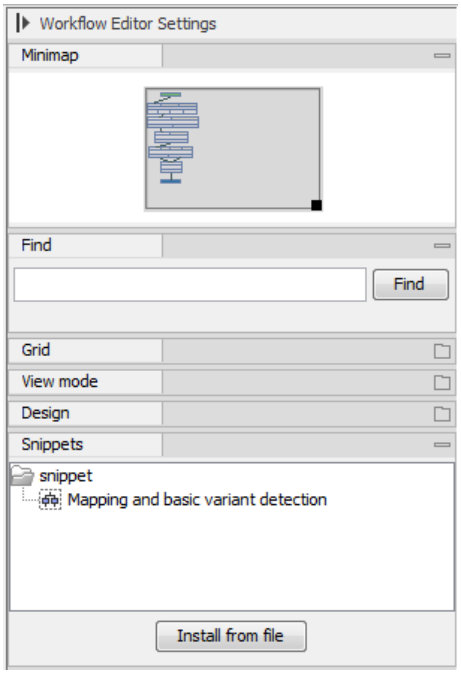


Figure 8.31: *When a snippet is installed, it appears in the Side Panel under the "Snippets" tab.*

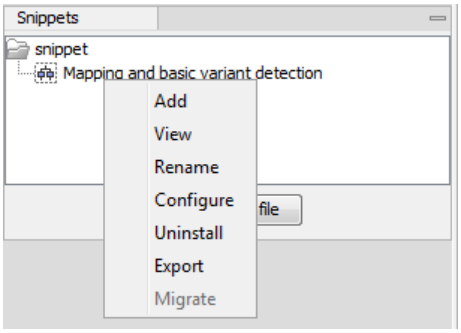


Figure 8.32: *Right-clicking on an installed snippet brings up a range of different options.*

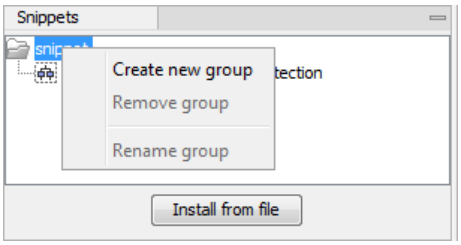


Figure 8.33: *Right-clicking on the snippet top-level folder makes it possible to manipulate the groups.*

shown in the Genome Browser View. To do this, display a view of the workflow layout, click once on the top part of the tool "Create New Genome Browser View" labeled "Tracks" followed by a right-click. In the pop up menu that appears (figure 8.35), choose the option "Order Workflow Inputs".

This opens a new pop up window (figure 8.36 where you can see a list of all the inputs that are connected with the input channel of the "Create New Genome Browser View" tool. Use the arrows found in the left-hand side to move the tracks up or down until you have the desired track order in your Genome Browser View.

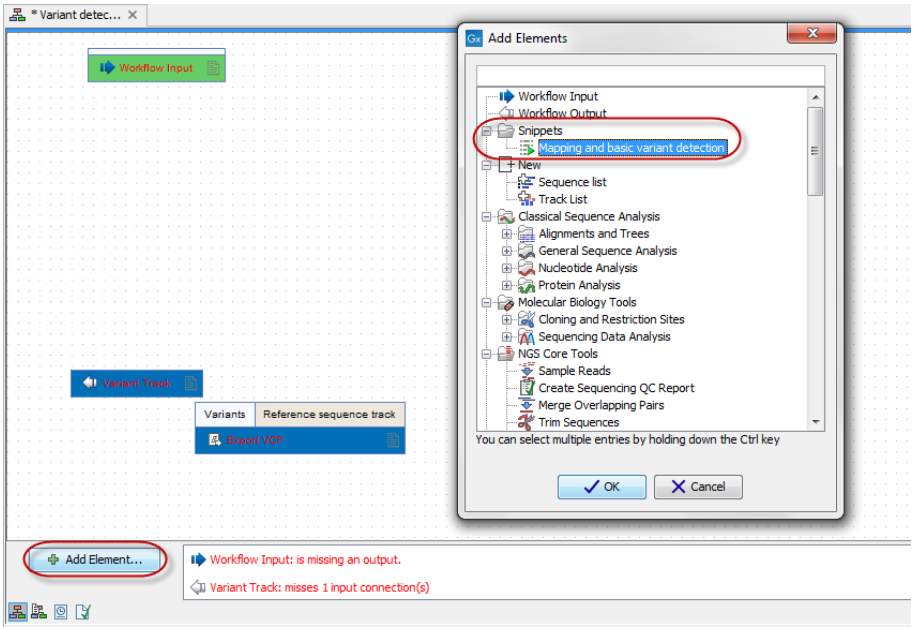


Figure 8.34: *Snippets can be added to a workflow in the workflow editor using the 'Add Elements' button found in the lower left corner.*

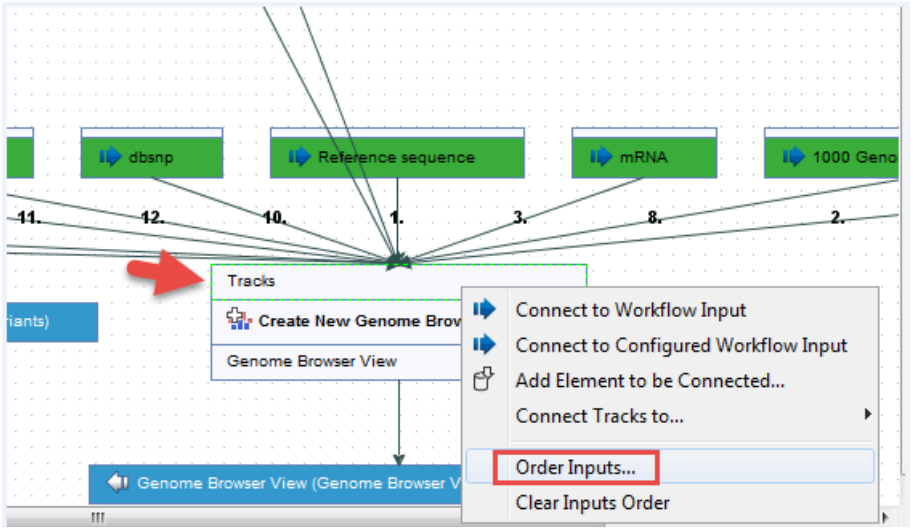


Figure 8.35: *Right click on the workflow layout and choose the option "Order Inputs...".*

If the workflow also generated several variant tracks, the variant table generated from the uppermost read mapping will open in split view with the Genome Browser View. By changing the Order of Inputs you can thus also influence which variant table should open in split view.

8.2 Distributing and installing workflows

Once the workflow has been configured, you can use the **Run** button (see section 8.1.8) to process data through the workflow, but the real power of the workflow is its ability to be distributed and installed in the **Toolbox** alongside the other tools that come with the *CLC Drug Discovery Workbench*, as well as the ability to install the same workflow on a *CLC Genomics Server*. The mechanism for distributing the workflow is a workflow installer file which can be created from the workflow editor and distributed and installed in any Workbench or Server.

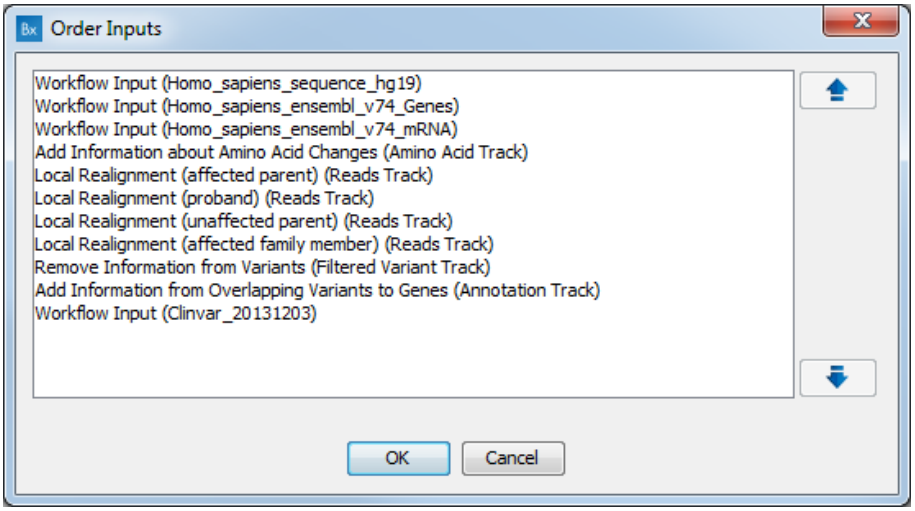


Figure 8.36: *Example of Order Inputs window that appears when choosing the option "Order Inputs...".*

8.2.1 Creating a workflow installation file

At the bottom of the workflow editor, click the **Create Installer** button (or use the shortcut Shift + Alt + I) to bring up a dialog where you provide information about the workflow to be distributed (see an example with information from a CLC bio workflow in figure 8.37).

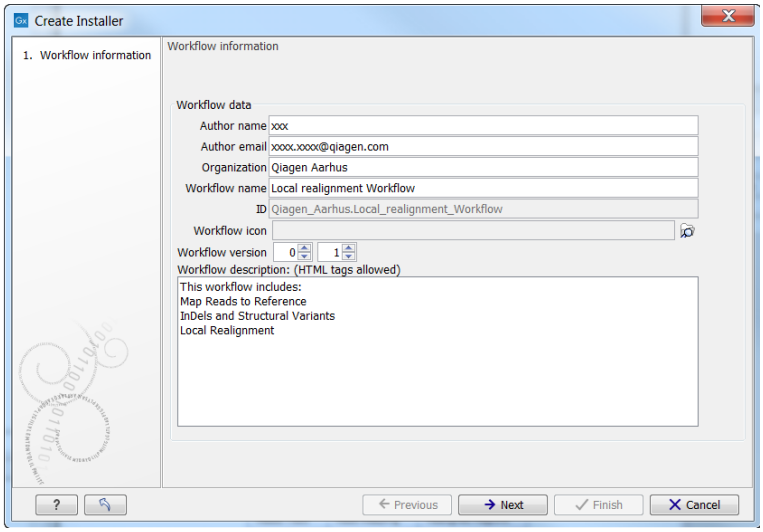


Figure 8.37: *Workflow information for the installer.*

Author information Providing name, email and organization of the author of the workflow. This will be visible for users installing the workflow and will enable them to look up the source of the workflow any time. The organization name is important because it is part of the workflow id (see more in section 8.2.4)

Workflow name The workflow name is based on the name used when saving the workflow in the **Navigation Area**. The workflow name is essential because it is used as part of the workflow id (see more in section 8.2.4). The workflow name can be changed during the installation of the workflow. This is useful whenever you have a workflow that you would like to use e.g.

with small variations. The original workflow name will remain the same in the **Navigation Area** - only the installed workflow will receive the customized name.

ID The final id of the workflow.

Workflow icon An icon can be provided. This will show up in the installation overview and in the **Toolbox** once the workflow is installed. The icon should be a 16 x 16 pixels gif or png file. If the icon is larger, it will automatically be resized to fit 16 x 16 pixels.

Workflow version A major and minor version can be provided.

Include original workflow file This will include the design file to be included with the installer. Once the workflow is installed in a workbench, you can extract the original workflow file and modify it.

Workflow description Provide a textual description of the workflow. This information will be displayed when a user mouses-over the name of the installed Workflow in the Workbench Toolbox, and is also presented in the Description tab for that Workflow in the Manage Workflows tool, described in section 8.2.3. Simple HTML tags are allowed (should be HTML 3.1 compatible, see http://www.w3.org/TR/REC-html32).

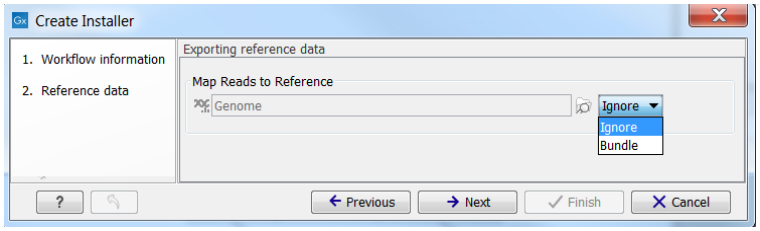


Figure 8.38: *Bundling data with the workflow installer.*

If you configured any of the workflow elements with data, clicking **Next** will give you the following options (see figure 8.38):

- **Ignore** This will exclude these reference data from the workflow.
- **Bundle** Includes data in the workflow by bundling the reference data with the workflow. **Note!** Bundling data should only be used to bundle small data sets with the workflow installer.

If you configured any of the workflow elements with data, clicking **Next** will give you the following options (see figure 8.39):

- **Ignore** This will exclude these reference data from the workflow.
- **Reference** This option can be used to include reference data from a shared CLC_References directory in a workflow without bundling the reference data with the workflow. Instead the reference data is included in the workflow by pointing at the shared CLC_References directory. This is particularly useful when working with large reference data.
- **Bundle** Includes data in the workflow by bundling the reference data with the workflow. **Note!** Bundling data should only be used to bundle small data sets with the workflow installer.

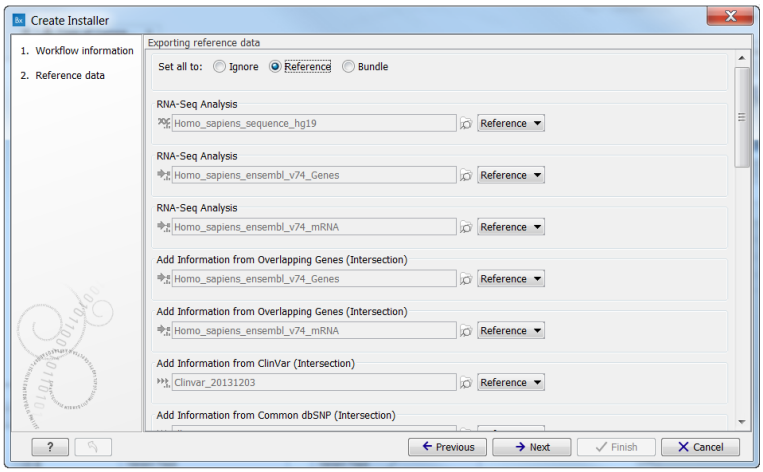


Figure 8.39: *Bundling data with the workflow installer.*

Click **Next** and you will be asked to specify where to install the workflow (figure 8.40). You can install your workflow directly on your local computer. If you are logged on a server and are the administrator, the option "Install the workflow on the current server" will be enabled. Finally, you can select to save the workflow as a .cpw file that can be installed on another computer. Click **Finish**. This will install the workflow directly on the selected destination. If you have selected to save the workflow for installation on another computer, you will be asked where to save the file after clicking **Finish**. If you chose to bundle data with your workflow installation, you will be asked for a location to put the bundled data on the workbench. Installing a workflow with bundled data on a server, the data will be put in a folder created in the first writeable persistence location. Should this location not suit your needs, you can always move it afterwards, using the normal persistence operations.

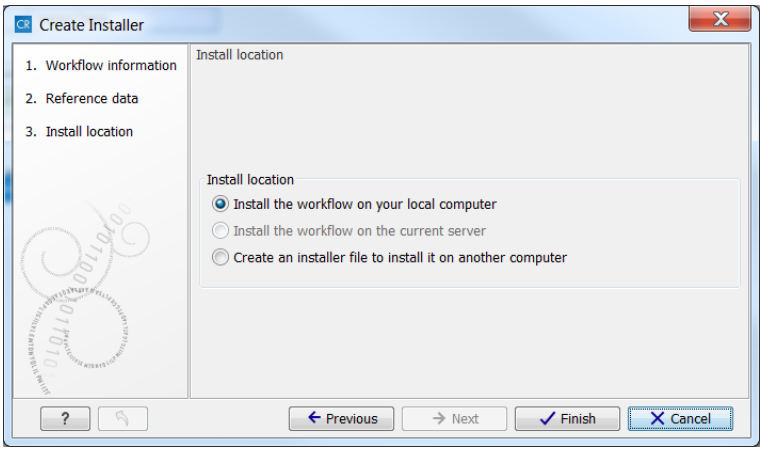


Figure 8.40: *Select whether the workflow should be installed on your local computer or on the current server. A third option is to create an installer file (.cpw) that can be installed on another computer.*

In cases where an existing workflow, that has already been installed, is modified, the workflow must be reinstalled. This can be done by first saving the workflow after it has been modified and then pressing the **Create Installer** button. Click through the wizard and select whether you wish to install the modified workflow on your local computer or on a server. Press **Finish**. This will open a pop-up dialog "Workflow is already installed" (figure 8.41) with the option that you can force the installation. This will uninstall the existing workflow and install the modified version of

the workflow. **Note!** When forcing installation of the modified workflow, the configuration of the original workflow will be lost.

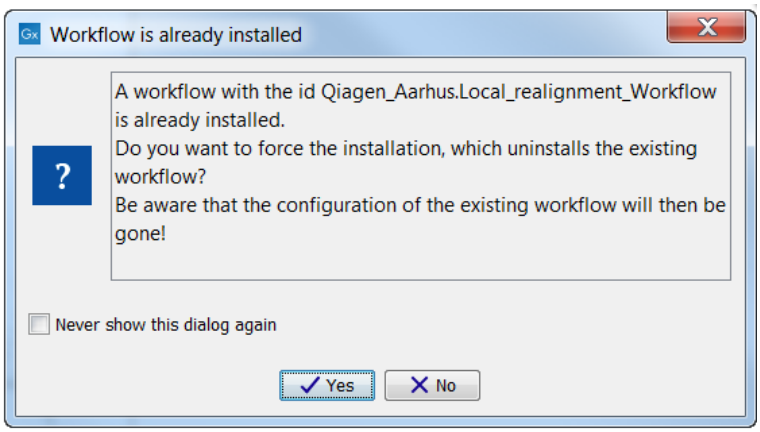


Figure 8.41: *Select whether you wish to force the installation of the workflow or keep the original workflow.*

8.2.2 Installing a workflow

Workflows are installed in the workflow manager (for information about installing a workflow on the *CLC Genomics Server*, please see the user manual at `http://www.clcbio.com/usermanuals`):

Help | **Manage Workflows ( )**

or press the "Workflows" button ( ) in the toolbar and then select "Manage Workflow..." ( ).

This will display a dialog listing the installed workflows. To install an existing workflow, click **Install from File** and select a workflow .cpw file .

Once installed, it will appear in the workflow manager as shown in figure 8.42.

If the workflow was bundled with data, installing it on the workbench will ask you for a location to put the bundled data. Installing a workflow with bundled data on a server, the data will be put in a folder created in the first writeable persistence location.

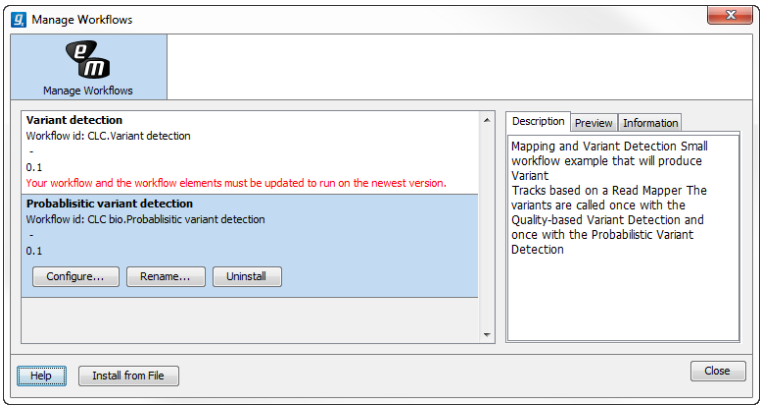


Figure 8.42: *Workflows available in the workflow manager. Note the alert on the "Variant detection" workflow, that means that this workflow needs to be updated.*

8.2.3 Managing workflows

Workflows can be managed from the workflow manager:

Help | Manage Workflows ( )

or using the "Workflows" button ( ) in the toolbar and then select "Manage Workflow..." ( ).

The workflow manager lists Custom workflows and Ready-to-Use workflows, but the functionalities described below (Configure, Rename, and Uninstall) are only available to custom workflows. You can always create a copy of a Ready-to-Use workflow (by opening the Ready-to-Use workflow and saving a copy in your Navigation Area) to enable the options described below.

Configure Select the workflow of interest and click on the button labeled Configure. You will be presented with a dialog listing all the reference data that need to be selected. An example is shown in figure 8.43.

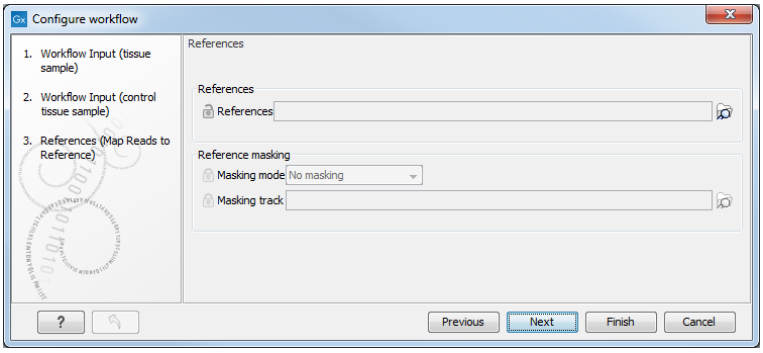


Figure 8.43: *Configuring parameters for the workflow.*

This dialog also allows you to lock parameters of the workflow (see more about locking in section 8.1.3).

Note that if the workflow is intended to be executed on a server, it is important to select reference data that is located on the server.

Rename In addition to the configuration option, it is also possible to rename the workflow. This will change the name of the workflow in the **Toolbox**. The workflow id (see below) remains the same. To rename an element right click on the element name in the Navigation Area and select "Rename" or click on the F2 button.

Uninstall Use this button to install a workflow.

Description, Preview and Information In the right side of the window, you will find three tabs. **Description** contains the description that was entered when creating the workflow installer (see figure 8.37), the **Preview** shows a graphical representation of the workflow (figure 8.44), and finally you can get **Information** about the workflow (figure 8.45).

The "Information" field (figure 8.45) contains the following:

Build id The date followed by the time

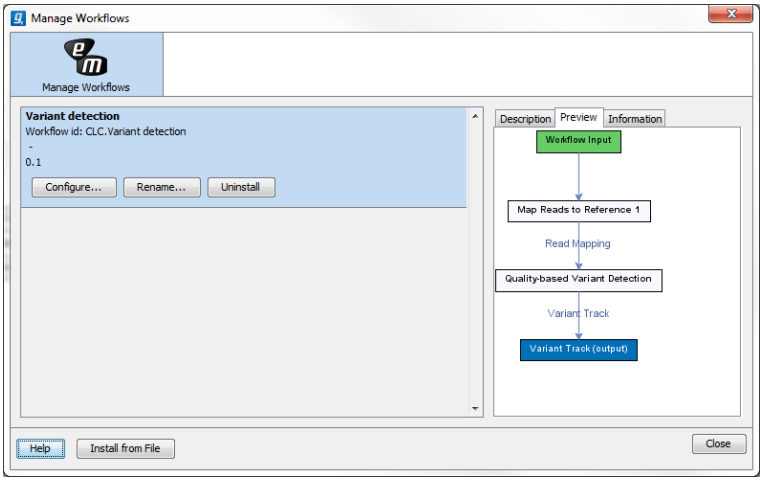


Figure 8.44: *Preview of the workflow.*

Download href The name of the workflow .cpw file

Id The unique id of a workflow, by which the workflow is identified

Major version The major version of the workflow

Minor version The minor version of the workflow

Name Name of workflow

Rev version Revision version. The functionality is activated but currently not in use

Vendor id ID of vendor that has created the workflow

Version <Major version>.<Minor version>

Workbench api version Workbench version

Workflow api version Workflow version (a technical number that can be used for troubleshooting)

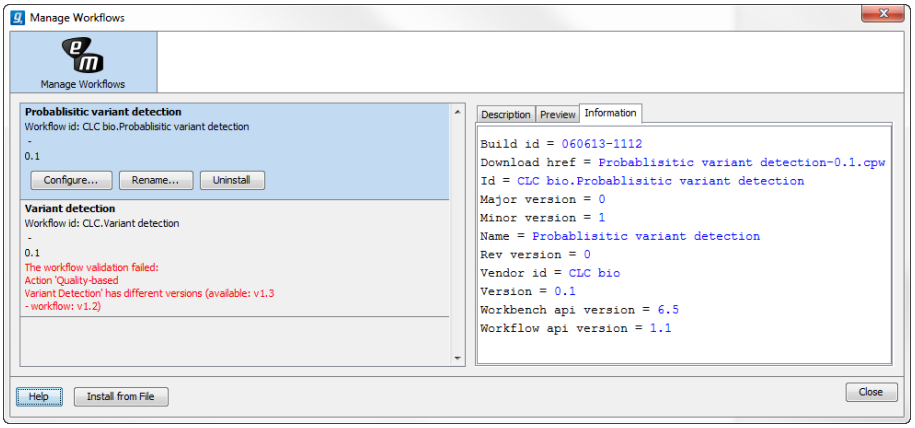


Figure 8.45: *With "Manage Workflows" it is possible to configure, rename and uninstall workflows.*

8.2.4 Workflow identification and versioning

A workflow has a version. The version is used to make it easy to distribute an improved version of the same workflow. To do this, create a new installer with an incremented version number. In order to install a new and updated version, the old one has to be uninstalled.

The way the *CLC Drug Discovery Workbench* checks whether a workflow already exists in a previous version is by looking at the workflow id. The id is a combination of the organization name and the name of the workflow itself as it is shown in the dialog shown in figure 8.37. Once installed this information is also available in the workflow manager (in figure 8.42 this is `CLC bio.Simple variant detection and annotation-1.2`).

If you create two different workflows with the same name and using the same organization name when creating the installer, they cannot both be installed.

8.2.5 Automatic update of workflow elements

When new versions of the *CLC Drug Discovery Workbench* are released, some of the tools that are part of a workflow may change. When this happens, the workflow may no longer be valid. This will happen both to the workflow configurations saved in the **Navigation Area** and the installed workflows.

When a workflow is opened from the **Navigation Area**, an editor will appear, if tools used in the workflow have been updated (see figure 8.46).

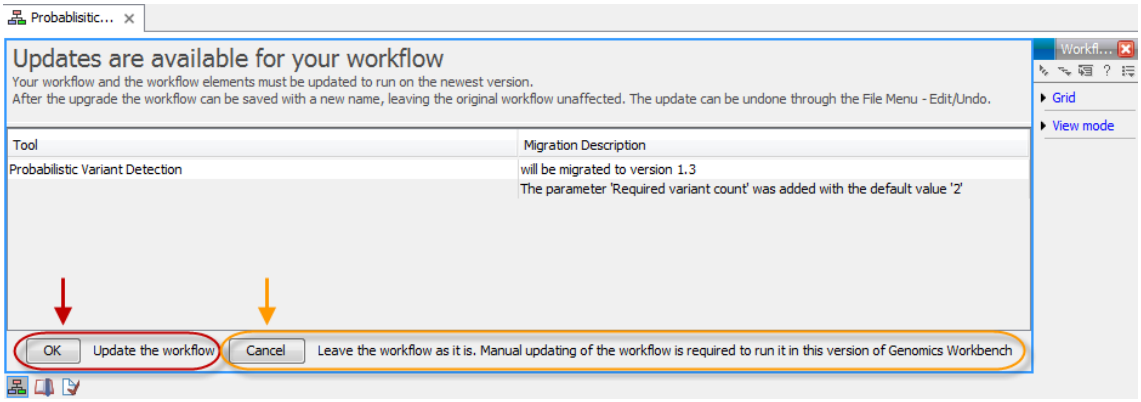


Figure 8.46: *When updates are available an editor appears with information about which tools should be updated. Press "OK" to update the workflow. The workflow must be updated to be able to run the workflow on the newest version of the Workbench.*

Updating a workflow means that the tools in your workflow is updated with the most recent version of these particular tools. To update your workflow, press the **OK** button at the bottom of the page.

There may be situations where it is important for you to keep the workflow in its original form. This could be the case if you have used a workflow to generate results for a publication. In such cases it may be necessary for you to be able to go back to the original workflow to e.g. repeat an analysis.

You have two options to keep the old workflow:

- If you do not wish to update the workflow at all, press the **Cancel** button. This will keep

the workflow unchanged. However, the next time you open the workflow, you will again be asked whether you wish to update the workflow. Please note that only updated workflows can run on the newest versions of the Workbench.

- Another option is to update the workflow and save the updated workflow with a new name. This will ensure that the old workflow is kept rather than being overwritten.

Note! In cases where new parameters have been added, these will be used with their default settings.

If you have used the toolbar "Workflow" button ( ) and "Manage Workflow..." ( ) to access a specific workflow in order to e.g. change the workflow configuration or are going to use the "Install from File" function, a button labeled "Update..." will appear whenever tools have been changed and the workflow needs to be updated (figure 8.47). When you click the button labeled "Update...", your workflow will be updated and the existing workflow will be overwritten.

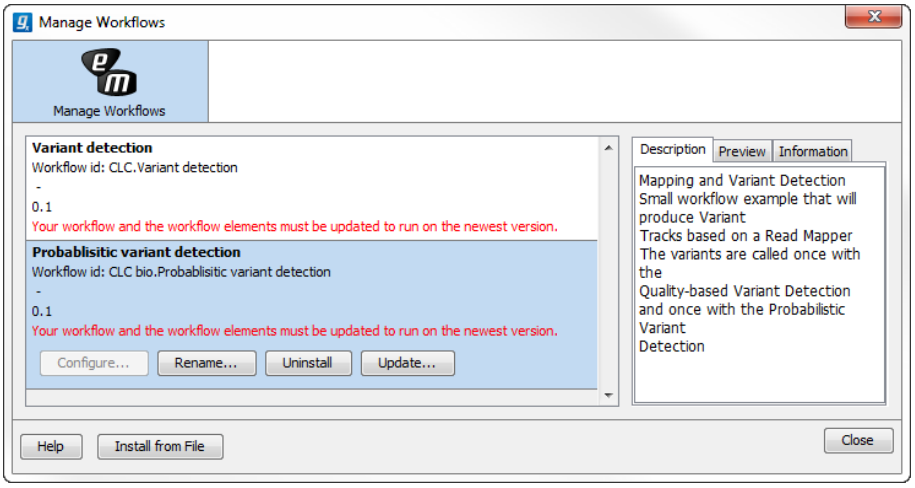


Figure 8.47: *Workflow migration.*

8.3 Executing a workflow

Once installed and configured, a workflow will appear in the **Toolbox** under **Workflows** ( ). If an icon was provided with the workflow installer this will also be shown (see figure 8.48).

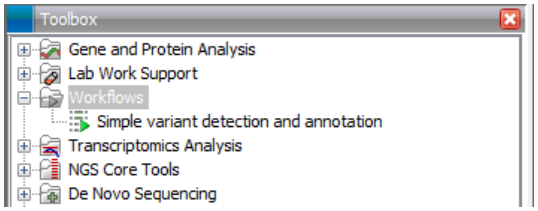


Figure 8.48: *A workflow is installed and ready to be used.*

The workflow is executed just as any other tool in the **Toolbox** by double-clicking or selecting it in the menu (or with the shortcut Ctrl + Enter). This will open a dialog where you provide input data and with options to run the workflow in batch mode (see section 7.3). In the last page of the dialog, you can preview all the parameters of the workflow, as well as the input data, before clicking "Next" to choose where to save the output, and then "Finish" to execute the workflow.

If you are connected to a *CLC Genomics Server*, you will be presented with the option to run the workflow locally on the Workbench or on the Server. When you are selecting where to run the workflow, you should also see a message should there be any missing configurations. There are more details about running Workflows on the Server in the Server manual (http://www.clcsupport.com/clcgenomicsserver/current/admin/index.php?manual=Workflows.html).

When the workflow is started, you can see the log file with detailed information from each step in the process.

If the workflow is not properly configured, you will see that in the dialog when the workflow is started [2].

8.4 Open copy of installed workflow

A copy of an installed and configured workflow found in the **Toolbox** under **Workflows** ( ) can be opened in the View Area by clicking once and then right-clicking on the name of the installed workflow in the toolbox (figure 8.49).

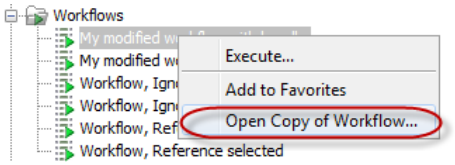


Figure 8.49: *A copy of an installed workflow can be opened from the Toolbox. The copied workflow will open in the View Area.*

An example of a copy of a workflow that has been opened in the **View Area** is shown in figure 8.50.

[2] If the workflow uses a tool that is part of a plugin, a missing plugin can also be the reason why the workflow is not enabled. A workflow can also become outdated because the underlying tools have changed since the workflow was created (see section 8.2.4)

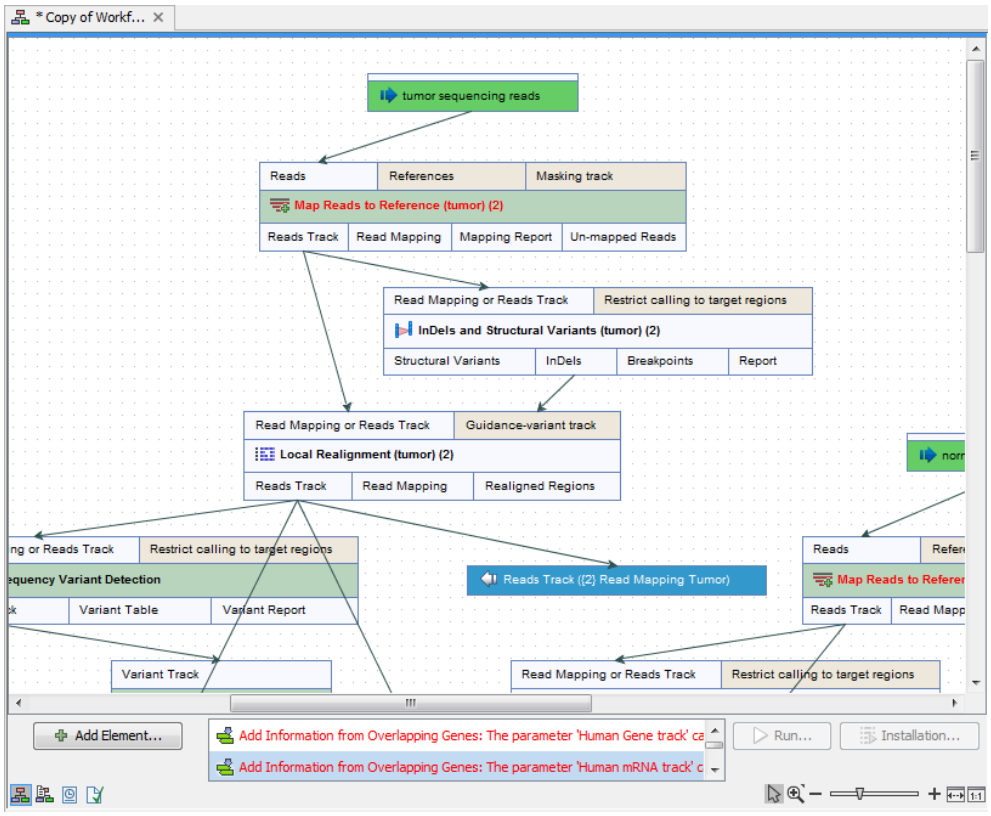


Figure 8.50: *A copy of an installed workflow after it has been opened in the View Area.*

Part III

Molecular modeling and sequence analysis

Chapter 9

Drug design

Contents

In this chapter you will find a description of the different options the *CLC Drug Discovery Workbench* offers for studying, visualizing, and manipulating molecule structures and properties as well as modeling protein-ligand complexes.

Basically, there are two ways to work with your molecules; you can either look at the 3D structure of the molecules in a **Molecule Project** or in table format in a **Molecule Table**.

The **Molecule Project** is accompanied by the **Project Tree**, which is found in the Side Panel. The **Project Tree** is a tree view of all molecules available in the 3D view, and this tree view gives access to a range of functionalities that can be used when visualizing and studying the molecules.

Molecule Tables can be used to store a large number of molecules together, like e.g. a screening library. The table can be connected to a **Molecule Project** view, to inspect and manipulate individual molecules in 3D.

Molecule structures can be imported in seven different ways, as described in section 6.2.

Tutorials are available that can introduce you to the workbench and guide you through the main applications. The tutorials can be found from the menu bar **Help | Tutorials** or downloaded in pdf-version from the web `http://www.clcbio.com/tutorials/#-drug-discovery`.

9.1 Viewing molecular structures in 3D

The **Molecule Project** editor has a range of different options that allow for custom visualization of the molecules in the 3D view. An example of molecules that have been opened as a **Molecule Project** is shown in figure 9.1.

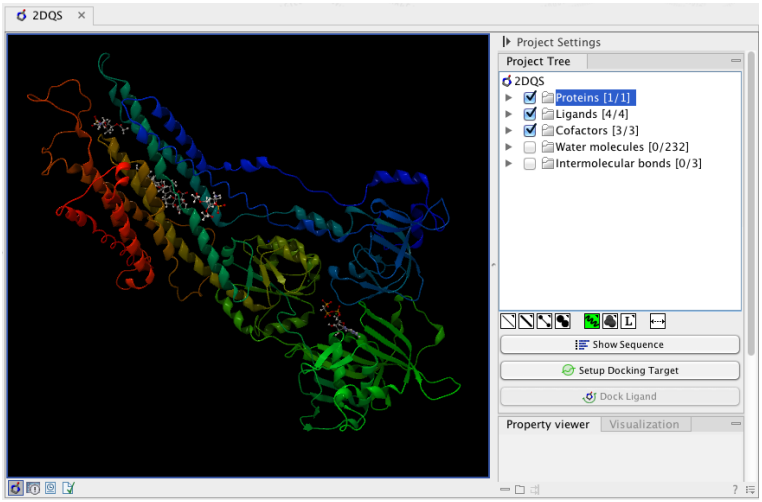


Figure 9.1: *3D view of a calcium ATPase. All molecules are shown in the Molecule Project. The Project Tree in the Side Panel lists the involved molecules.*

Molecules in **Molecule Tables** can also be viewed in 3D by connecting the table to a **Molecule Project** editor from the Select View button in the **Molecule Table** Side Panel (see section 9.7.3).

The following sections will address how to use the **Molecule Project** editor for visualizing molecules. Section 9.7 will address how to visualize molecules in **Molecule Tables**.

9.1.1 Moving and rotating

The molecules can be rotated by holding down the left mouse button while moving the mouse. The right mouse button can be used to move the view.

Zooming can be done with the scroll-wheel or by holding down both left and right buttons while moving the mouse up and down.

All molecules in the **Molecule Project** are listed in categories in the **Project Tree**. The individual molecules or whole categories can be hidden from the view by un-cheking the boxes next to them.

It is possible to bring a particular molecule or a category of molecules into focus by selecting the molecule or category of interest in the **Project Tree** view and double-click on the molecule or category of interest. Another option is to use the zoom-to-fit button (⟷) at the bottom of the **Project Tree** view.

9.1.2 Troubleshooting 3D graphics errors

The 3D viewer uses OpenGL graphics hardware acceleration in order to provide the best possible experience. If you experience any graphics problems with the 3D view, please make sure that the drivers for your graphics card are up-to-date.

If the problems persist after upgrading the graphics card drivers, it is possible to change to a

rendering mode, which is compatible with a wider range of graphic cards. To change the graphics mode go to Edit in the menu bar, select "Preferences", Click on "View", scroll down to the bottom and find "Molecule Project 3D Editor" and uncheck the box "Use modern OpenGL rendering".

Finally, it should be noted that certain types of visualization are more demanding than others. In particular, using multiple molecular surfaces may result in slower drawing, and even result in the graphics card running out of available memory. Consider creating a single combined surface (by using a selection) instead of creating surfaces for each single object. For molecules with a large number of atoms, changing to wireframe rendering and hiding hydrogen atoms can also greatly improve drawing speed.

9.2 Customizing the visualization

The molecular visualization of all molecules in the Molecule Project can be customized using different visualization styles. The styles can be applied to one molecule at a time, or to a whole category (or a mixture), by selecting the name of either the molecule or the category. Holding down the Ctrl (Cmd on Mac) or shift key while clicking the entry names in the **Project Tree** will select multiple molecules/categories.

The six leftmost quick-style buttons below the **Project Tree** view give access to the molecule visualization styles, while context menus on the buttons (accessible via right-click or left-click-hold) give access to the color schemes available for the visualization styles. Visualization styles and color schemes are also available from context menus directly on the selected entries in the **Project Tree**. Other quick-style buttons are available for displaying hydrogen bonds between Project Tree entries, for displaying labels in the 3D view and for creating custom atom groups. They are all described in detail below.

Note! Whenever you wish to change the visualization styles by right-clicking the entries in the **Project Tree**, please be aware that you must first click on the entry of interest, and ensure it is highlighted in blue, before right-clicking.

9.2.1 Visualization styles and colors

Wireframe, Stick, Ball and stick, Space-filling/CPK

() () () ()

Four different ways of visualizing molecules by showing all atoms are provided: Wireframe, Stick, Ball and stick, and Space-filling/CPK.

The visualizations are mutually exclusive meaning that only one style can be applied at a time for each selected molecule or atom group.

Six color schemes are available and can be accessed via right-clicking on the quick-style buttons:

- Color by Element. Classic CPK coloring based on atom type (e.g. oxygen red, carbon gray, hydrogen white, nitrogen blue, sulfur yellow).
- Color by Temperature. For PDB files, this is based on the b-factors. For structure models created with tools in a CLC workbench, this is based on an estimate of the local model quality. The color scale goes from blue (0) over white (50) to red (100). The b-factors as

well as the local model quality estimate are measures of uncertainty or disorder in the atom position; the higher the number, the higher the uncertainty.

- Color Carbons by Entry. Each entry (molecule or atom group) is assigned its own specific color. Only carbon atoms are colored by the specific color, other atoms are colored by element.
- Color by Entry. Each entry (molecule or atom group) is assigned its own specific color.
- Custom Color. The user selects a molecule color from a palette.
- Custom Carbon Color. The user selects a molecule color from a palette. Only carbon atoms are colored by the specific color, other atoms are colored by element.

Backbone

()

For the molecules in the Proteins and Nucleic Acids categories, the backbone structure can be visualized in a schematic rendering, highlighting the secondary structure elements for proteins and matching base pairs for nucleic acids. The backbone visualization can be combined with any of the atom-level visualizations.

Five color schemes are available for backbone structures:

- Color by Residue Position. Rainbow color scale going from blue over green to yellow and red, following the residue number.
- Color by Type. For proteins, beta sheets are blue, helices red and loops/coil gray. For nucleic acids backbone ribbons are white while the individual nucleotides are indicated in green (T/U), red (A), yellow (G), and blue (C).
- Color by Backbone Temperature. For PDB files, this is based on the b-factors for the Cα atoms (the central carbon atom in each amino acid). For structure models created with tools in the workbench, this is based on an estimate of the local model quality. The color scale goes from blue (0) over white (50) to red (100). The b-factors as well as the local model quality estimate are measures of uncertainty or disorder in the atom position; the higher the number, the higher the uncertainty.
- Color by Entry. Each chain/molecule is assigned its own specific color.
- Custom Color. The user selects a molecule color from a palette.

Surfaces

()

Molecular surfaces can be visualized.

Five color schemes are available for surfaces:

- Color by Charge. Charged amino acids close to the surface will show as red (negative) or blue (positive) areas on the surface, with a color gradient that depends on the distance of the charged atom to the surface.

- Color by Element. Smoothed out coloring based on the classic CPK coloring of the heteroatoms close to the surface.
- Color by Temperature. Smoothed out coloring based on the temperature values assigned to atoms close to the surface (See the "Wireframe, Stick, Ball and stick, Space-filling/CPK" section above).
- Color by Entry. Each surface is assigned its own specific color.
- Custom Color. The user selects a surface color from a palette.

A surface spanning multiple molecules can be visualized by creating a custom atom group that includes all atoms from the molecules (see section 9.2.1)

It is possible to adjust the opacity of a surface by adjusting the transparency slider at the bottom of the menu.

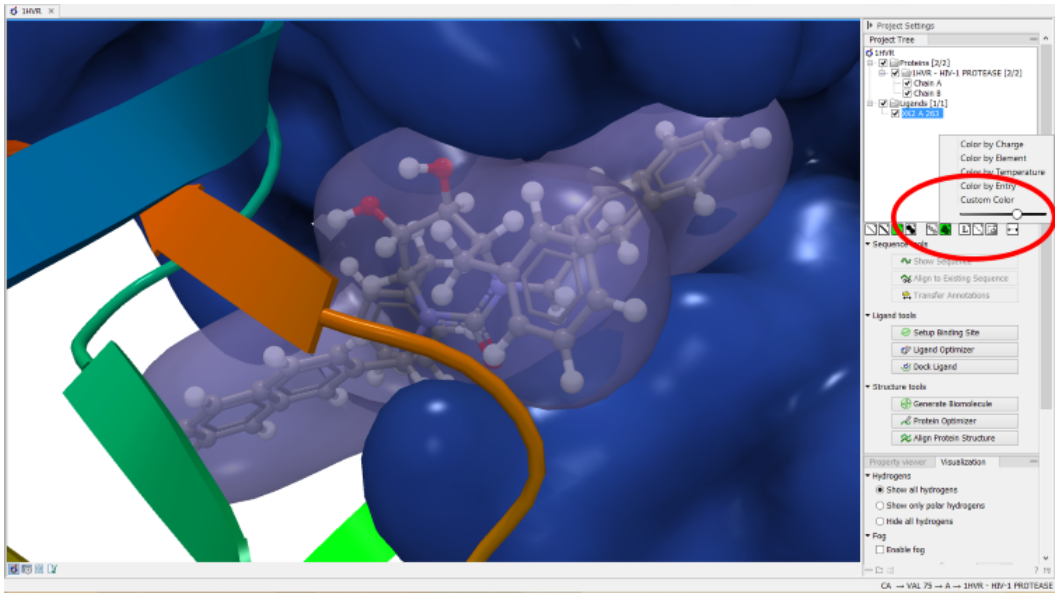


Figure 9.2: *Transparent surfaces*

Notice that visual artifacts may appear when rotating a transparent surface. These artifacts disappear as soon as the mouse is released.

Labels

(L)

Labels can be added to the molecules in the view by selecting an entry in the Project Tree and clicking the label button at the bottom of the Project Tree view. The color of the labels can be adjusted from the context menu by right clicking on the selected entry (which must be highlighted in blue first) or on the label button in the bottom of the Project Tree view (see figure 9.3).

- For proteins and nucleic acids, each residue is labelled with the PDB name and number.
- For ligands, each atom is labelled with the atom name as given in the input.

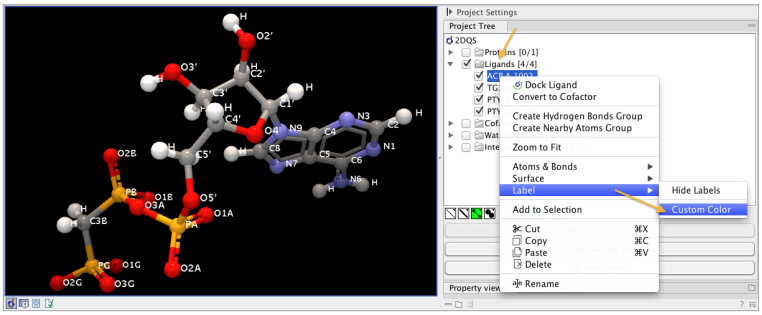


Figure 9.3: *The color of the labels can be adjusted in two different ways. Either directly using the label button by right clicking the button, or by right clicking on the molecule or category of interest in the Project Tree.*

- For cofactors and water, one label is added with the name of the molecule.
- For atom groups including protein atoms, each protein residue is labelled with the PDB name and number.
- For atom groups not including protein atoms, each atom is labelled with the atom name as given in the input.

Labels can be removed again by clicking on the label button.

Hydrogen bonds

()

The Show Hydrogen Bond visualization style may be applied to molecules and atom group entries in the project tree. If this style is enabled for a project tree entry, hydrogen bonds will be shown to all other currently visible objects. The hydrogen bonds are updated dynamically: if a molecule is toggled off, the hydrogen bonds to it will not be shown.

It is possible to customize the color of the hydrogen bonds using the context menu.

Create atom group

()

Often it is convenient to use a unique visualization style or color to highlight a particular set of atoms, or to visualize only a subset of atoms from a molecule. This can be achieved by creating an atom group. Atom groups can be created based on atoms selected in the 3D view or entries selected in the Project Tree. When an atom group has been created, it appears as an entry in the Project Tree in the category "Atom groups". The atoms can then be hidden or shown, and the visualization changed, just as for the molecule entries in the Project Tree.

Note that an atom group entry can be renamed. Select the atom group in the Project Tree and invoke the right-click context menu. Here, the Rename option is found.

Create atom group based on atoms selected in 3D view

When atoms are selected in the 3D view, brown spheres indicate which atoms are included in the selection. The selection will appear as the entry "Current" in the Selections category in the Project Tree.

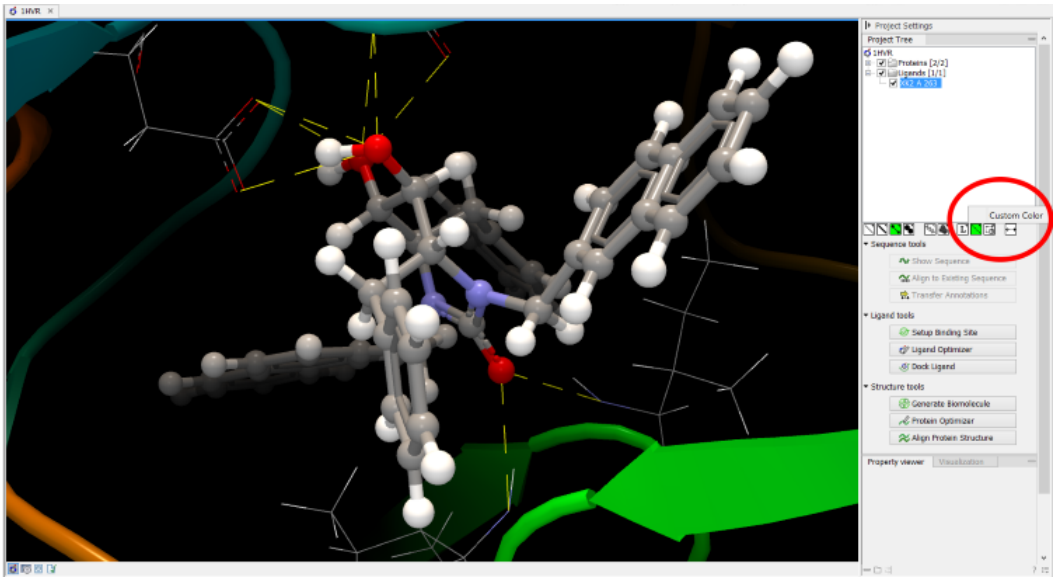


Figure 9.4: *The hydrogen bond visualization setting, with custom bond color*

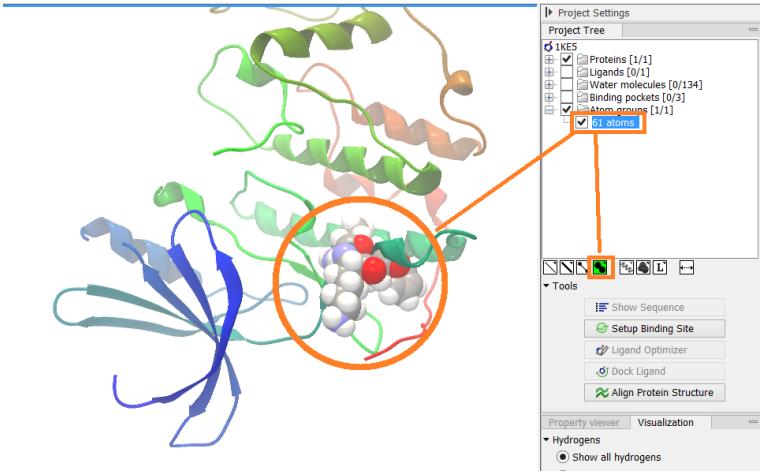


Figure 9.5: *An atom group that has been highlighted by adding a unique visualization style.*

Once a selection has been made, press the "Create Atom Group" button and a context menu will show different options for creating a new atom group based on the selection:

- Selected Atoms. Creates an atom group containing exactly the selected atoms (those indicated by brown spheres). If an entire molecule or residue is selected, this option is not displayed.
- Selected Residue(s)/Molecules. Creates an atom group that includes all atoms in the selected residues (for entries in the protein and nucleic acid categories) and molecules (for the other categories).
- Nearby Atoms. Creates an atom group that contains residues (for the protein and nucleic acid categories) and molecules (for the other categories) within 5 Å of the selected atoms. Only atoms from currently visible Project Tree entries are considered.
- Hydrogen Bonded Atoms. Creates an atom group that contains residues (for the protein and nucleic acid categories) and molecules (for the other categories) that have hydrogen

bonds to the selected atoms. Only atoms from currently visible Project Tree entries are considered.

There are several ways to select atoms in the 3D view:

- Double click to select. Click on an atom to select it. When you double click on an atom that belongs to a residue in a protein or in a nucleic acid chain, the entire residue will be selected. For small molecules, the entire molecule will be selected.
- Adding atoms to a selection. Holding down Ctrl while picking atoms, will pile up the atoms in the selection. All atoms in a molecule or category from the Project Tree, can be added to the "Current" selection by choosing "Add to Current Selection" in the context menu. Similarly, entire molecules can be removed from the current selection via the context menu.
- Spherical selection. Hold down the shift-key, click on an atom and drag the mouse away from the atom. Then a sphere centered on the atom will appear, and all atoms inside the sphere, visualized with one of the all-atom representations will be selected. The status bar (lower right corner) will show the radius of the sphere.
- Show Sequence. Another option is to select protein or nucleic acid entries in the Project Tree, and click the "Show Sequence" button found below the Project Tree (section 9.4.1). A split-view will appear with a sequence editor for each of the sequence data types (Protein, DNA, RNA) (figure 9.6). If you then select residues in the sequence view, the backbone atoms of the selected residues will show up as the "Current" selection in the 3D view and the Project Tree view. Notice that the link between the 3D view and the sequence editor is lost if either window is closed, or if the sequence is modified.
- Align to Existing Sequence. If a single protein chain is selected in the Project Tree, the "Align to Existing Sequence" button can be clicked (section 9.4.2). This links the protein sequence with a sequence or sequence alignment found in the Navigation Area. A split-view appears with a sequence alignment where the sequence of the selected protein chain is linked to the 3D structure, and atoms can be selected in the 3D view, just as for the "Show Sequence" option.

Create atom group based on entries selected in the Project Tree

Select one or more entries in the Project Tree, and press the "Create Atom Group" button, then a context menu will show different options for creating a new atom group based on the selected entries:

- Nearby Atoms. Creates an atom group that contains residues (for the protein and nucleic acid categories) and molecules (for the other categories) within 5 Å of the selected entries. Only atoms from currently visible Project Tree entries are considered. This option is also available on binding pocket entries (binding pockets can only be created in *CLC Drug Discovery Workbench*).
- Hydrogen Bonded Atoms. Creates at atom group that contains residues (for the protein and nucleic acid categories) and molecules (for the other categories) that have hydrogen bonds to the selected entries. Only atoms from currently visible Project Tree entries are considered.

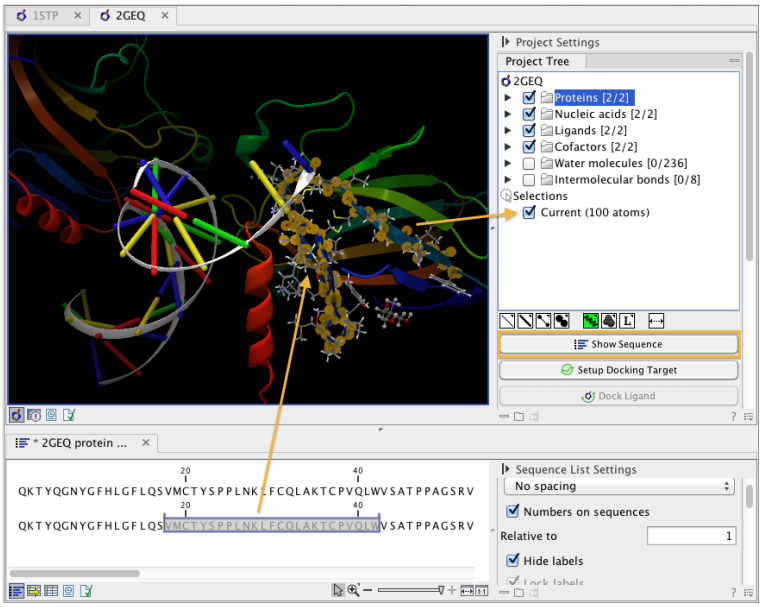


Figure 9.6: *The protein sequence in the split view is linked with the protein structure. This means that when a part of the protein sequence is selected, the same region in the protein structure will be selected.*

If a Binding Site Setup is present in the Project Tree (A Binding Site Setup can only be created using *CLC Drug Discovery Workbench*), and entries from the Ligands or Docking results categories are selected, two extra options are available under the header **Create Atom Group (Binding Site)**. For these options, atom groups are created considering all molecules included in the Binding Site Setup, and thus not taking into account which Project Tree entries are currently visible.

Zoom to fit

(⇠⇢)

The "Zoom to fit" button can be used to automatically move a region of interest into the center of the screen. This can be done by selecting a molecule or category of interest in the Project Tree view followed by a click on the "Zoom to fit" button (⇠⇢) at the bottom of the Project Tree view (figure 9.7). Double-clicking an entry in the Project Tree will have the same effect.

9.2.2 Project settings

A number of general settings can be adjusted from the **Side Panel**. Personal settings as well as molecule visualizations can be saved by clicking in the lower right corner of the **Side Panel** (☰). This is described in detail in section 4.5.

Project Tree Tools

Just below the Project Tree, the following tools are available

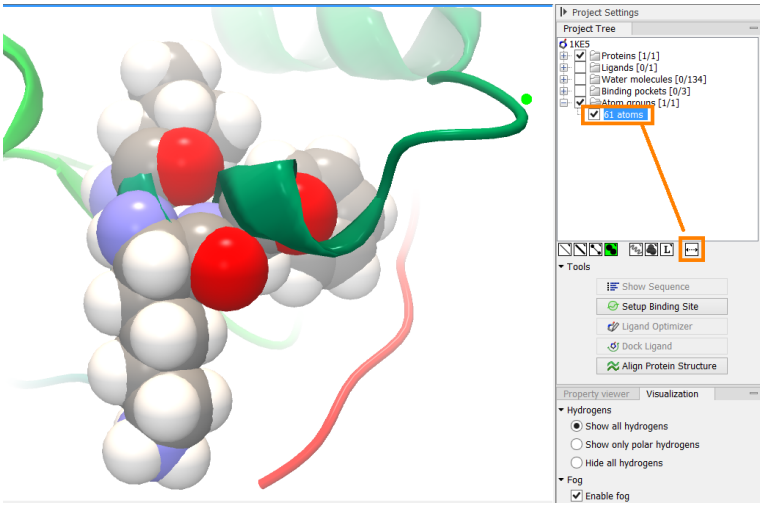


Figure 9.7: *The "Fit to screen" button can be used to bring a particular molecule or category of molecules in focus.*

- **Show Sequence** Select molecules which have sequences associated (Protein, DNA, RNA) in the Project Tree, and click this button. Then, a split-view will appear with a sequence list editor for each of the sequence data types (Protein, DNA, RNA). This is described in section 9.4.1.

- **Align to Existing Sequence** Select a protein chain in the Project Tree, and click this button. Then protein sequences and sequence alignments found in the Navigation Area, can be linked with the protein structure. This is described in section 9.4.2.

- **Transfer Annotations** Select a protein chain in the Project Tree, that has been linked with a sequence using either the "Show Sequence" or "Align to Existing Sequence" options. Then it is possible to transfer annotations between the structure and the linked sequence. This is described in section 9.4.3.

- **Setup Binding Site** Clicking this button will invoke the Binding Site Setup dialog box, which will do an automatic binding site setup, display the setup settings and give access to modify them (see section 9.12.1).

- **Ligand Optimizer** This opens an interactive dialogue to use for modifying a ligand or docking result. If a Binding Site Setup is present, the modifications will adjust to the binding site, and the interactions can be visualized. This is further described in section 9.11.

- **Dock Ligand** Select ligands that you wish to dock to the Binding Site Setup, and click this button to start the docking (see 9.12.3).

- **Align Protein Structure** This will invoke the dialog for aligning protein structures based on global alignment of whole chains or local alignment of e.g. binding sites defined by atom groups. This is described in section 9.5.

- **Generate Biomolecule** This will invoke the dialog for generating biomolecules based on information available from imported PDB files. This is described in section 9.6.

Property viewer

The Property viewer, found in the Side Panel, lists detailed information about the atoms that the mouse hovers over. For all atoms the following information is listed:

- **Molecule** The name of the molecule the atom is part of.
- **Residue** For proteins and nucleic acids, the name and number of the residue the atom belongs to is listed, and the chain name is displayed in parentheses.
- **Name** The particular atom name, if given in input, with the element type (Carbon, Nitrogen, Oxygen...) displayed in parentheses.
- **Hybridization** The atom hybridization assigned to the atom.
- **Charge** The atomic charge as given in the input file. If charges are not given in the input file, some charged chemical groups are automatically recognized and a charge assigned.

For atoms in molecules imported from a PDB file, extra information is given:

- **Temperature** Here is listed the b-factor assigned to the atom in the PDB file. The b-factor is a measure of uncertainty or disorder in the atom position; the higher the number, the higher the disorder.
- **Occupancy** For each atom in a PDB file, the occupancy is given. It is typically 1, but if atoms are modeled in the PDB file, with no foundation in the raw data, the occupancy is 0. If a residue or molecule has been resolved in multiple positions, the occupancy is between 0 and 1.

For atoms in protein models created by tools in the workbench, the following extra information is given:

- **Temperature** For structure models, the temperature value is an estimate of local structure uncertainty. The three aspects contributing to the assigned atom temperature is also listed, and described in section 13.1.1. The temperature value is a measure of uncertainty or disorder in the atom position; the higher the number, the higher the disorder.
- **Occupancy** For modeled structures and atoms, the occupancy is set to zero.

If an atom is selected, the Property view will be frozen with the details of the selected atom shown. If then a second atom is selected (by holding down Ctrl while clicking), the distance between the two selected atoms is shown. If a third atom is selected, the angle for the second atom selected is shown. If a fourth atom is selected, the dihedral angle measured as the angle between the planes formed by the three first and three last selected atoms is given.

If a molecule is selected in the Project Tree, the Property view shows information about this molecule. Two measures are always shown:

- **Atoms** Number of atoms in the molecule.
- **Weight** The weight of the molecule in Daltons.

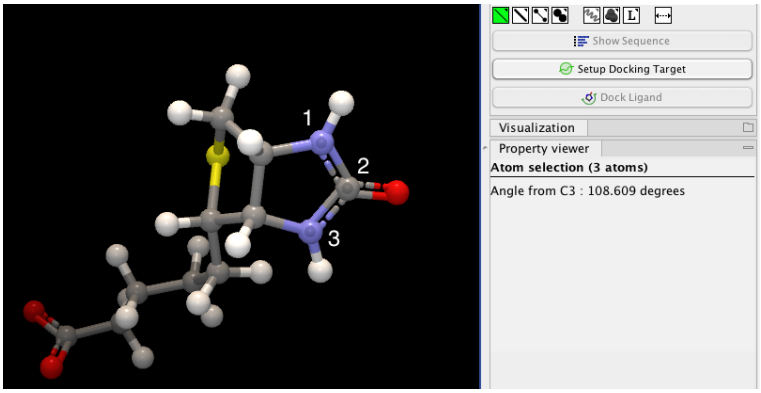


Figure 9.8: *Selecting two, three, or four atoms will display the distance, angle, or dihedral angle, respectively.*

Visualization settings

Under "Visualization" five options exist:

- **Hydrogens** Hydrogen atoms can be shown (Show all hydrogens), hidden (Hide all hydrogens) or partially shown (Show only polar hydrogens).
- **Fog** "Fog" is added to give a sense of depth in the view. The strength of the fog can be adjusted or it can be disabled.
- **Clipping plane** This option makes it possible to add an imaginary plane at a specified distance along the camera's line of sight. Only objects behind this plane will be drawn. It is possible to clip only surfaces, or to clip surfaces together with proteins and nucleic acids. Small molecules, like ligands and water molecules, are never clipped.
- **3D projection** The view is opened up towards the viewer, with a "Perspective" 3D projection. The field of view of the perspective can be adjusted, or the perspective can be disabled by selecting an orthographic 3D projection.
- **Coloring** The background color can be selected from a color palette by clicking on the colored box.

9.3 Snapshots of the molecule visualization

To save the current view as a picture, right-click in the **View Area** and select "File" and "Export Graphics". Another way to save an image is by pressing the "Graphics" button in the Workbench toolbar ( ). Next, select the location where you wish to save the image, select file format (PNG, JPEG, or TIFF), and provide a name, if you wish to use another name than the default name.

You can also save the current view directly on data with a custom name, so that it can later be applied (see section 4.5).

9.4 Tools for linking sequence and structure

The *CLC Drug Discovery Workbench* has functionality that allows you to link a protein sequence to a protein structure. Selections made on the sequence will show up on the structure. This allows

you to explore a protein sequence in a 3D structure context. Furthermore, sequence annotations can be transferred to annotations on the structure and annotations on the structure can be transferred to annotations on the sequence (see section 9.4.3).

9.4.1 Show sequence associated with molecule

From the Side Panel, sequences associated with the molecules in the Molecule Project can be opened as separate objects by selecting protein or nucleic acid entries in the Project Tree and clicking the button labeled "Show Sequence" (figure 9.9). This will generate a Sequence or Sequence List for each selected sequence type (protein, DNA, RNA). The sequences can be used to select atoms in the Molecular Project as described in section 9.2.1. The sequences can also be used as input for sequence analysis tools or be saved as independent objects. You can later re-link to the sequence using "Align to Existing Sequence" (see section 9.4.2).

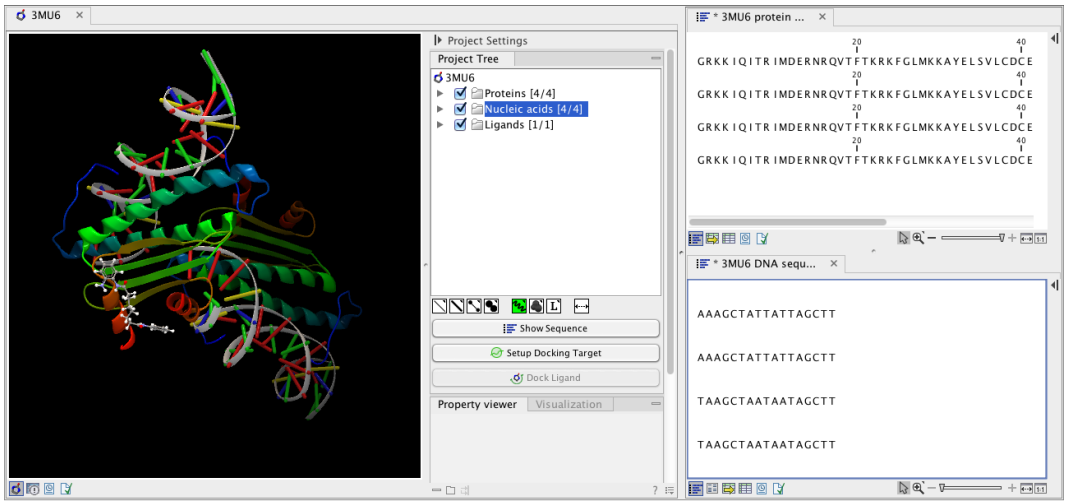


Figure 9.9: *Protein chain sequences and DNA sequences are shown in separate views.*

9.4.2 Link sequence or sequence alignment to structure

The "Align to Existing Sequence" button can be used to map and link existing sequences or sequence alignments to a protein structure chain in a Molecule Project (3D view). It can also be used to reconnect a protein structure chain to a sequence or sequence alignment previously created by Show Sequence (section 9.4.1) or Align to Existing Sequence.

Select a single protein chain in the project tree (see figure 9.10). Pressing "Align to Existing Sequence" then opens a Navigation Area browser, where it is possible to select one or more Sequence, Sequence Lists, or Alignments, to link with the selected protein chain.

If the sequences or alignments already contain a sequence identical to the protein chain selected in the Molecule Project (i.e. same name and amino acid sequence), this sequence is linked to the protein structure. If no identical sequence is present, a sequence is extracted from the protein structure (as for Show Sequence - section 9.4.1), and a sequence alignment is created between this sequence and the sequences or alignments selected from the Navigation Area. The new sequence alignment is created (see section 14.1) with the following settings:

- Gap open cost: 10.0

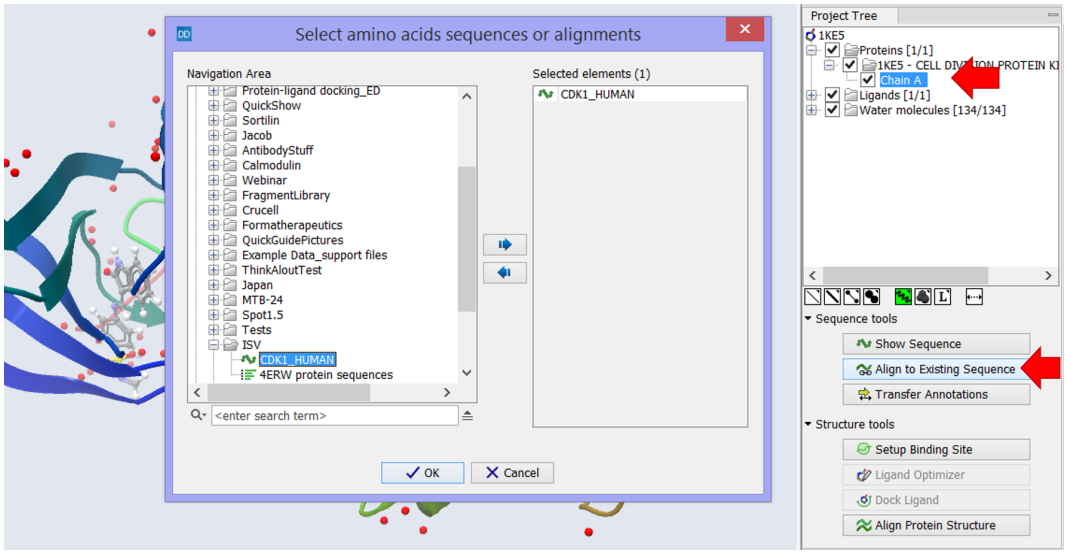


Figure 9.10: *Select a single protein chain in the Project Tree and invoke "Align to Existing Sequence".*

- Gap Extension cost: 1.0
- End gap cost: free
- Existing alignments are not redone

When the link is established, selections on the linked sequence in the sequence editor will create atom selections in the 3D view, and it is possible to transfer annotations between the linked sequence and the 3D protein chain (see section 9.4.3). Notice, that the link will be broken if either the sequence or the 3D protein chain is modified.

Two tips if the link is to a sequence in an alignment:

1. Read about how to change the layout of sequence alignments in section 14.2
2. It is only annotations present on the sequence linked to the 3D view that can be transferred to atom groups on the structure. To transfer sequence annotations from other sequences in the alignment, first copy the annotations to the sequence in the alignment that is linked to the structure (see figure 9.13 and section 14.3.4).

9.4.3 Transfer annotations between sequence and structure

The Transfer Annotations dialog makes it possible to create new atom groups (annotations on structure) based on protein sequence annotations and vice versa.

You can read more about sequence annotations in section 10.3 and more about atom groups in section 9.2.1.

Before it is possible to transfer annotations, a link between a protein sequence editor and a Molecule Project (a 3D view) must be established. This is done either by opening a sequence associated with a protein chain in the 3D view using the 'Show Sequence' button (see section 9.4.1) or by mapping to an existing sequence or sequence alignment using the 'Align to Existing Sequence' button (see section 9.4.2).

Invoke the Transfer Annotations dialog by selecting a linked protein chain in the Project Tree and press 'Transfer Annotations' (see figure 9.11).

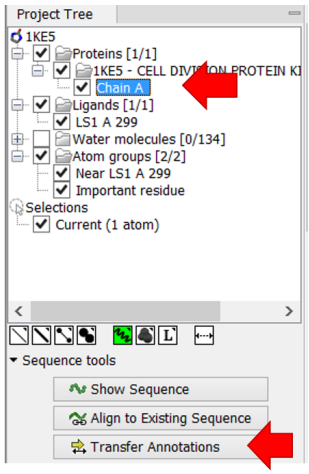


Figure 9.11: *Select a single protein chain in the Project Tree and invoke "Transfer Annotations".*

The dialog contains two tables (see figure 9.12). The left table shows all atom groups in the Molecule Project, with at least one atom on the selected protein chain. The right table shows all annotations present on the linked sequence. While the Transfer Annotations dialog is open, it is not possible to make changes to neither the sequence nor the Molecule Project, however, changes to the visualization styles are allowed.

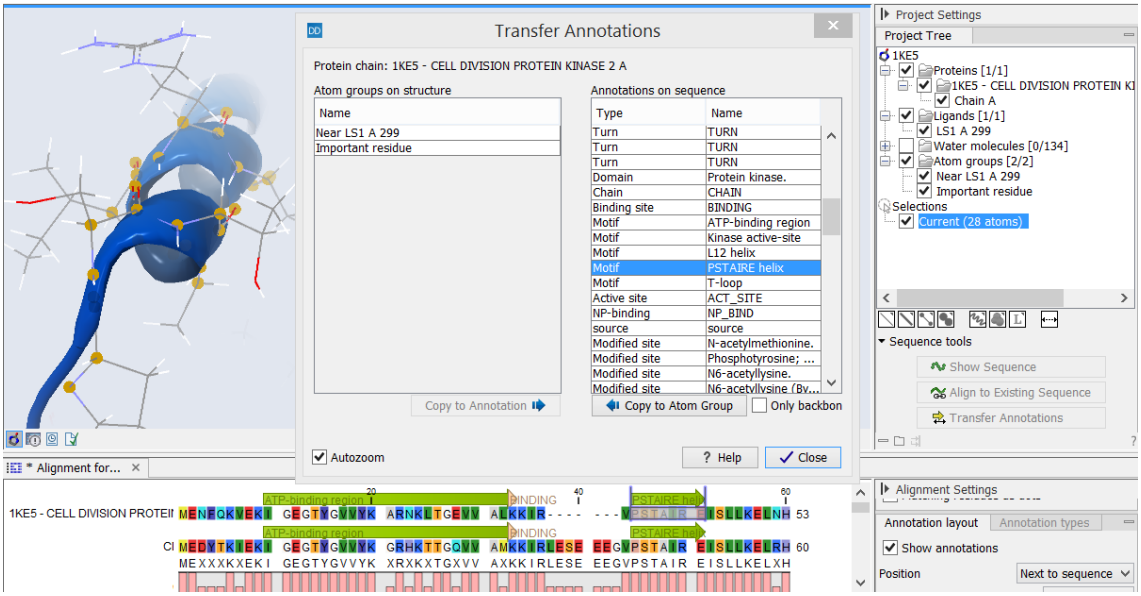


Figure 9.12: *The Transfer Annotations dialog allow you to select annotations listed in the two tables, and copy them from structure to sequence or vice versa.*

How to undo annotation transfers

In order to undo operations made using the Transfer Annotations dialog, the dialog must first be closed. To undo atom groups added to the structure, activate the 3D view by clicking in it and press Undo in the Toolbar. To undo annotations added to the sequence, activate the sequence view by clicking in it and press Undo in the Toolbar.

Transfer sequence annotations from aligned sequences

It is only annotations present on the sequence linked to the 3D view that can be transferred to atom groups on the structure. If you wish to transfer annotations that are found on other sequences in a linked sequence alignment, you need first to copy the sequence annotations to the actual sequence linked to the 3D view (the sequence with the same name as the protein structure). This is done by invoking the context menu on the sequence annotation you wish to copy (see figure 9.13 and section 14.3.4).

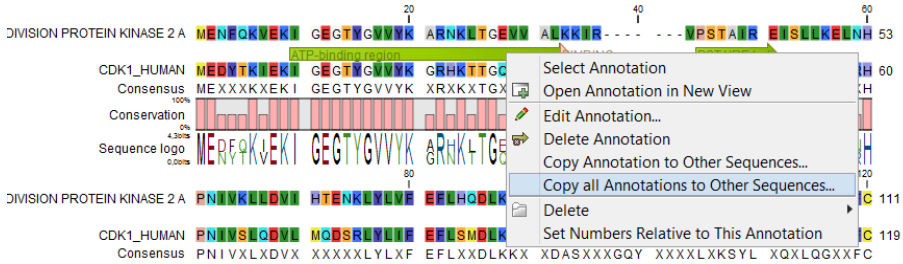


Figure 9.13: *Copy annotations from sequences in the alignment to the sequence linked to the 3D view.*

9.5 Protein structure alignment

The Align Protein Structure tool allows you to compare a protein or binding pocket in a **Molecule Project** with proteins from other **Molecule Projects**. The tool is invoked using the ( ) Align Protein Structure action from the **Molecule Project Side Panel**. This action will open an interactive dialog box (figure 9.14). By default, when the dialog box is closed with an "OK", a new **Molecule Project** will be opened containing all the input protein structures laid on top of one another. All molecules coming from the same input Molecule Project will have the same color in the initial visualization.

Note! Care should be taken when using the Setup Binding Site tool on a Molecule Project containing a structure alignment. This is because, by default, all the protein chains lying on top of each other will be included in the setup. Make sure to un-select those protein chains in the Binding Site Setup dialog box that should not be included in the Binding Site Setup.

9.5.1 The Align Protein Structure dialog box

The dialog box contains three fields:

- **Select reference (protein chain or atom group)** This drop-down menu shows all the protein chains and residue-containing atom groups in the current **Molecule Project**. If an atom group is selected, the structural alignment will be optimized in that area. The 'All chains from *Molecule Project* option will create a global alignment to all protein chains in the project, fitting e.g. a dimer to a dimer.
- **Molecule Projects with molecules to be aligned** One or more **Molecule Projects** containing protein chains may be selected.
- **Output options** The default output is a single **Molecule Project** containing all the input projects rotated onto the coordinate system of the reference. Several alignment statistics,

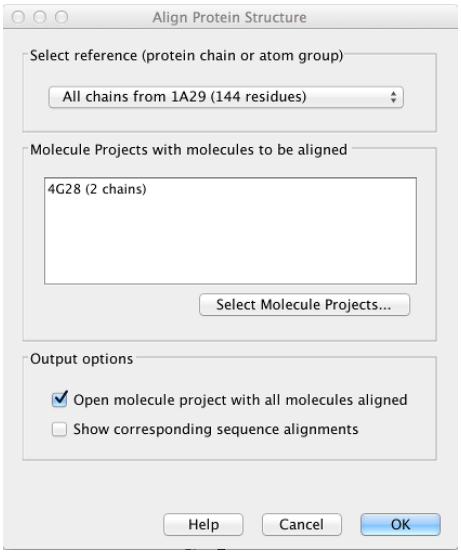


Figure 9.14: *The Align Protein Structure dialog box.*

including the RMSD, TM-score, and sequence identity, are added to the **History** of the output **Molecule Project**. Additionally, a sequence alignments of the aligned structures may be output, with the sequences linked to the 3D structure view.

9.5.2 Example: alignment of calmodulin

Calmodulin is a calcium binding protein. It is composed of two similar domains, each of which binds two calcium atoms. The protein is especially flexible, which can make structure alignment challenging. Here we will compare the calcium binding loops of two calmodulin crystal structures – PDB codes 1A29 and 4G28.

Initial global alignment The 1A29 project is opened and the Align Protein Structure dialog is filled out as in figure 9.14. Selecting "All chains from 1A29" tells the aligner to make the best possible global alignment, favoring no particular region. The output of the alignment is shown in figure 9.15. The blue chain is from 1A29, the brown chain is the corresponding calmodulin chain from 4G28 (a calmodulin-binding chain from the 4G28 file has been hidden from the view). Because calmodulin is so flexible, it is not possible to align both of its domains (enclosed in black boxes) at the same time. A good global alignment would require the brown protein to be translated in one direction to match the N-terminal domain, and in the other direction to match the C-terminal domain (see black arrows).

Focusing the alignment on the N-terminal domain To align only the N-terminal domain, we return to the 1A29 project and select the **Show Sequence** action from beneath the **Project Tree**. We highlight the first 62 residues, then convert them into an atom group by right-clicking on the "Current" selection in the **Project Tree** and choosing "Create Group from Selection" (figure 9.16). Using the new atom group as the reference in the alignment dialog leads to the alignment shown in figure 9.17. In addition to the original input proteins, the output now includes two Atom Groups, which contain the atoms on which the alignment was focused. The **History** of the output **Molecule Project** shows that the alignment has 0.9 Å RMSD over the 62 residues.

Aligning a binding site Two bound calcium atoms, one from each calmodulin structure, are

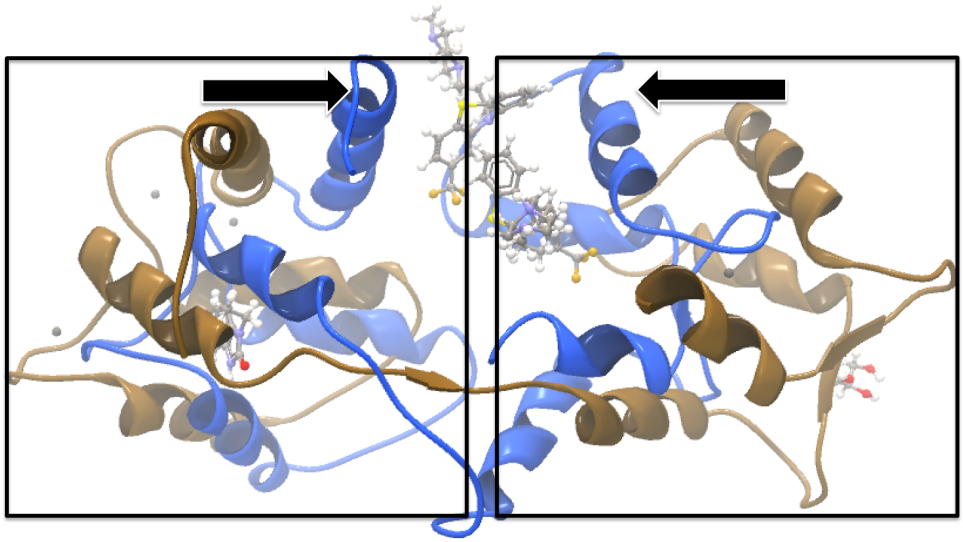


Figure 9.15: *Global alignment of two calmodulin structures (blue and brown). The two domains of calmodulin (shown within black boxes) can undergo large changes in relative orientation. In this case, the different orientation of the domains in the blue and brown structures makes a good global alignment impossible: the movement required to align the brown structure onto the blue is shown by arrows – as the arrows point in opposite directions, improving the alignment of one domain comes at the cost of worsening the alignment of the other.*

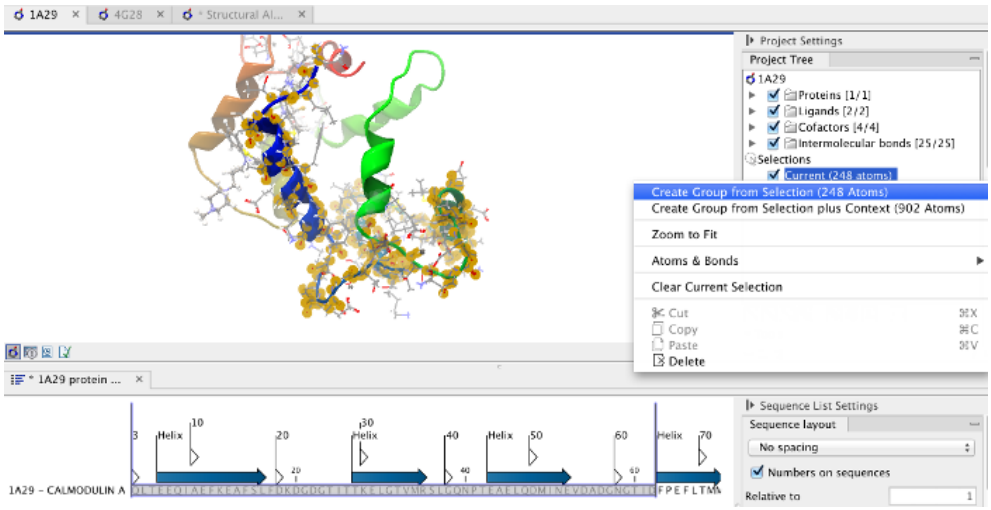


Figure 9.16: *Creation of an atom group containing the N-terminal domain of calmodulin.*

shown in the black box of figure 9.17. We now wish to make an alignment that is as good as possible about these atoms so as to compare the binding modes. We return to the 1A29 project, right-click the calcium atom from the cofactors list in the **Project Tree** and select "Create Nearby Atoms Group". Using the new atom group as the reference in the alignment dialog leads to the alignment shown in figure 9.18.

9.5.3 The Align Protein Structure algorithm

Any approach to structure alignment must make a trade-off between alignment length and alignment accuracy. For example, is it better to align 200 amino acids at an RMSD of 3.0 Å

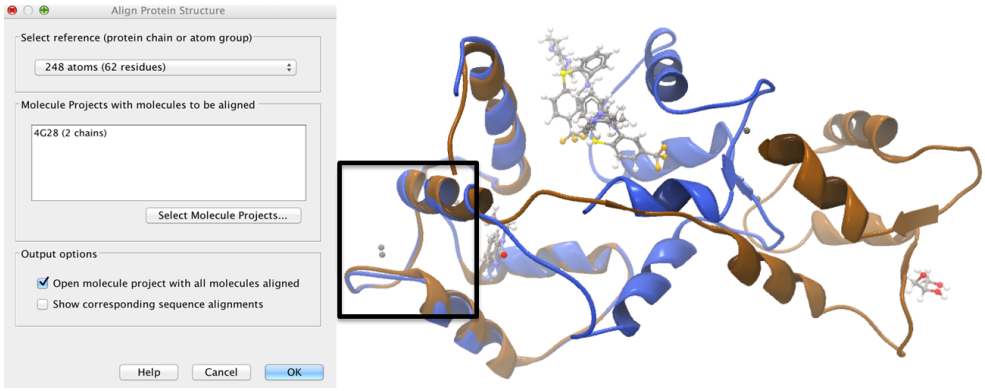


Figure 9.17: *Alignment of the same two calmodulin proteins as in figure 9.15, but this time with a focus on the N-terminal domain. The blue and brown structures are now well-superimposed in the N-terminal region. The black box encloses two calcium atoms that are bound to the structures.*

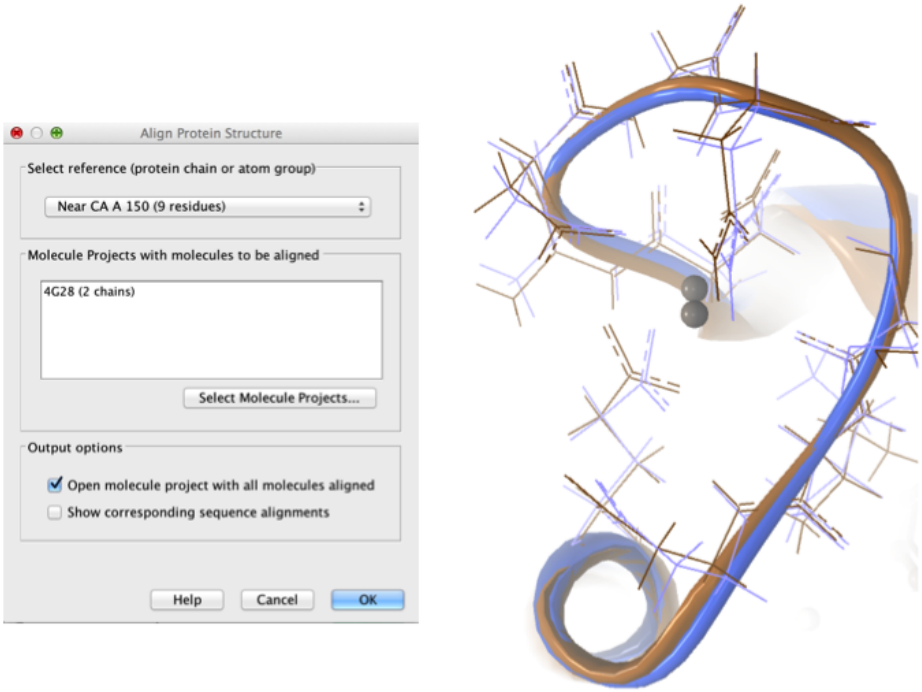


Figure 9.18: *Alignment of the same two calmodulin domains as in figure 9.15, but this time with a focus on the calcium atom within the black box of figure 9.17. The calcium atoms are less than 1 Å apart – compatible with thermal motion encoded in the atoms' temperature factors.*

or 150 amino acids at an RMSD of 2.5 Å? The Align Protein Structure algorithm determines the answer to this question by taking the alignment with the higher TM-score. For an alignment focused on a protein of length L, this is:

$$\text{TM-score} = \frac{1}{L}\sum_i \frac{1}{1+\frac{d_i}{d(L)}^2}$$

where i runs over the aligned pairs of residues, d_i is the distance between the i^{th} such pair, and $d(L)$ is a normalization term that approximates the average distance between two randomly

chosen points in a globular protein of length L [Zhang and Skolnick, 2004]. A perfect alignment has a TM-score of 1.0, and two proteins with a TM-score $>$0.5 are often said to show structural homology [Xu and Zhang, 2010].

The Align Protein Structure Algorithm attempts to find the *structure alignment* with the highest TM-score. This problem reduces to finding a *sequence alignment* that pairs residues in a way that results in a high TM-score. Several sequence alignments are tried including an alignment with the BLOSUM62 matrix, an alignment of secondary structure elements, and iterative refinements of these alignments.

The Align Protein Structure Algorithm is also capable of aligning entire protein complexes. To do this, it must determine the correct pairing of each chain in one complex with a chain in the other. This set of chain pairings is determined by the following procedure:

1. Make structure alignments between every chain in one complex and every chain in the other. Discard pairs of chains that have a TM-score of < 0.4
2. Find all pairs of structure alignments that are consistent with each other i.e. are achieved by approximately the same rotation
3. Use a heuristic to combine consistent pairs of structure alignments into a single alignment

The heuristic used in the last step is similar to that of MM-align [Mukherjee and Zhang, 2009], whereas the first two steps lead to both a considerable speed up and increased accuracy. The alignment of two 30S ribosome subunits, each with 20 protein chains, can be achieved in less than a minute (PDB codes 2QBD and 1FJG).

9.6 Generate Biomolecule

Protein structures imported from a PBD file show the tertiary structure of proteins, but not necessarily the biologically relevant form (the quaternary structure). Oftentimes, several copies of a protein chain need to arrange in a multi-subunit complex to form a functioning biomolecule. In some PDB files several copies of a biomolecule are present and in others only one chain from a multi-subunit complex is present. In many cases, PDB files have information about how the molecule structures in the file can form biomolecules.

When a PDB file with biomolecule information available has been either downloaded directly to the workbench using the *Search for PDB Structures at NCBI* or imported using *Import Molecules with 3D Coordinates*, the information can be used to generate biomolecule structures in *CLC Drug Discovery Workbench*.

The "Generate Biomolecule" dialog is invoked from the Side Panel of a Molecule Project (figure 9.19). The button () is found in the Structure tools section below the Project Tree.

There can be more than one biomolecule description available from the imported PDB files. The biomolecule definitions have either been assigned by the crystallographer solving the protein structure (Author assigned = "Yes") or suggested by a software prediction tool (Author assigned = "No"). The third column lists which protein chains are involved in the biomolecule, and how many copies will be made.

Select the preferred biomolecule definition and click OK.

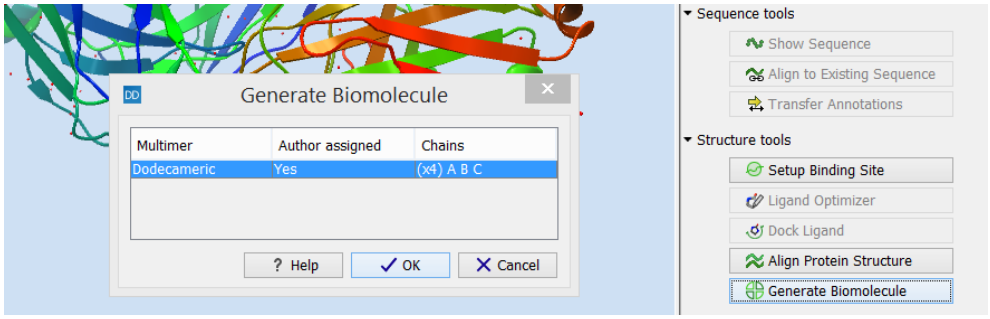


Figure 9.19: *The Generate Biomolecule dialog lists all possibilities for biomolecules, as given in the PDB files imported to the Molecule Project. In this case, only one biomolecule option is available. The Generate Biomolecule button that invokes the dialog can be seen in the bottom right corner of the picture.*

A new Molecule Project will open containing the molecules involved in the selected biomolecule (example in figure 9.20). If required by the biomolecule definition, copies are made of protein chains and other molecules, and the copies are positioned according to the biomolecule information given in the PDB file. The copies will in that case have "s1", "s2", "s3" etc. at the end of the molecule names seen in the Project Tree.

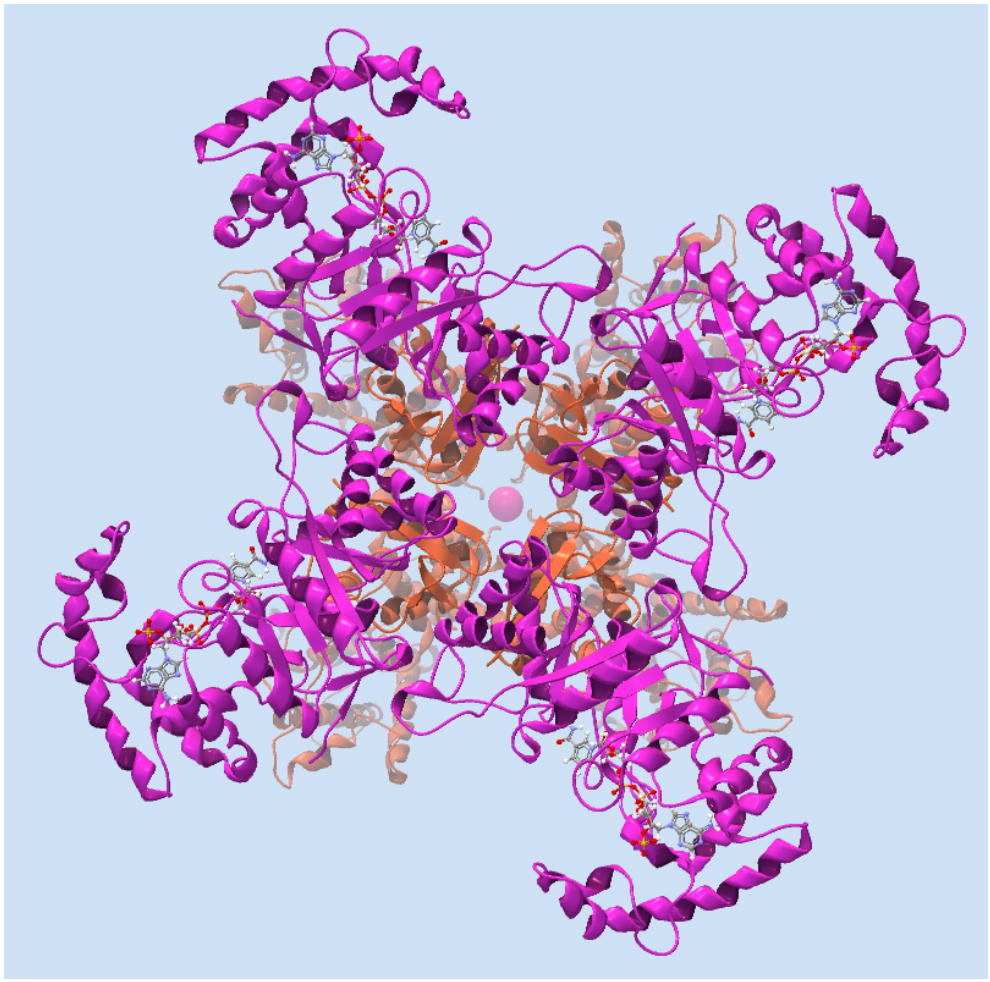


Figure 9.20: *One of the biomolecules that can be generated after downloading the PDB 2R9R to* **CLC Drug Discovery Workbench**. *It is a voltage gated potassium channel.*

If the proteins in the Molecule Project already are present in their biomolecule form, the message "The biological unit is already shown" is displayed, when the "Generate Biomolecule" button is clicked.

If the PDB files imported or downloaded to a Molecule Project did not hold biomolecule information, the message "No biological unit is associated with this Molecule Project" is shown, when the Generate Biomolecule button is clicked.

9.7 Molecule Tables

When molecules are imported to a **Molecule Table**, each molecule is shown as a row in the table. This row contains information about the molecule such as its name, 2D depiction, number of atoms, and molecular weight. Information about the 3D structure of the molecule is also present, and can be visualized by connecting the Molecule Table to a 3D view (section 9.7.3).

9.7.1 Create Molecule Table

There are several ways to go about creating a library of molecules in a Molecule Table.

- **Molecules with 3D coordinates** in the format SDF or Mol2 can be imported directly to a Molecule Table. SDF and Mol2 are both plain text formats, and you can have several molecules one after the other in a Mol2 or SDF file. Each molecule will get its own row in a Molecule Table. One Mol2/SDF file will turn into one Molecule Table (section 6.2.2).
- **Molecules with 2D coordinates** in the format SDF or Mol2 (the z-coordinate set to zero for all atoms) can be imported directly to a Molecule Table. SDF and Mol2 are both plain text formats, and you can have several molecules one after the other in a Mol2 or SDF file. Each molecule will get its own row in a Molecule Table (section 6.2.6).
- **Molecules given in SMILES notation** can be imported directly to a Molecule Table. Several SMILES can be listed in the same .smi file, one SMILES string per line. If the SMILES string is followed by a text string, this will be used as the name of the molecule on import. Each molecule will get its own row in a Molecule Table (section 6.2.6).

  Example of how a .smi file with eight molecules from the ZINC database could look:

  Cc1ccc(o1)C(F)(F)F ZINC15442277

  Cc1[nH]c(cn1)C(F)(F)F ZINC31176470

  Cc1c[nH]nc1C(F)(F)F ZINC00066375

  c1cnn(c1)CC(F)(F)F ZINC12396097

  C1CN[C@H](C1(F)F)C(=O)[O-] ZINC36458630

  COC(=O)C(C(=O)OC)F ZINC00158096

  c1cc2c(c(c1)F)c(=O)c2=O ZINC00367517

  [C@H]([C@@H](C(=O)N)F)(C(=O)[O-])[NH3+] ZINC01568479

- If you have **molecules in a Molecule Project** that you would like to have in a Molecule Table, use the Extract Ligands tool from the Toolbox (section 9.17).
- If you have a **Molecule Table that you would like to extend** with extra molecules:

1. The molecules should be found in a Molecule Table or Molecule Project.
2. Select the molecules in the Molecule Table or Molecule Project they currently are found in.
3. Use the right-click context menu (or Ctrl-C) to copy the molecules.
4. Go to the Molecule Table you would like to extend, and select any row.
5. Use the right-click context menu (or Ctrl-V) to paste the molecules to the table.

- If you are **building a molecule library from molecules copy-pasted from a 2D sketcher**:

1. Paste the molecules to a Molecule Project as usual (section 6.2.7).
2. Inspect the molecules in 3D and make corrections to them if necessary.
3. When you have pasted all the molecules you wish to have in a Molecule Table, use the Extract Ligands tool to transfer them to a table (section 9.17).
4. If you wish to add them to another table, you can select all rows in the extracted Molecule Table (Ctrl-A), and copy-paste them into another Molecule Table, as described above.

9.7.2 Grid view of molecule 2D depictions

The 2D depictions of the molecules in the table can be seen in a grid layout by selecting the Molecule Grid view from below the table view () , or from right-click context menu option **Show | Molecule Grid**. From the Side Panel, the number of columns in the grid view can be adjusted, as well as which parameter is used to sort the molecules in the view. Two labels can be specified for the molecules; an upper and a lower label. The molecules in the view can be printed to a PDF document by clicking the Print option in the toolbar. If the table contains many molecules, the option *Show only selected* found in the Side Panel, will only show and print the molecules selected from the table or the grid view.

9.7.3 Viewing Molecule Table structures in 3D

Each row in a **Molecule Table** contains information about all the 3D atom positions in the molecule, just as when a molecule is imported to a **Molecule Project**. To see the 3D information, the **Molecule Table** must be connected to a **Molecule Project** view. This is done using the "Select View" action found in the **Side Panel**. This will open the "View in 3D" dialog box shown in figure 9.21.

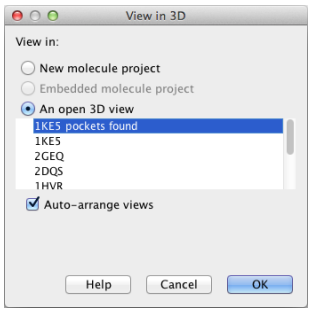


Figure 9.21: *The "View in 3D" dialog box.*

The following options are available:

- **New molecule project** will open a new, blank **Molecule Project**, and show the molecules selected in the table as guests in this project.
- **Embedded molecule project** is only an option for Docking Results tables (see section 9.7.3), and only if the option to embed the input molecule project in the output table was selected in the Dock Ligands or Screen Ligands wizards (see section 9.12.2 and 9.13). The embedded **Molecule Project** will in that case contain the Binding Site Setup used for the docking.
- **An open 3D view** will connect the table to one of the **Molecule Projects** already open in the view area. If no **Molecule Projects** are currently open, the list will be empty.

If the "Auto-arrange views" option is checked, a split-screen will be arranged with the **Molecule Table** at the top and the 3D view at the bottom. The entries selected in the **Molecule Table** will now appear as "guests" in the selected **Molecule Project**. The visualization of the guest molecules can be changed as for the molecules belonging to the **Molecule Project**, and they can be hidden by un-checking the boxes next to them (see figure 9.22).

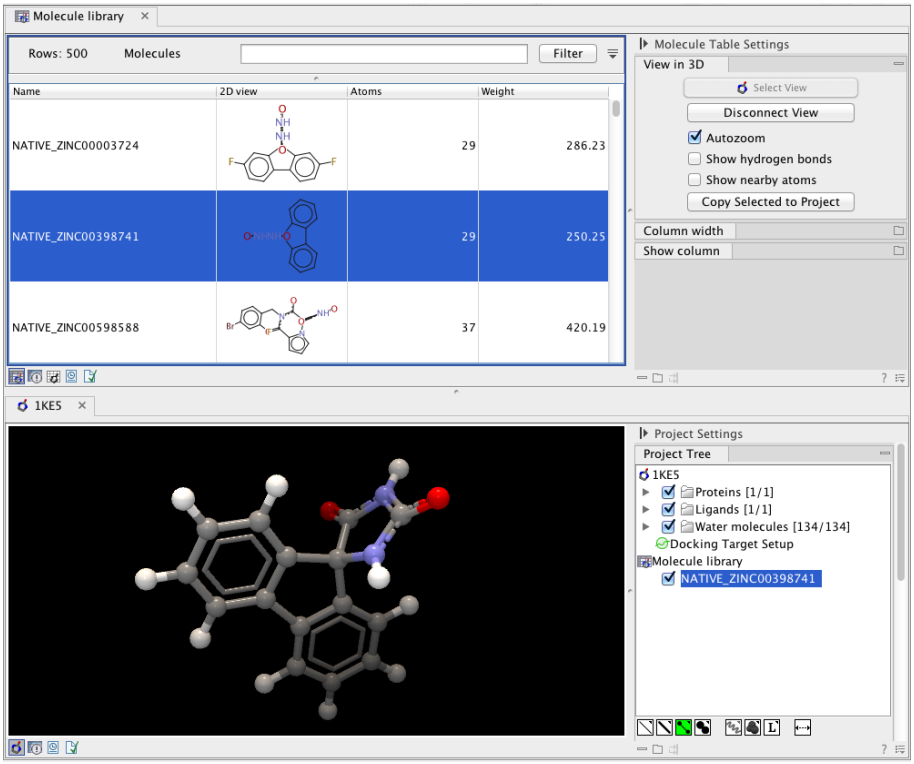


Figure 9.22: *A split-screen arrangement with the Molecule Table on top and the 3D view in the bottom.*

In the **Molecule Table** Side Panel, the "View in 3D" category has several options that relates to the behavior of the molecules of the 3D view:

- **Auto zoom** will zoom in on the molecules that have been selected in the table. This is very handy if you wish to browse through a list of molecules in a table, by using the arrow-keys.
- **Show hydrogen bonds** will find and display potential hydrogen bond interactions between the molecules selected in the table, and the molecules included in the Binding Site Setup

in the connected **Molecule Project**. If no Binding Site Setup is present, hydrogen bonds to all molecules in all categories except Ligands and Docking Results will be shown.

- **Show nearby atoms** will display all residues and molecules within 5 Å of the molecule that was selected in the table. If a Binding Site Setup is present, only molecules included in the setup will be shown, otherwise all molecules except for Ligands and Docking Results are shown.
- **Copy Selected to Project** will copy the molecules selected in the table, to their respective categories in the connected **Molecule Project**. They can then be saved together with the **Molecule Project**.

When showing hydrogen bond interactions or the nearby atoms for molecules in a table, the visualization of the interacting atoms will always be in Wireframe, and only the ligand representation will change when another molecule representation is selected for the "guest" molecule. If you wish to customize the visualization of the interacting atoms as well, you should copy the molecule from the **Molecule Table** to the **Molecule Project** using the "Copy Selected to Project" action in the **Molecule Table** Side Panel (section 9.9), and then follow the directions for custom atom group selection of interacting atoms in the section 'How to select a particular group of atoms' in section 9.2.1.

9.8 Docking Results Tables

The output from the tools Dock Ligands (section 9.12.2) and Screen Ligands (section 9.13) is a Docking Results Table. This table has the same options in the Side Panel as a **Molecule Table** and extra columns with output data for the docked ligands. What makes a Docking Results Table special is the ability to have the **Molecule Project**, with the binding site setup used for docking, embedded in the table. The option to embed the **Molecule Project** is found in the Dock Ligands and Screen Ligands wizards (figure 9.23).

Figure 9.23: *The option to embed the Molecule Project is found in the Dock Ligands and Screen Ligands wizards.*

This will ensure that the setup used for the docking simulation is saved together with the output, and that the ligand binding modes resulting from the dockings can always be seen with the binding site in the way it was set up (see section 9.7.3). Be aware that for tables with only a few molecules, the embedded **Molecule Project** will increase the disc size of the tables considerably.

9.9 Editing molecule objects

Individual molecules can be moved and copied between **Molecule Tables** and **Molecule Projects**.

In **Molecule Tables**, select the molecule entries you wish to copy, move or delete. From the right-click context menu, you can then Cut, Copy or Delete the entries from the *Edit* menu. If Cut or Copy is selected, it is now possible to paste the molecules into another open **Molecule Table** or **Molecule Project**. To paste the molecules into a **Molecule Project**, invoke the context menu from the free space in the 3D view or from the **Project Tree** view, or use Ctrl + V (Cmd + V on Mac).

In **Molecule Projects**, the molecules you wish to copy, move or delete should be selected in the **Project Tree** view. The right-click context menu will allow you to Cut, Copy or Delete the entries. If Cut or Copy is selected, it is now possible to paste the molecules into another **Molecule Project** or a **Molecule Table**, using the right-click context menu or Ctrl + V (Cmd + V on Mac).

If a **Molecule Table** is connected to a **Molecule Project** (see section 9.7.3), the molecules selected in the table can be copied to the **Molecule Project** simply be clicking the "Copy Selected to Project" action in the table Side Panel.

9.9.1 Editing atom and bond properties

In some cases, the molecule input contains errors or the automatic assignments are imperfect (see section 6.2). As the assigned hybridization, hydrogen assignment and bond order is translated into atom types used in the docking simulation, a correct representation is of particular importance for molecules used for docking. If the molecule representation is incorrect, the simulated protein-ligand interactions will also be incorrect. Therefore, molecules can be edited manually, to make the necessary corrections to the representation.

For small molecules (ligands), there are two ways to modify the representation. The Ligand Optimizer (section 9.11) is a convenient tool for this purpose, but notice that it will sometimes make changes to atom coordinates while doing modifications. Another option, which is available for all categories of molecules, is to edit atom and bond properties directly in the 3D view. The original heavy atom coordinates will always be maintained during these manipulations. To make changes to an atom, right-click the atom, and a list of properties to be used for editing will appear (figure 9.24).

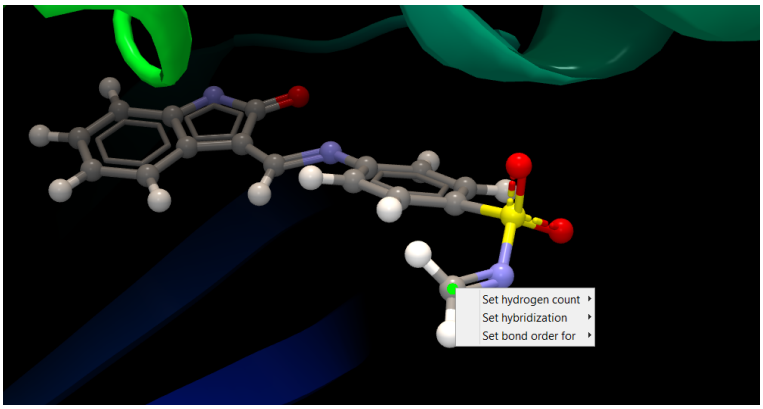


Figure 9.24: *To make changes to an atom, right-click the atom, and a list of properties to be used for editing will appear.*

- **Set hydrogen count** will give you a list of the possible number of attached hydrogens to choose from.
- **Set hybridization** will give you the option of SP1, SP2, and SP3 hybridization. Changing the hybridization will change the position of attached hydrogens, to best adhere to the geometric arrangement around an atom with the requested hybridization. It is not possible to change the atom hybridization for atoms in aromatic rings.
- **Set bond order for** will give you a list that shows the bonds to neighboring heavy atoms. For each of these, you can pick the order of the bond from a list with the options single,

double, triple, and delocalized. It is not possible to change bond order between atoms in aromatic rings.

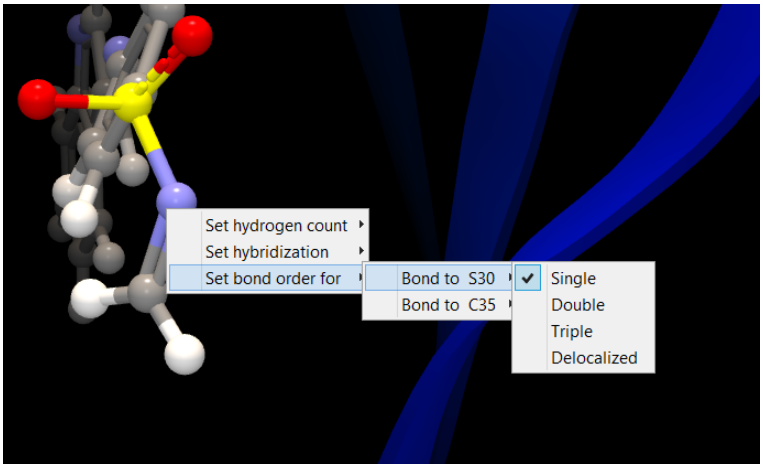


Figure 9.25: *To make changes to an atom, right-click the atom, and a list of properties to change will appear.*

The Issues view (see section 6.2.9) lists all the concerns that were raised during import of the molecule files. The list is updated whenever changes are made to molecules. The issues listed in the Chemistry category, point out aspects of the molecule representation that seems out of the ordinary. It is therefore a valuable help to have the Issues list in split screen view together with the 3D view, while manually editing molecules. Not having any Chemistry issues connected to a molecule is not a guarantee that the representation is correct, but if there are Chemistry issues, it is relevant to check if they call for changes being made to the molecule.

Please note! If the input file has charges assigned to the atoms, these will also be imported. If the input file has no charges assigned, particular chemical groups (e.g. acids) will be assigned charges during import. However, no charges will or can be assigned at a later stage, and the charges are ignored in the docking simulation. Chemistry issues related to a missing charge can therefore not be amended manually. In such situations it is safe to ignore "Chemistry" issues if the bonding pattern for the atom seems otherwise correct.

Please also note! If you are making changes to a molecule from a **Molecule Table**, the changes are related to the table, and not the 3D view that was used to specify the changes. It is therefore the Issues list on the **Molecule Table** that lists the chemistry issues of the molecule, and not the Issues list of the **Molecule Project** used for visualizing the molecule. Likewise, if you would like to undo a change, the focus should be shifted to the **Molecule Table**, before clicking the undo button or pressing Ctrl + Z.

9.9.2 Converting molecules to Cofactors or Ligands

Molecules are automatically assigned to a category (Proteins, Nucleic acids, Ligands, Cofactors, Docking results, Water molecules) on import. As it is only possible to dock molecules from the Ligands category, molecules assigned to the Proteins, Nucleic acids, Cofactors, and Docking results categories can be manually re-assigned to the Ligands category. This is done from the right-click context menu when the molecule entries are selected in the Project Tree.

Likewise, it is only possible to include molecules in the Proteins, Nucleic acids, Water molecules, and Cofactors categories in the Binding Site Setup. Molecules automatically assigned to the

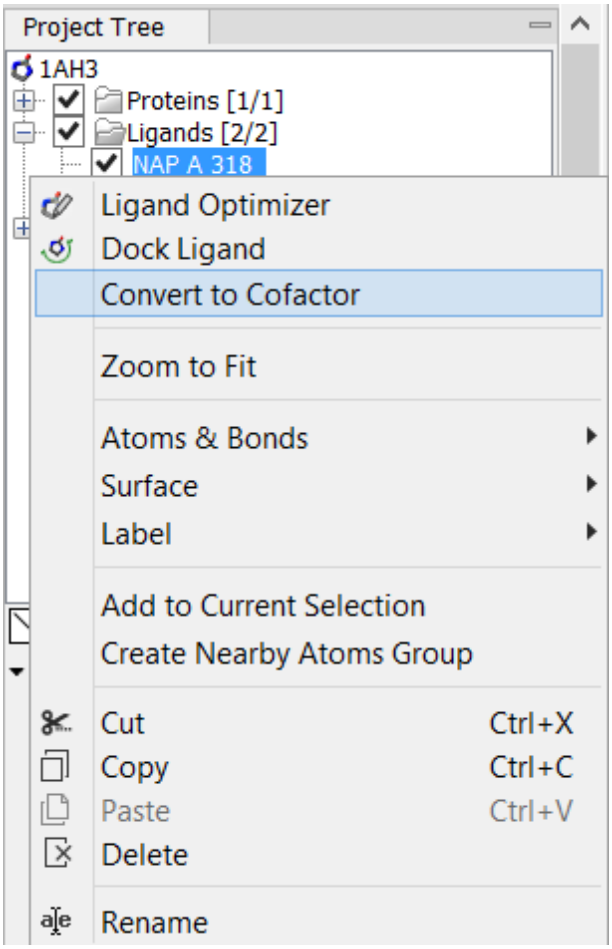


Figure 9.26: *From the context menu, molecules can be converted to Ligands or Cofactors in the Project Tree.*

Ligands category can therefore be moved to the Cofactors category instead. This is done from the context menu when a ligand has been selected in the Project Tree.

9.10 The Protein Optimizer

The **Protein Optimizer** ( ) makes it possible to modify protein side chains while taking surrounding molecules into account.

It can be used to replace missing atoms in a protein structure, to refine the results of modeling from the Find and Model Structure tool (section 13.1), or to investigate how mutations in a binding pocket may affect binding.

To invoke the **Protein Optimizer** dialog, click the **Protein Optimizer** button in the Structure tools panel below the Project Tree. Then select a residue to be changed.

Note: It is possible to undo and redo operations done while using the dialog. If the dialog is canceled, all changes done are rolled back.

Also note: If multiple mutations are desired, it is best to make them with the Find and Model Structure tool, which will usually give more accurate results (see Section 9.10.6 for details). The output of Find and Model Structure may then be refined in the **Protein Optimizer**.

The **Protein Optimizer** dialog (figure 9.27) is composed of the following five panels:

1. The selection panel shows the protein and residue currently being optimized.
2. The issues panel lists any chemical issues for the current residue.
3. The modify residue panel performs modifications on the current residue.
4. The modify surrounding residues panel performs modifications on a local region of the protein.
5. The visualization panel controls the view.

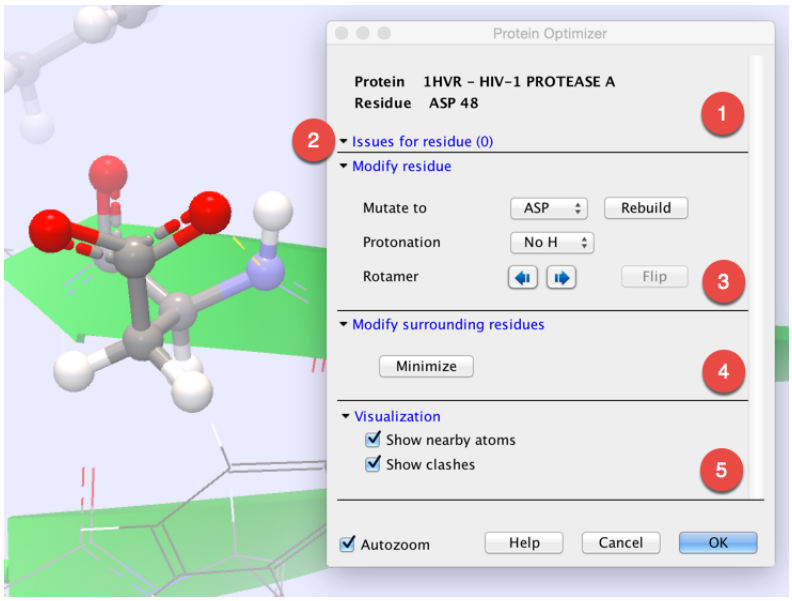


Figure 9.27: *The Protein Optimizer dialog*

9.10.1 The selection panel

The selection panel shows the protein and residue currently being optimized (figure 9.28). Residues may be selected in a variety of ways, for example by clicking within the 3D view, via a selection on a linked sequence, or from a Molecule Project's Issue Editor.

Note: mutating a residue in the current protein will break connections to its linked sequence.

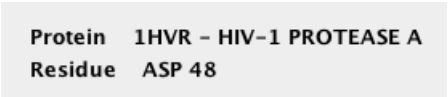


Figure 9.28: *The selection panel.*

9.10.2 The issues panel

This panel shows any chemical issues specific to the current residue. These issues are taken from the issues list (see Section 6.2.9).

9.10.3 The modify residue panel

The modify residue panel (figure 9.29) offers the following operations:

- Mutate a residue. The current amino acid is selected in the drop down box. Select a new amino acid to mutate the residue. The mutated residue always has the best minimized rotamer.
- Rebuild a residue that is missing side chain atoms. To do this, click the rebuild button. The residue is rebuilt completely, without taking existing positions into account.
- Change the protonation state. The current state is shown in the drop drop down box. Select a new state to change the protonation.
- Choose a rotamer for a residue. Rotamers are minimized when selected.
- Flip a residue. Asparagine (ASN), Glutamine (GLN) and Histidine (HIS) residues contain pairs of atoms that are often hard to distinguish in protein crystal structures. The flip button swaps these pairs of atoms, but keeps their coordinates fixed. A flip may be required if i) rebuilding the residue favors the flipped conformation, ii) flipping improves the local hydrogen bond network, or iii) hydrogen atoms clash without the flip.

Figure 9.29: *The modify residue panel*

The atoms in the modeled residue will be assigned temperature values as described in section 13.1.1. That means that atoms with steric clashes will be colored bright red when visualized using the "Color by Temperature" color scheme (section 9.2.1).

What is seen during a modification?

Modifications to a residue take into account all visible molecules in the residue's neighborhood. For example, a mutation in a binding pocket may choose different rotamers as the bound ligand is toggled on and off (figure 9.30).

9.10.4 The modify surrounding residues panel

A mutation to a residue may cause knock-on effects to the positions of atoms in the surrounding region of the protein. The minimize button preserves the selected residue, while minimizing the positions of surrounding side chain atoms on the same protein chain. As with the modify panel, all visible molecules are taken into account during minimization, such that toggling a ligand on and off may change the results.

Note: Minimization is a stochastic process, meaning that pressing the minimize button twice may give slightly different results.

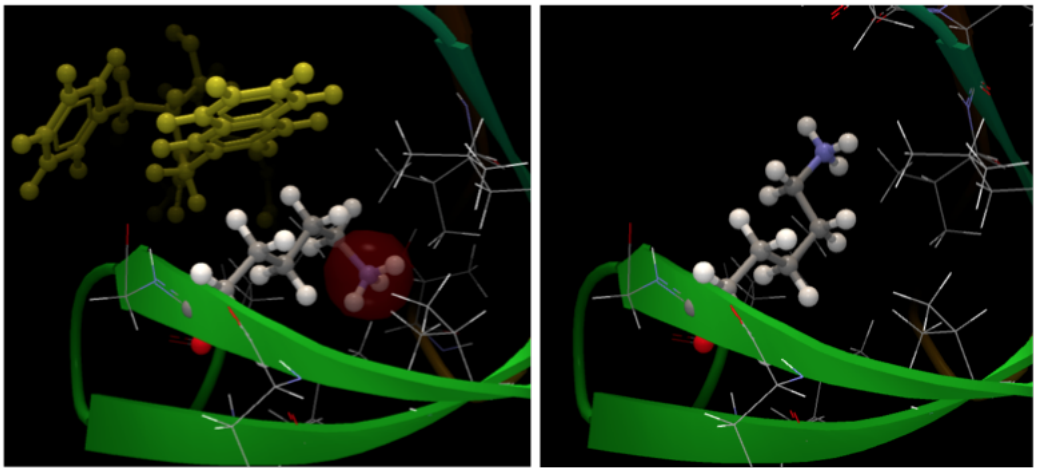


Figure 9.30: *Left: A ligand (yellow) is visible. The side chain is placed taking the position of the ligand into account. This leads to a clash (red) between atoms in the protein chain. Right: The ligand is toggled off, and the side chain is rebuilt. It can now occupy the space previously occupied by the ligand, removing the earlier clash.*

9.10.5 The visualization panel

The Protein Optimizer shows the selected residue as **Ball and Stick**. Additional visualization styles can be checked to visualize interactions between the selected residue and its surroundings. **Show nearby atoms** shows atoms in steric contact with the residue's side chain. Hydrogen bonds from the residue are marked with dashed blue lines, and potential hydrogen bonds – hydrogen bonds which would be formed if the protonation was changed – are marked with dashed yellow lines.

Show clashes shows the parts of a residue that clash with the surrounding structure. Clashes are only shown for atoms that have been modeled, either with the Find and Model Structure tool, or using the modify residue panel. Clashes are marked as transparent red regions around the affected atoms.

9.10.6 How side chains are modeled

Amino acid side chains tend to assume one of a discrete number of "rotamer" conformations. The rotamers used in *CLC Drug Discovery Workbench* have been calculated from a non-redundant set of high-resolution crystal structures.

Side chains are modeled by their lowest energy rotamer. In the **Protein Optimizer** this is found by trying each possible rotamer in turn. For creating homology model structures with the **Find and Model Structure** tool, a heat bath Monte Carlo simulated annealing algorithm is used, similar to the OPUS-Rota method [Lu et al., 2008]. The algorithm consists of approximately 100 cycles of simulation. In a single cycle, rotamers are selected for each side chain with a probability according to their energy. As the simulation proceeds, the selection increasingly favors the rotamers with the lowest energy, and the algorithm converges.

A local minimization of the modeled side chains is then carried out, to reduce unfavorable interactions with the surroundings.

Calculating the energy of a side chain rotamer

The total energy is composed of several terms:

- Statistical potential: This score accounts for interactions between the given side chain and the local backbone, and is estimated from a database of high-resolution crystal structures. It depends only on the rotamer and the local backbone dihedral angles ϕ and ψ.
- Atom interaction potential: This score is used to evaluate the interaction between a given side chain atom and its surroundings. It is the same score that is used to evaluate protein-ligand interactions for docking and ligand optimization in *CLC Drug Discovery Workbench*.
- Disulfide potential: Only applies to cysteines. It follows the form used in the RASP program [Miao et al., 2011] and serves to allow disulfide bridges between cysteine residues. It penalizes deviations from ideal disulfide geometry. A distance filter is applied to determine if the disulfide potential should be used, and when it is applied the atom interaction potential between the two sulfur atoms is turned off. Note that disulfide bridges are not formed between separate chains.

Note: For manual mutations, the atom interaction potential considers only interactions within the mutated protein chain. In the case where side chains are modeled with the Find and Model Structure tool, the interactions with all molecules in the template PDB file (except water) are considered.

Local minimization of side chain

After applying a side chain rotamer from the library to the backbone, a local minimization may be carried out for rotations around single bonds in the side chain.

The potential to minimize with respect to bond rotation is composed of the following terms:

- Atom interaction potential: Same as for calculating the energy of a rotamer.
- Disulfide potential: Same as for calculating the energy of a rotamer.
- Harmonic potential: This penalizes small deviations from ideal rotamers according to a harmonic potential. This is motivated by the concept of a rotamer representing a minimum energy state for a residue without external interactions.

9.11 The Ligand Optimizer

The **Ligand Optimizer** ( ) makes it possible to modify small molecules while taking the surrounding protein environment into account.

It can be used when a specific binding mode is already known (for instance from a docking simulation, or from a co-crystallized ligand) to investigate how changes in the ligand chemistry will affect the binding. It is also possible to use the **Ligand Optimizer** without any protein context, for instance in order to prepare small molecules before a docking simulation.

To invoke the **Ligand Optimizer** dialog, select a single ligand or docking result from the **Project Tree** and click the **Ligand Optimizer** button in the Structure tools panel below the Project Tree. It is possible to convert molecules from other categories to ligand (see section 9.9.2).

Note: Whenever a change is performed on the ligand, all atomic charges are cleared and reassigned using built-in preparation. Likewise, some chemical groups (such as carboxylate) are automatically turned into a delocalized bond representation when recognized.

Also note: It is possible to undo and redo operations done while using the dialog. If the dialog is canceled, all changes done are rolled back.

The **Ligand Optimizer** dialog is composed of the following five panels:

9.11.1 The 2D depiction panel

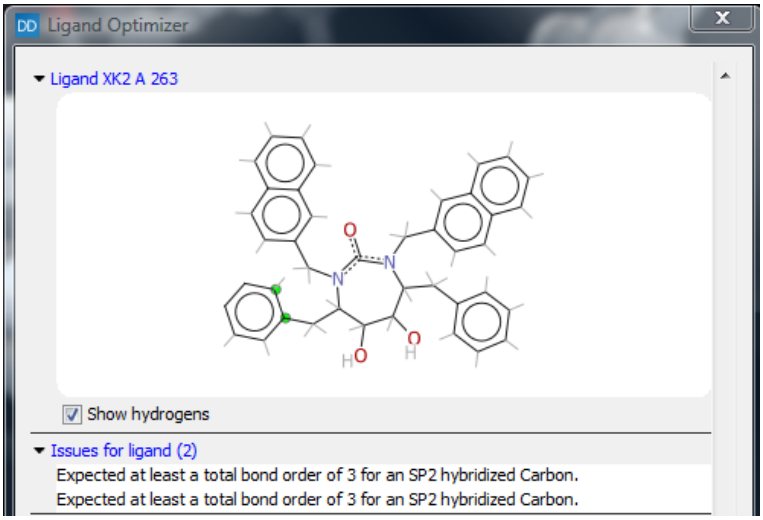


Figure 9.31: *The 2D depiction panel, shown together with the issue list panel. The two green dots on the 2D depictions indicates the selection of a covalent bond.*

The 2D depiction panel shows the molecule currently being optimized (figure 9.31). It is possible to select atoms from either the 2D or the 3D view. A bond can be selected by selecting the two atoms it connects.

A checkbox makes it possible to toggle hydrogens on and off, which can be useful when specifying the attachment point when adding new fragments. It is possible to zoom in and out on the 2D depiction using the mouse wheel, making it easy to navigate larger ligands.

9.11.2 The issues panel

This panel shows any chemical issues for the current ligand (figure 9.31). These issues are also shown in the issues list (see section 6.2.9), but the issues shown here are limited to those concerning the ligand being optimized. Clicking on an issue highlights the relevant position on the molecule in 2D and 3D.

9.11.3 The modify panel

The modify panel offers the following operations:

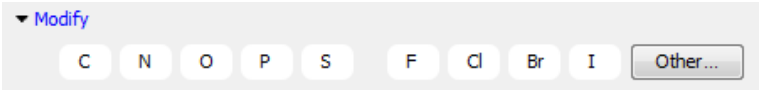


Figure 9.32: *The element buttons make it possible to change the element type for an atom, or to add an element to the ligand (this is done by selecting a hydrogen atom and choosing the desired element).*

1) **Change the element for an existing atom** (figure 9.32). To do this select a non-hydrogen atom on the ligand, and click the desired element button from the top row. It is also possible to select more exotic elements using the **Other...** button. Replacing an element will automatically adjust bond lengths and the number of attached hydrogens.

2) **Add new elements to the ligand** (figure 9.32). This is done by selecting a hydrogen atom, which will serve as the attachment point. After selecting the hydrogen atom, click on an element button to replace the hydrogen atom. As above, bond lengths and hydrogens are automatically adjusted.

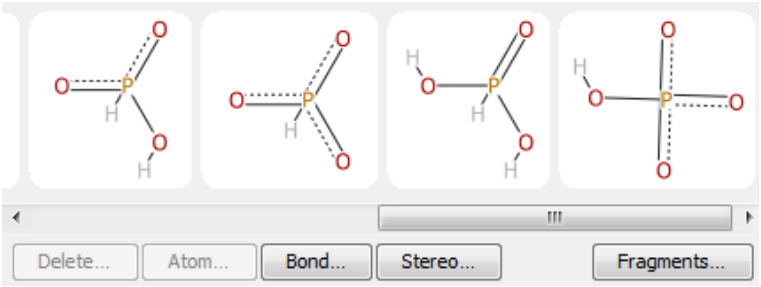


Figure 9.33: *Fragments are attached by selecting a hydrogen atom on the ligand, and a hydrogen atom on the fragment. Any Molecule Project may be used as a source for fragments.*

3) **Attach a new fragment to the ligand** (figure 9.33). This works similar to adding a new element, but requires selecting an attachment point on the fragment as well: first select a hydrogen on the ligand, then select a hydrogen on the fragment. The fragment will be attached to the ligand by replacing the selected hydrogens with a covalent bond. As above, bond lengths and hydrogen counts are automatically adjusted. It is also possible to just select a heavy atom with at least one hydrogen attached on the ligand. In this case an arbitrary hydrogen on the heavy atom will be used as the attachment point for the fragment.

It is possible to choose between different fragment libraries using the **Fragments...** button. Two small libraries are provided with the **Ligand Optimizer** (one for ring systems, and one for common chemical groups, such as carboxylate and phosphate groups), but it is possible to use molecules from any **Molecule Project** as fragments. To do this, choose the **Fragments...** | **From Molecule Project...** menu entry. Any ligands in the selected **Molecule Project** will appear as attachable

fragments in the **Ligand Optimizer**. This also makes it possible to attach fragments drawn in a 2D sketcher: simply paste the SMILES string from the 2D sketcher into a **Molecule Project** (as described in section 6.2.7), and use the **Molecule Project** as a fragment library.

Fragment modifications are previewed in real-time in the 3D view. After a fragment has been added, the newly attached part of the ligand will be subjected to a quick geometry optimization in order to place the fragments in the most energetically favorable way with regard to the protein context.

4) **Delete parts of the ligand**

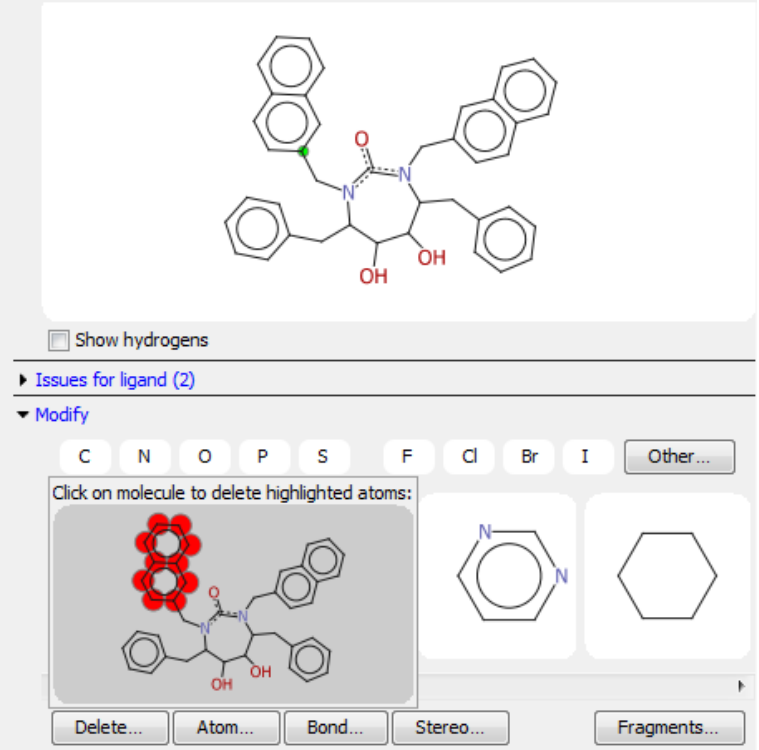


Figure 9.34: *Choose an atom, and press the Delete... button in order to remove a part containing the selected atom.*

The **Delete...** button (figure 9.33) makes it possible to delete any part of the molecule containing the selected atom. However, a part can only be deleted if the remaining part of the molecule is still connected, and no rings are broken (whole rings may be deleted though). The atoms connected to removed parts are automatically repaired by adding hydrogens.

5) **Change atom type**

The **Atom...** button (figure 9.35) makes it possible to change the atom type (the number of hydrogens and formal charges). If the atom type needs to be changed into a different hybridization, it might be necessary to change the bonds connected to it by using the **Bond...** tool described below.

6) **Change bond order**

The **Bond...** button (figure 9.36) makes it possible to change bond orders. Start by choosing two covalently bonded atoms, and press the **Bond...** button. The different possible bond orders are then shown (e.g. single, double, triple). If the **Adjust atoms** checkbox is checked, neighboring atoms will automatically have their hybridization and hydrogen count corrected. In cases where

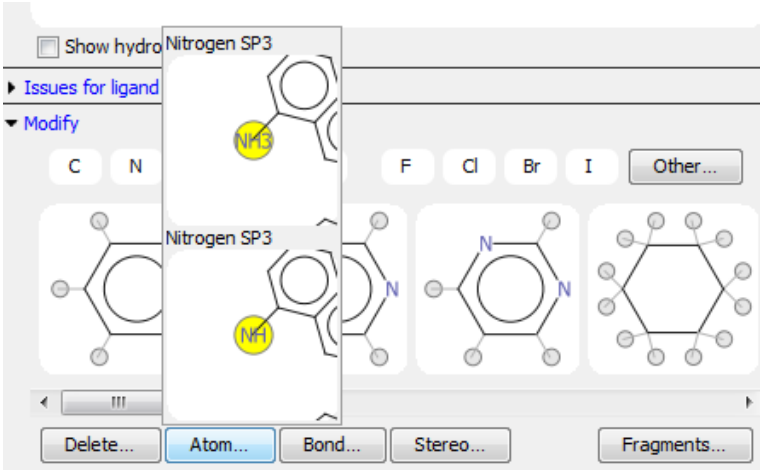


Figure 9.35: *Choose an atom and press the Atom... button to change the number of hydrogens or hybridization.*

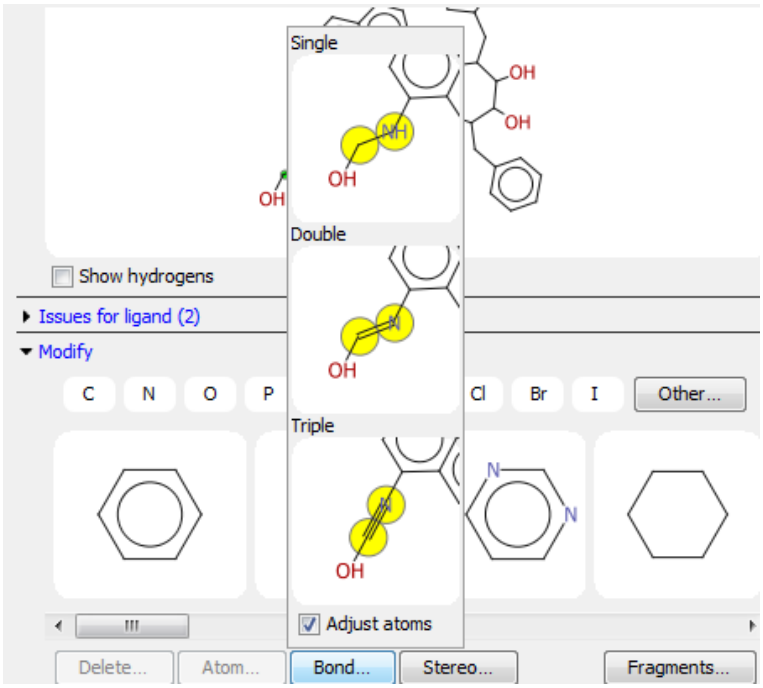


Figure 9.36: *Choose two covalently connected atoms and press the Bond... button to change the bond order.*

the automatic adjustments are unwanted, the **Adjust atoms** option can be disabled, which makes it possible to change bond order without updating atom hybridizations and geometry even though it might lead to chemistry issues.

7) **Change stereochemistry**

The **Stereo...** button (figure 9.37) makes it possible to change the stereo chemistry of a ligand. In order to do this, a stereo center must be selected on the ligand. If the selected atom is not a stereo center (or there is no atom currently selected), any valid stereo centers will be automatically shown using cyan overlays on the 2D depiction.

Two types of stereo chemistry changes are possible: E/Z swaps (for SP2 hybridized atoms), where the groups on one side of a double (or delocalized bond) are rotated 180 degrees, and

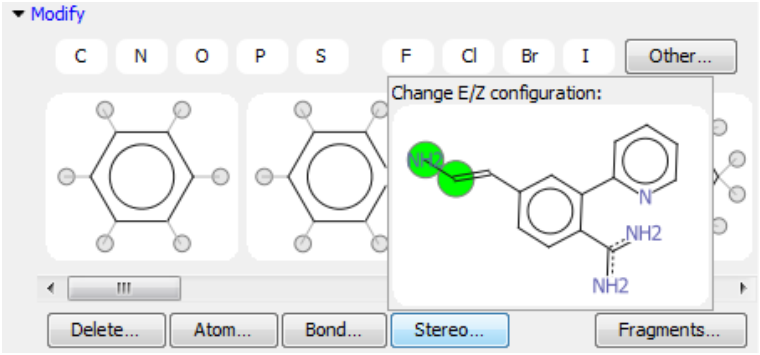


Figure 9.37: *press the Stereo... button to change stereo chemistry.*

R/S inversions (for SP3 hybridized atoms) where two different groups are swapped. For R/S inversions, we swap the two smallest attached groups. The stereochemistry options are not enabled on atoms with two identically attached groups.

8) **Rotate around bonds**

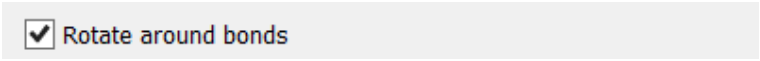


Figure 9.38: *Tick off the* Rotate around bonds *check box, to enable manual rotation around bonds in the ligand.*

It is possible to rotate part of a ligand relative to the rest, around a covalent bond.

In order to do this, first check the *Rotate around bonds* option. Then select a single atom. Circular double-headed arrows will then appear on rotatable bonds next to the selected atom in the 3D view (figure 9.39).

Click the arrow belonging to the part of the molecule you wish to move, and hold the mouse button while dragging. The part of the ligand containing the selected atom will stay in place. Release the mouse button when the rotated part of the molecule is in the wanted position.

While dragging, the arrow will turn green and snap to the most natural rotamers. For instance, two covalently bonded SP3 atoms will snap to the staggered positions with a 120 degrees rotational symmetry.

Notice that it is not possible to rotate around bonds in ring structures or bonds that only connect to terminal atoms.

9) **Modifying rings**

It is possible to modify ring geometry in the Ligand Optimizer. This is somewhat slower than other operations, since the small molecule geometry is subsequently minimized using a MMFF94-like force field. After that, the molecule is minimized with regards to the binding pocket geometry, similar to other Ligand Optimizer modifications.

Ring modifications may be performed by selecting two atoms and pressing the **Bond...** button. The following operations are possible:

Breaking a ring It is possible to break a ring by selecting two covalently bonded heavy atoms in a ring.

Creating a ring A ring may be formed by selecting two heavy atoms and pressing the **Bond...** button. (It is also possible to select hydrogens: a selected hydrogen will be treated as a selection

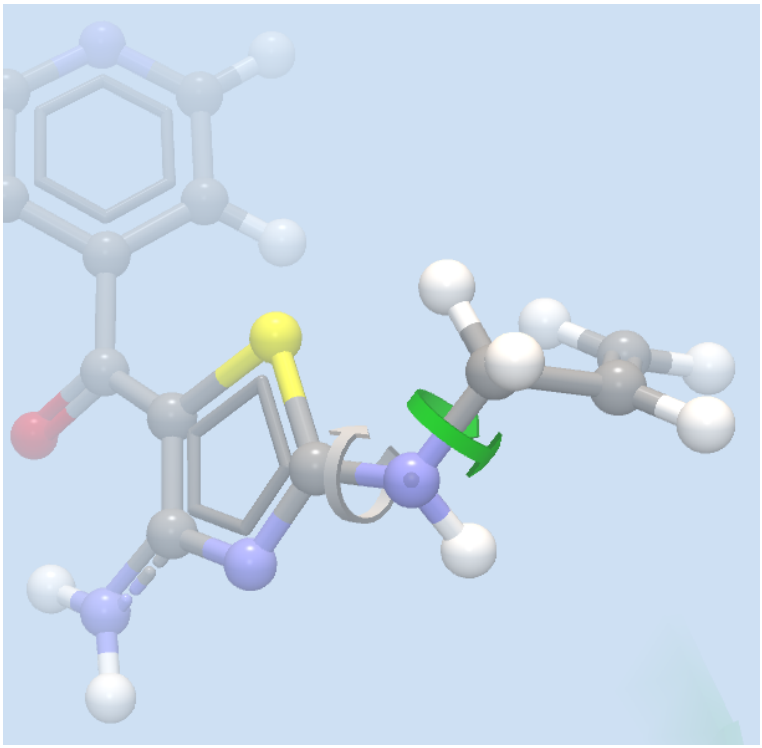


Figure 9.39: *Click-hold-drag from an arrow in the 3D view, to rotate that part of the molecule with respect to the rest.*

on the heavy atom it belongs to.). Atoms inside an existing ring may also be fused. The ligand optimizer will present options for creating either an aromatic or aliphatic ring. Notice that aromatic ring creation is not always possible.

Changing bond order It is possible to modify the bond order by selecting two covalently bond heavy atoms, which are part of the same aliphatic ring structure.

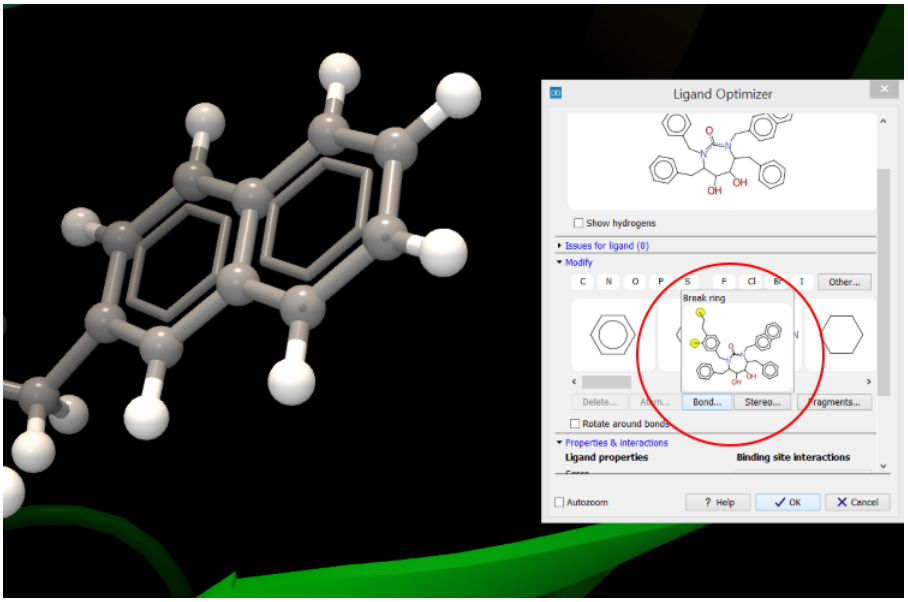


Figure 9.40: *Breaking a ring in the Ligand Optimizer*

9.11.4 The properties & interactions panel

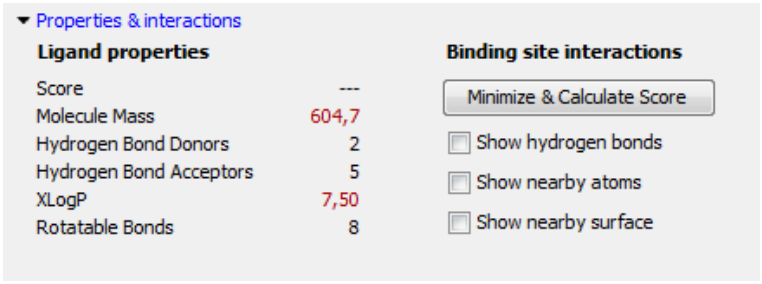


Figure 9.41: *The properties on the left side of the properties & interactions panel are updated dynamically. The binding site interactions features are available only if a binding site has been added to the project.*

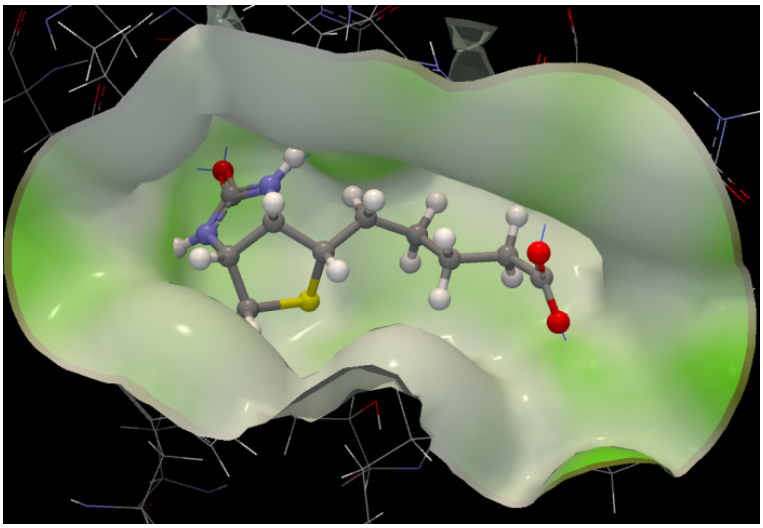


Figure 9.42: *A surface showing nearby parts of the protein. The surface is colored according to hydrophobicity - green spots are hydrophilic, polar regions of the protein.*

The left part of the properties & interactions panel displays a number of standard descriptors, which are updated automatically whenever the ligand is modified (figure 9.41). These descriptors include the Lipinski's rule of five descriptors. Lipinski rule of five violations will be indicated with red text.

Note: The right part of the properties & interactions panel is only enabled when a binding site is present in the workspace.

The **Minimize & Calculate Score** button makes it possible to perform an energy minimization with regards to the protein environment defined by the **Setup Binding Site** dialog. This minimization only changes the overall position, rotation, and any dihedral angles in the ligand. It will not change bond lengths and angles, or change ring conformations. The score mimics the potential energy change when the protein and ligand come together. This means that a very negative score corresponds to a strong binding and a less negative or even positive score corresponds to a weak or non-existing binding. You can read more about the scoring function in section 9.12.4.

The checkboxes on the right part of the properties & interactions panel make it possible to visualize interactions between the ligand and the binding site as defined in the Binding Site Setup. **Show hydrogen bonds** toggles on the display of hydrogen bonds between the ligand and the binding site molecules. **Show nearby atoms** shows atoms in steric contact with the ligand.

Show nearby surface shows the surface boundary for the parts of the binding site in contact with the ligand (figure 9.42). The surface is two-sided: it is opaque when viewed from the inside, but transparent from the outside in order to make it possible to manipulate ligands buried in deep binding pockets. The surface is colored according to a hydrophobicity measure: green areas correspond to hydrophilic regions (regions with polar atoms capable of forming hydrogen bonds), while the white areas are purely hydrophobic.

9.11.5 The constraints panel

Selected atoms in the ligand can be constrained to their current position during optimization of the ligand. All operations, such as changing bond orders or stereochemistry, keep the constrained atoms of the ligand in place. When minimizing the score of the protein-ligand interaction, deviations from the constrained atom positions will be heavily penalized.

Constraints may be used to prevent undesired conformational changes during larger modifications of a ligand and to keep known protein-ligand interactions in place. Constraints may also be used to keep ligand atoms at otherwise unfavorable positions, for instance atoms known to covalently bind to the target receptor.

To place a constraint, select an atom, and press the "Add" button in the constraints panel. A green transparent sphere in the 3D view will indicate that the atom is constrained, and the atom name will be added to the list of constraints in the constraints panel (see figure 9.43). To delete a constraint, select it from the list in the constraints panel and press the "Delete" button.

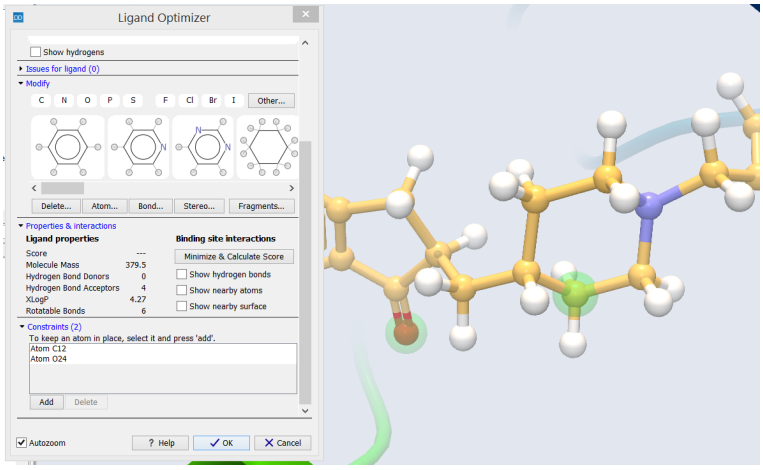


Figure 9.43: *Atoms with position constraints are marked by a green transparent sphere in the 3D view, and the atom names are listed in the constraints panel.*

If multiple atom constraints are placed on a molecule, it may not be possible to satisfy all of them while modifying the ligand. In this case, some constraints may end up in an unsatisfied state. An unsatisfied constraint is shown as a red sphere at the desired atom position. A red line connects the red sphere with the constrained atom (figure 9.44). When doing a "Minimize & Calculate Score", the heavy penalty on deviations from constraints will often force the molecule back into a conformation where the constraints are satisfied. If this is not the intention, the unsatisfied constraint should be deleted from the list of constraints in the dialog before doing a minimization.

Note that when minimizing the score, the penalty term from constraint violations are not included as part of the final 'Score' value. However, the 'Score' may become positive, it the constraints

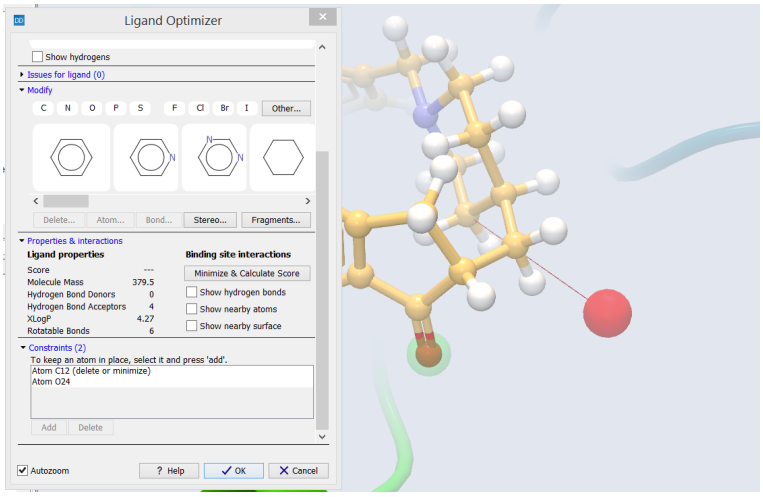


Figure 9.44: *When an atom is forced away from its constrained position, a red sphere indicates the desired position, and a red line connects it visually with the atom.*

force the ligand into energetically unfavorable conformations.

9.12 Molecular docking

In a docking simulation, one or more small molecules (ligands) are fitted into an expected binding site on the surface of a protein target. The protein-ligand interaction is scored, and the best scoring binding mode is returned for each ligand, together with the score. The steps to go through to explore protein-ligand interaction using docking, are described in the following:

- Setup the binding site in a **Molecule Project** (section 9.12.1).
- Dock ligands imported to a **Molecule Table** (section 9.12.2) or imported to the **Molecule Project** holding the binding site setup (section 9.12.3).
- Inspect the docking results (section 9.12.5).

9.12.1 Setup Binding Site

Before a small molecule can be fitted to a protein binding pocket in a docking simulation, you must first specify where on the protein the binding site is found, to limit the search for the optimal protein-ligand complex. This is done using the Setup Binding Site action from the **Molecule Project** Side Panel. This action will open an interactive dialog box, and when the dialog box is closed with an "OK", a Binding Site Setup will appear in the **Project Tree**, and the **Molecule Project** is now ready to use for docking. The binding site setup can later be inspected and altered, by re-invoking the action from the Side Panel or by double-clicking the Binding Site Setup entry in the **Project Tree** (figure 9.45).

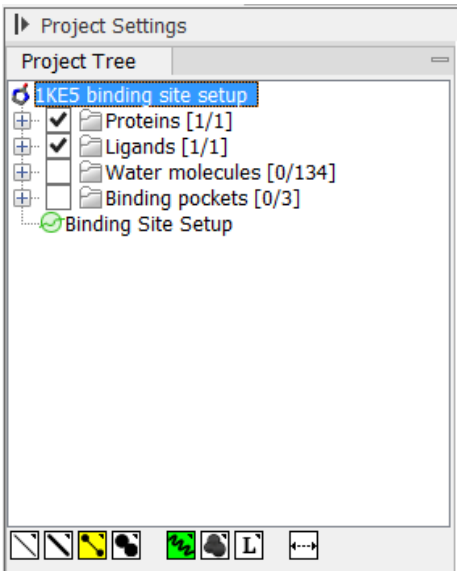


Figure 9.45: *"Binding Site Setup" in the Project Tree.*

In the following, each category in the Setup Binding Site dialog box will be described in detail. The 3D view will zoom to the different elements, when they are selected in the dialog box. This can be avoided by unchecking the Auto-zoom box in the lower left corner of the dialog box.

Binding site

The ligand binding mode search is carried out inside the binding site volume. It is specified using a sphere, which should be centered around the expected binding pocket, and have a radius

large enough to accommodate all ligands docked to the protein. The binding site is shown as a transparent green sphere when invoking the Setup Binding Site action from the **Molecule Project** Side Panel (figure 9.46).

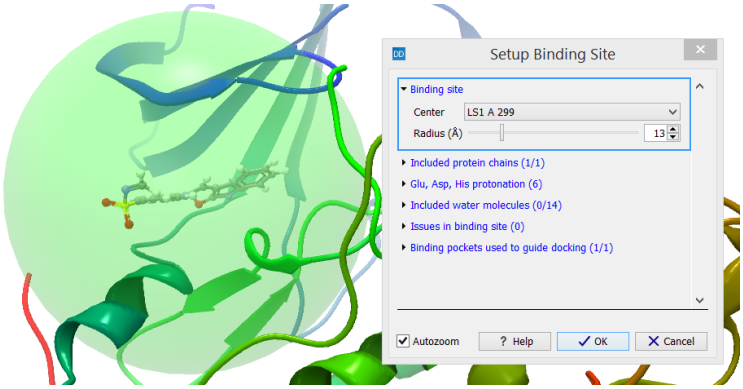


Figure 9.46: *The binding site is shown as a transparent green sphere when invoking the Setup Binding Site action.*

The center of the binding site can be selected from a list of the ligands, binding pockets and atom groups contained in the **Molecule Project**. Alternatively, the mouse can be used to select a number of atoms from the 3D view, e.g. residues known to be involved in binding using the 'Current atom selection' option. The radius of the sphere can also be adjusted. The sphere should be as small as possible, to improve the docking search efficiency, but large enough to ensure that proper ligand binding modes will not be discarded due to atoms not being able to fit inside the sphere.

Included protein chains (X/Y)

The category header indicates the number of protein chains that are included in the binding site setup (X) as well as the number of protein chains that have atoms inside the binding site volume (Y) (figure 9.47).

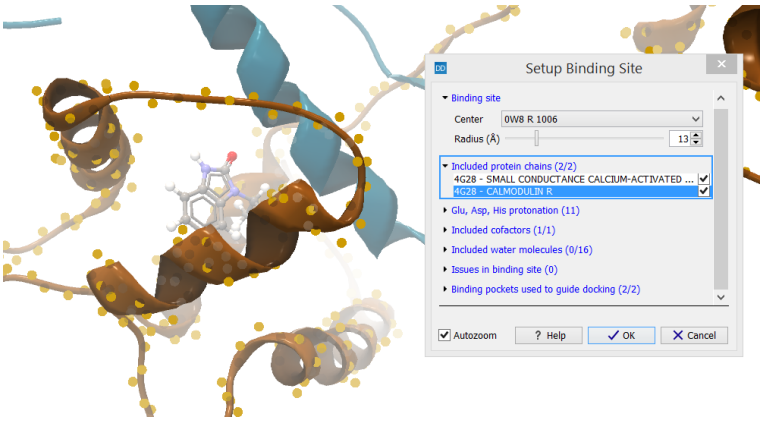


Figure 9.47: *In the category header is indicated how many protein chains are selected to be part of the binding site setup (X) and how many protein chains have atoms inside the binding site volume (Y).*

The protein chains found inside the binding site volume are listed. Selecting a row in the list will fit the view around the protein chain and highlight it. All protein chains found inside the binding site volume are included per default, but individual chains can be removed from the setup by unchecking the boxes next to them in the dialog box. This will not remove them from the **Molecule**

Project, only from the binding site setup.

Glu, Asp, His protonation (X)

The "Glu, Asp, His protonation" category header indicates how many Glu, Asp, and His residues (X) that are found inside the binding site volume on the included protein chains. These residues are interesting as the protonation state of the side chain depends on the local environment. Per default, Glu and Asp (pKa~4) are setup with no protons on the acid group, and His (pKa~6) with a single proton positioned at Nδ (figure 9.48).

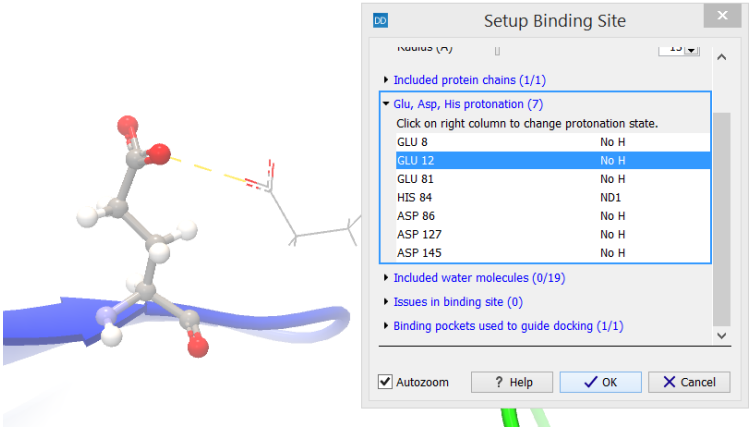


Figure 9.48: *The protonation state is shown in the right-most column.*

The residues are listed, and selecting an entry in the list will zoom to that residue and show it in a ball-and-sticks representation. The protonation of that residue can then be changed by clicking on the right-most column, which will make a list of protonation choices appear. Selecting an alternative protonation will immediately update the view (figure 9.49). If the selected side chain can form hydrogen bonds, these are automatically shown with blue dashed lines. Hydrogen bonds that could be formed if the protonation was changed are shown with yellow dashed lines.

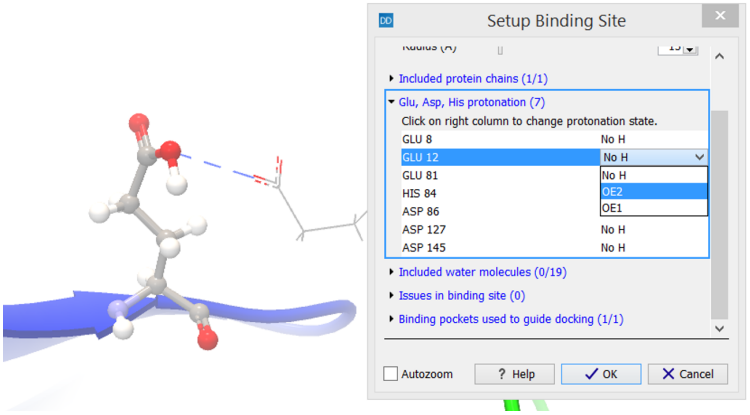


Figure 9.49: *The protonation of a residue can be changed by clicking on the right-most column, which will make a list of protonation choices appear.*

You can see more residues at the same time by selecting multiple entries from the list.

Note! Changing the protonation of residues will take immediate effect and will not revert if you press "Cancel" on the Setup Binding Site dialog box. However, like all other changes to molecules, it can be reverted using the "Undo" button.

Included cofactors (X/Y)

The category header indicates the number of cofactors that are included in the binding site setup (X) as well as the number of cofactors that have atoms inside the binding site volume (Y) (figure 9.50).

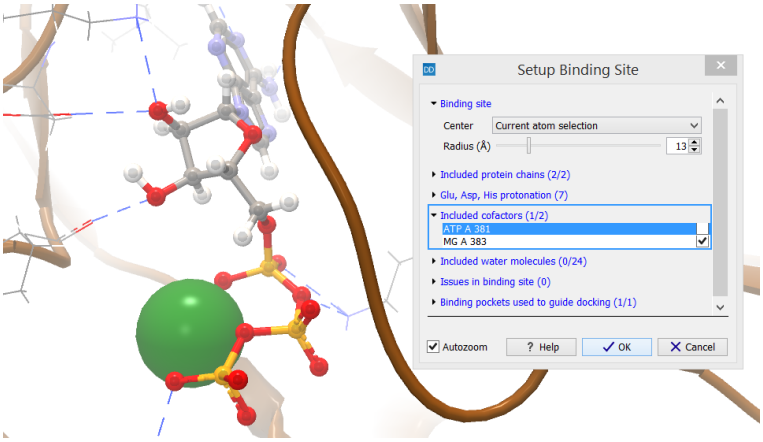


Figure 9.50: *Cofactors within the binding site volume are included in the setup per default.*

The cofactors found inside the binding site are listed. Selecting a row in the list will zoom the view around it and high-light it. All cofactors found inside the binding site are included per default, but individual cofactors can be removed from the setup by unchecking the boxes next to them in the dialog box. This will not remove them from the **Molecule Project**, only from the binding site setup.

Included water molecules (X/Y)

The category header indicates the number of water molecules that are included in the binding site setup (X) as well as the number of water molecules that are found inside the binding site volume (Y) (figure 9.51).

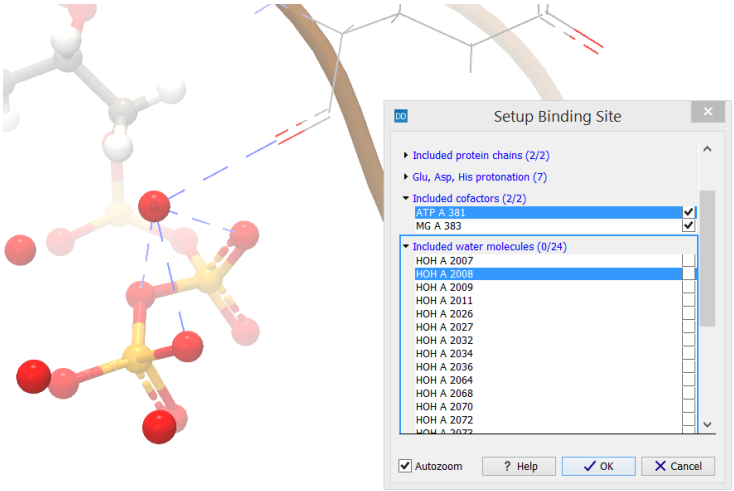


Figure 9.51: *The category header indicates the number of water molecules that are included in the binding site setup (X) as well as the number of water molecules that are found inside the binding site volume (Y).*

The water molecules found inside the binding site are listed. Selecting one or more rows in the list will zoom the view around the selected water molecules and high-light them. If the

water molecules are currently not displayed, their position will be seen as brown atom-selection dots. Hydrogen bond interactions involving the selected water molecules and other molecules included in the binding site setup will automatically be shown. Per default, water molecules are not included in the binding site setup, but individual molecules can be included by checking the boxes next to them in the dialog box.

Note As all molecules included in the setup will stay rigid and take up space in the binding pocket during docking, only include water molecules that you know is important for binding.

Included nucleic acids (X/Y)

The category header indicates the number of nucleic acid chains that are included in the binding site setup (X) and how many nucleic acid chains have atoms inside the binding site volume (Y) (figure 9.52).

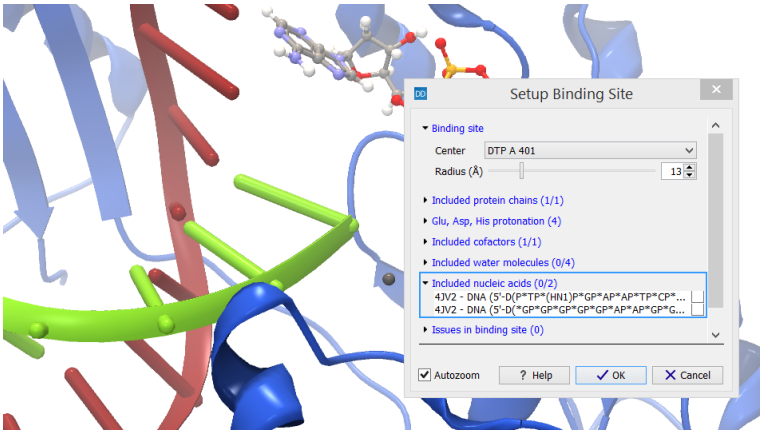


Figure 9.52: *Select nucleic acid chains to include in the target.*

The nucleic acids found inside the binding site are listed. Selecting a row in the list will zoom the view around the chain and high-light it. All nucleic acids found inside the binding site volume are included per default, but individual chains can be removed from the setup by un-checking the boxes next to them in the dialog box.

Note The docking score has not been optimized to dock ligands against nucleic acids, and it is not advisable to have nucleic acids as the main target for the docking.

Issues in binding site (X)

The category header indicates the number (X) of the issues in the Issues list (see section 6.2.9) that relates to atoms or molecules found inside the binding site volume and included in the Binding Site Setup (figure 9.53).

The issues found inside the binding site volume are listed. Selecting a row in the list will zoom the view around the implicated atoms and high-light them. If the issues relate to serious deficiencies in the binding pocket, another protein structure should be used to set up the binding site.

Binding pockets used to guide docking (X/Y)

The category header indicates the number of binding pockets that are included in the binding site setup (X) as well as the number of binding pockets that are found inside the binding site volume (Y) (figure 9.54).

Binding pockets are used to guide the docking simulations. Binding modes are enforced to have

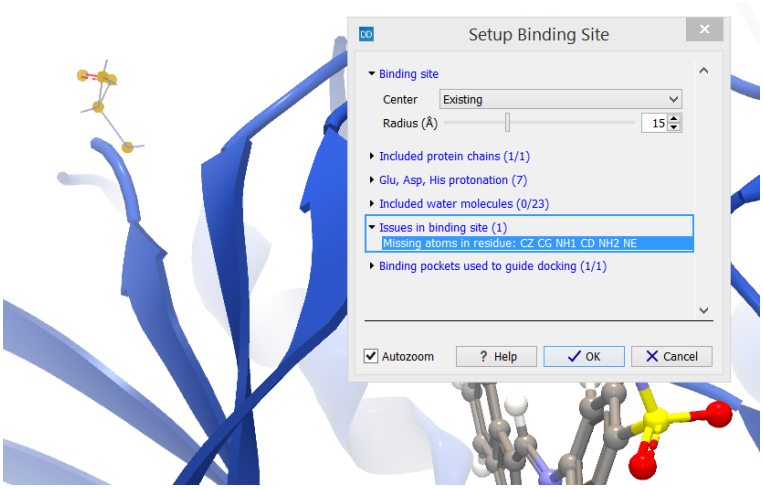


Figure 9.53: *The category header indicates the number (X) of the issues listed in the Issues list that relates to atoms or molecules found inside the binding site volume.*

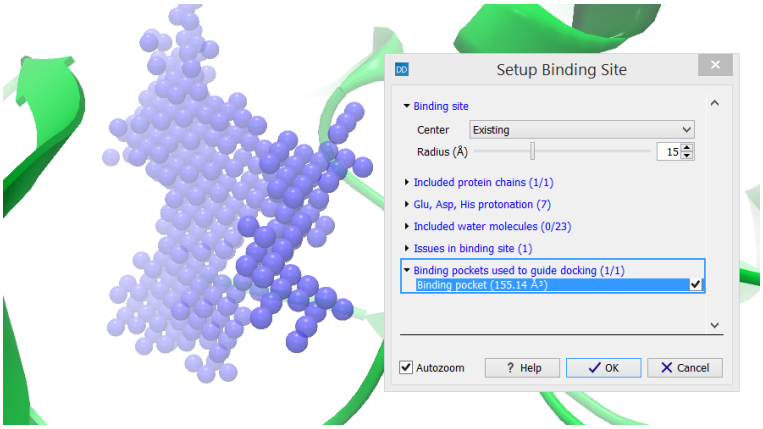


Figure 9.54: *The category header indicates the number of binding pockets that are included in the binding site setup (X) as well as the number of binding pockets that are found inside the binding site volume (Y).*

at least one ligand heavy atom inside a binding pocket included in the setup. It is therefore important that the binding pockets included in the setup actually correspond to where the ligand is expected to bind. All found binding pockets are per default included in the setup, but they can be removed from the setup by un-checking the boxes next to them. The algorithm evaluating the binding pockets is described in section 9.15.

9.12.2 Ligand docking using the Dock Ligands tool

To run the "Dock Ligands" tool:

Toolbox | **Drug Design** ( ) | **Dock Ligands** ( )

The Dock Ligands tool takes a **Molecule Table** containing small molecules (ligands), and a **Molecule Project** containing a Binding Site Setup (see section 9.12.1) as input. For each of the small molecules in the **Molecule Table**, the docking simulation searches for optimal binding modes to the binding site. The optimal binding mode together with a score of the binding mode are returned in a Docking Results Table (see section 9.8). The docking algorithms are described in section 9.12.4.

In the Dock Ligands wizard step 1 (figure 9.55), select the **Molecule Table** that holds the small molecules to dock. More than one table can be selected at the same time.

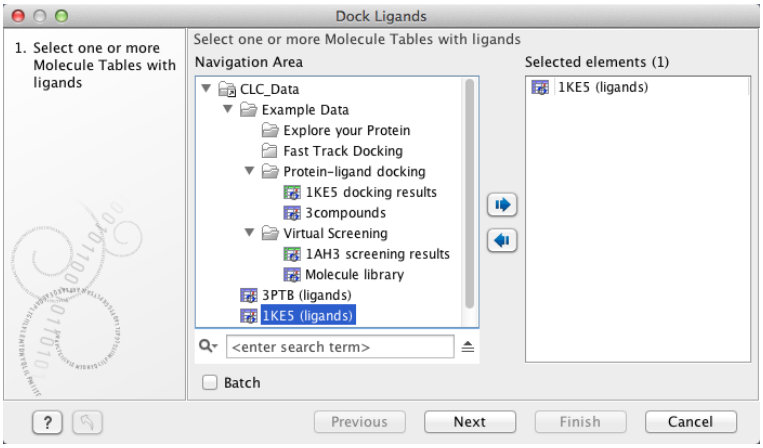


Figure 9.55: *The "Dock Ligands" dialog. Select the "Molecule Table" that hold the small molecules to dock.*

Click on the button labeled **Next** to select the **Molecule Project** and specify the docking parameters (figure 9.56).

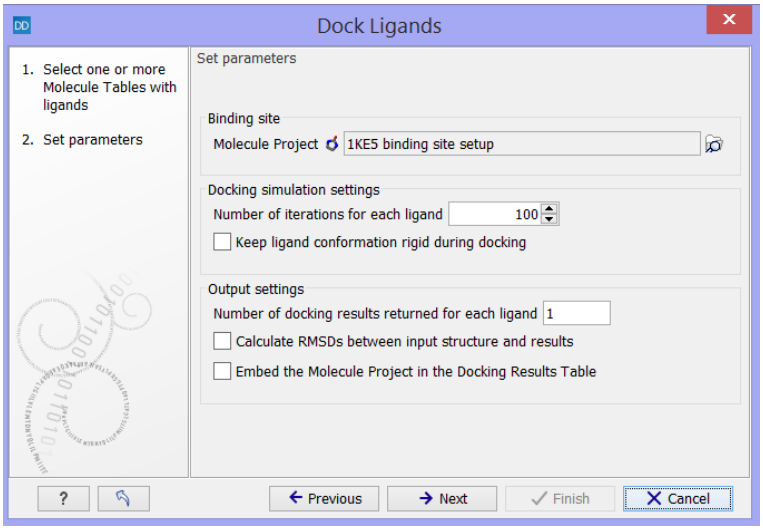


Figure 9.56: *Select the "Molecule Project" with the binding site setup and specify the docking parameters.*

This wizard has the following options:

- **Binding site**
  - **Molecule Project**. The **Molecule Project** holding the Binding Site Setup to dock against should be specified.
- **Docking simulation settings**
  - **Number of iterations for each ligand**. For each ligand, a number of iterations are carried out in the search for an optimal binding mode. The default value is considered to be a good balance between search completeness and cost. Increasing the number will

lead to a more extensive binding mode search, at the cost of increased computational time.

  - **Keep ligand conformation rigid during docking**. During docking, the conformation of the ligand is changed via rotation around flexible bonds. Some molecule libraries contain pre-generated conformations of the ligands, and in that case, the ligand conformations should not be altered during the docking, and this option should be checked.

- **Output settings**

  - **Number of docking results returned for each ligand**. Per default, only the best scoring binding mode is returned for each ligand. As the scoring is a simple description of molecular interaction, and thus not perfect (see section 9.12.4), it can be relevant and interesting to inspect more than just the top ranked binding mode. If the number of docking results to return for each ligand is set higher than one, other optimized top ranked binding modes will be returned, together with the best ranked, in the Docking Results Table. If the returned binding modes are very similar, it is a sign that alternative binding modes with good scores could not be found in the binding mode search.
  - **Calculate root mean square deviations (RMSDs) between input structure and results**. Usually, you will start a docking study by docking the co-crystallized ligand, if such exists, in the binding site. If the binding mode resulting from the docking is similar to the binding mode of the co-crystallized ligand, it is a sign that the binding site setup is safe to use for docking small molecules similar to the ligand co-crystallized with the protein. The root mean square deviation (RMSD) is measured between the heavy atom positions of the ligand in the co-crystallized and docked binding modes.
  - **Embed the Molecule Project in the Docking Results Table**. Selecting this option will save a copy of the **Molecule Project** used as input together with the Docking Results Table output from the simulation. Then it will always be possible to inspect the resulting binding modes in the binding site setup used for the docking (see section 9.7.3), and to see which molecules were included in the setup (see section 9.12.1). However, it will take up more disk space, and it is therefore not set as default.

Note! The wizard will remember the user defined settings from run to run. If you wish to bring the settings back to the default settings, this can be done by clicking on the small arrow button in the lower left corner of the dialog.

Click on the button labeled **Next** and consider whether the Docking Results Table should open when the docking is done, or if it should be saved in the Navigation Area. A log file of the simulation process can also be made.

If you choose to save the results, the next step will allow you to specify where to save it, otherwise click on the button labeled **Finish**, and the docking simulation will start. You can see how it proceeds in the Processes tab in the Toolbox area.

9.12.3 Ligand docking from the Project Tree

The Dock Ligands tool can be invoked from the context menu on one or more ligands selected in the **Project Tree** in a **Molecule Project** or from clicking the Dock Ligand button below the Project

Tree. The docking simulation will then start right away, without access to the wizard settings (figure 9.57).

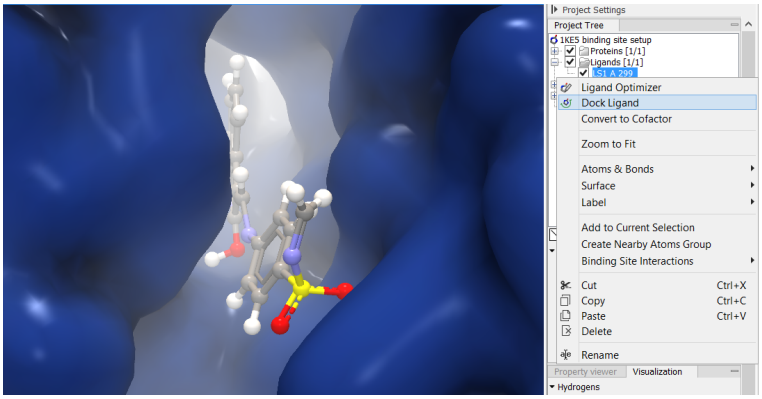


Figure 9.57: *Dock ligands with one click.*

The selected ligands will be docked to the Binding Site Setup found in the same **Molecule Project**. The number of iterations for each ligand will be as the default in the Dock Ligands wizard. No Docking Results Table will be created, instead the best scoring docking result is returned in the Docking results category in the **Project Tree**. Selecting a docking result in the **Project Tree** will show the docking score in the Property viewer in the **Molecule Project** Side Panel, together with other molecule information.

9.12.4 The docking algorithms

In the docking simulation, a number of variables defining the ligand binding mode are optimized, to achieve a good docking score. There are thus two important aspects of the docking simulation, namely the search algorithm and the docking scoring function.

The docking scoring function

A perfect search will return the binding mode with the best possible docking score. Docking results will therefore never get better than the scoring function used. As the scoring function has to be evaluated for numerous different binding modes, the complexity of the function influences greatly how much time is needed to do the docking simulation. The scoring function should therefore be as simple as possible, while still being able to distinguish between favorable and poor protein-ligand interactions. The docking score used in the Drug Discovery Workbench is the $\text{PLANTS}_{\text{PLP}}$ score [Korb et al., 2009]. This score has a good balance between accuracy and evaluation time. The score mimics the potential energy change, when the protein and ligand come together. This means that a very negative score corresponds to a strong binding and a less negative or even positive score corresponds to a weak or non-existing binding.

$$\text{Score} = S_{target-ligand} + S_{ligand}$$

The score is listed in the Docking Results Table as "Score". The $S_{target-ligand}$ term is a sum over contributions from all heavy atom contacts between the ligand and the molecules included in the binding site setup. It scores the complementarity between binding site and ligand by rewarding and punishing different types of heavy atom contacts (inter atom distance below ~5.5 Å).

Five different types of contacts are defined:

Rewarded contacts

1. Hydrogen bond interactions
2. Lone-pair - metal ion interactions
3. Non-polar interactions

Punished contacts

4. Non-polar - polar contacts
5. Repulsive contacts:
    - Hydrogen bond donor-donor contacts
    - Hydrogen bond donor-metal contacts
    - Hydrogen bond acceptor-acceptor contacts

These contact types are associated with a pairwise linear potential (PLP), which determines the distance-dependence of the contribution to the score. This is shown in figure 9.58.

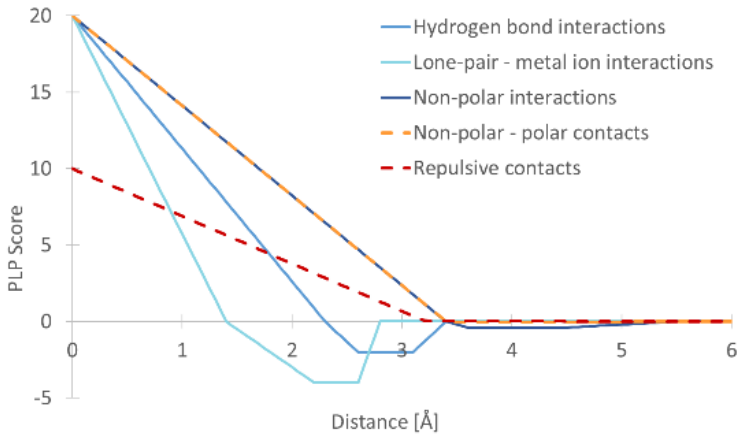


Figure 9.58: *The pairwise linear potentials defining the score contributions for different types of heavy atom contacts.*

Some of these contributions to the score can be displayed in the Docking Results Table. They are accessible from the 'Show column' category in the Side Panel. The contribution from hydrogen bond interactions is listed as 'Hydrogen bond score'. The contribution from lone-pair - metal ion interactions is listed as 'Metal interaction score'. The non-polar interactions, non-polar - polar contacts, and the repulsive contacts are combined in the 'Steric interaction score' contribution. The S_{ligand} term punishes internal heavy atom clashes in the ligand and strain resulting from unfavorable bond rotations. This contribution to the score is also accessible from the table Side Panel, and is listed as 'Ligand conformation penalty'.

As the score is a sum over contributions, a large ligand can get a better score than a small one, simply due to its size. When comparing scores for different molecules, this effect has to be considered and kept in mind.

The search algorithm

In a docking simulation, the variables to optimize are those that define a binding mode. Namely, the rotation angle for all rotatable (flexible) bonds in the ligand (figure 9.59), the position of the ligand within the binding site (translation), and the overall rotation of the ligand with respect to the protein. It is not feasible to do an exhaustive search with this number of variables.

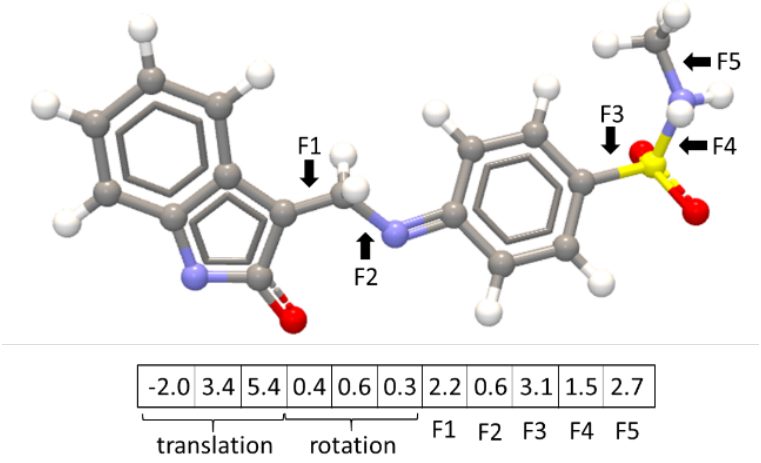


Figure 9.59: *Variables defining a binding mode.*

It is the docking scoring function, described above, that should be minimized with respect to the variables. The minimization problem can be envisioned as a landscape with hills and valleys, and the job is to find the deepest valley while being blindfolded, and without having to visit every point on the landscape. If there were only two variables to optimize, it would be like a familiar landscape with a latitude and a longitude to specify a point. When there are more variables, as in the case of a binding mode, the landscape is much more complex to search, with many hills and valleys in many dimensions. Fortunately, the same methods can be used as when searching for a minimum in two dimensions. The problem can be divided in two; finding a point close to the deepest valley (global optimization problem) and finding the deepest point in a valley close-by (local optimization problem).

In the docking simulation, the global search space is sampled in a number of iterations, combined with local optimizations. The number of sample iterations can be adjusted in the Dock Ligands and Screen Ligands wizards (Number of iterations for each ligand). Each iteration is independent of the others, and below is described the operations carried out for each sample iteration.

A sample iteration: A population of 20 potential binding modes is generated, by assignment of random values to the variables defining a binding mode. If binding pockets are included in the Binding Site Setup (section 9.12.1), the binding modes are required to have at least one heavy atom inside a pocket. A binding mode initially positioned outside the binding pockets is therefore translated so as to fulfil this requirement. For each of the 20 binding modes, a local optimization with respect to the docking score is carried out, using the simplex method for function minimization [Nelder and Mead, 1965] (the simplex method is described in detail below). This is a search for the deepest point in a valley close to the initial random position. The best scoring of the 20 resulting binding modes is selected for a more refined simplex minimization. Finally, this optimized binding mode is returned from the iteration. The binding mode returned from the iteration is compared to the best scoring binding mode found so far in all iterations. If the new found binding mode has an even better score, it is saved, otherwise it is discarded. Now

a new iteration can start. Each sample iteration starts from a random point in the search space. The overall search is therefore stochastic in nature, and the final result will not be exactly the same between executions, even when the same input and settings are used. However, if the results are not highly similar, it is a sign that the sampling is not sufficient (increase the 'Number of iterations for each ligand' parameter), or that a number of binding modes are equally valid.

As the iterations are independent, the overall number of iterations can be split in several 'runs', each taking their share of the iterations. This is done automatically to exploit all available CPU cores. If 'Number of docking results returned for each ligand' is set to more than one in the Dock Ligands or Screen Ligands wizards, the docking simulation is divided in at least this number of runs, and the best scoring binding modes from each of these runs are kept and returned, instead of only returning the overall best scoring binding mode.

Simplex method for function minimization: The simplex search starts from the point specified by random values [*i*, *j*, ...] assigned to the variables. To initialize the simplex minimization, *N* new binding modes should be generated (where *N* is the number of variables, see figure 9.59). These binding modes should not be too different from the input binding mode, so they are generated by shifting each variable in the initial binding mode with a small offset, δ:

$$\begin{bmatrix} i \\ j \\ \vdots \end{bmatrix}; \begin{bmatrix} i+\delta_i \\ j \\ \vdots \end{bmatrix}; \begin{bmatrix} i \\ j+\delta_j \\ \vdots \end{bmatrix}; \ldots$$

For all *N* + 1 binding modes (points in the search space), the docking score (the function value) is evaluated.

The example below illustrates the fundamental steps in the search process. The example only has two variables (*N* = 2), and the color scale indicates the function values (the score, in our case). Hilltops are dark blue and valleys are dark red.

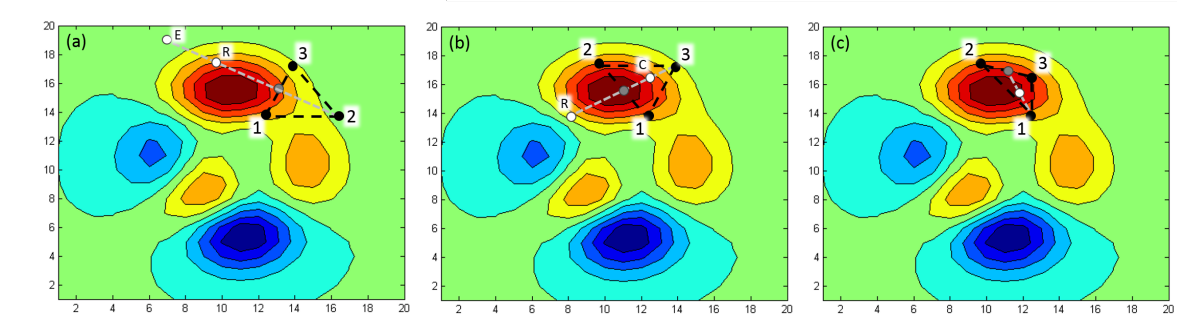


Figure 9.60: *Schematic of steps taken in a Simplex local optimization.*

In (a), the initial three (*N* + 1) points have been evaluated.

For each step in the simplex minimization, the point with the highest function value is moved to a new position.

To find the new position, the point with the highest function value is reflected in the centroid formed by all the other *N* points. The centroid is the mean position of the *N* points, in all of the coordinate directions. In a two dimensional search, the centroid is therefore the point on the line, midway between the two best scoring points (the gray points in figure 9.60). In (a), point 2 has the highest function value, and is reflected through the centroid to the white point marked

"R". The function is evaluated at this point, and if it is lower than the original point, the step is taken. However, if the value is lower than for all the other points, as in this case, it could be a sign that the step was straight downhill. The reflection is therefore "extended", so as to take a larger step.

In (a), the extended point is white and marked with an "E". The function is evaluated in the extended point, and if this point is better than the reflected point, the extended step is taken instead. In figure 9.60, the extended point is worse than the reflected, and therefore the reflected point is kept.

In (b), point 2 has moved to its new position, and the point with the highest function value is now 3. When 3 is reflected in the centroid between the other points to the white point marked "R", the function value is slightly higher than in the original point. The reflected step is therefore not taken. Instead, a "contracted" step is taken, shown in (b) as a white point marked "C". The contracted point is on the reflection line, but halfway between the centroid point and the original point.

In (c), point 3 has moved to its new position, and the point with the highest function value is now 1. The reflected point is not indicated on the figure, but it is clearly uphill from the other points, and is therefore rejected. Instead, the contracted step is taken, indicated as the white point.

In this way, the search continues until the difference between the highest and lowest docking scores (for the $N + 1$ points), divided by the size of the lowest score, is below 0.01 (convergence), or until 2000 steps have been taken.

The refined simplex minimization, carried out for the best scoring binding mode of the 20 in the iteration population, is different from the process described above in two ways: 1) The initial point is not based on random values, it is instead the optimized binding mode returned from the initial simplex minimization. 2) The search is continued until the difference between the highest and lowest scores (for the $N + 1$ points), divided by the size of the lowest score, is below 0.0001, or until 2000 steps have been taken.

9.12.5 Inspecting docking results

A docking result consists of a ligand binding mode and the connected docking score. No matter if the docking result is found in a **Docking Results Table** or in a **Molecule Project**, there is easy access to both aspects of the result.

Docking score and other data

In a **Docking Results Table**, each entry represents a docking result, and the docking score is found in the column with the title "Score". From the table Side Panel, columns are accessible that show how the score can be broken down into the contributions described in section 9.12.4. In the table you can also find a column listing how many flexible bonds each ligand have. The conformation of the ligand is changed in the docking simulation, through rotations around the ligand flexible bonds. Furthermore, the number of flexible bonds in a molecule is related to the ligand entropy loss on binding. If the "Calculate RMSDs between input structure and results" is selected in the Dock Ligands wizard, the Docking Results Table will also contain a column with RMSD values. Please note that the RMSD values are only relevant for ligands that are docked into their initial position, as is the case for the docking of a co-crystallized ligand.

In a **Molecule Project**, the docking score and number of flexible bonds are displayed in the Side

Panel **Property viewer**, when the docking result is selected in the **Project Tree**.

Ligand binding mode

In a **Molecule Project**, a docking result entry can be displayed together with the protein and other molecules in the project, to visualize the binding mode of the ligand in the binding site. To visualize the ligand interaction in atomic detail, select the relevant docking result entry in the Project Tree and invoke the right-click context menu. From the menu, pick **Binding Site Interactions**. If this option is not found in the menu, click the Setup Binding Site button below the Project Tree and define the binding site. There are two options for showing binding site interactions; **Show Hydrogen Bonds** and **Create Interacting Atoms Group**. Use the **Show Hydrogen Bonds** option to display protein residues forming hydrogen bonds to the ligand (in wireframe), with the hydrogen bonds shown as blue dashed lines. The hydrogen bonds can be hidden again invoking the context menu on the docking result and selecting **Binding Site Interactions | Hide Hydrogen Bonds**. Use the **Create Interacting Atoms Group** option to generate a custom atom group consisting of protein residues and molecules, which have at least one heavy atom within 5 Å of a ligand heavy atom. The atom group appears in the Atom groups category in the **Project Tree**, and can be hidden using the check box next to it, and the visualization changed using the quick-style buttons found at the bottom of the Project Tree.

In a **Docking Results Table**, connect to a 3D view using the Select View action in the table Side Panel. Select either the "Embedded molecule project" if available, to see the ligand binding modes in a read-only version of the **Molecule Project** used as input for the docking, or select an open **Molecule Project**, where the binding site setup is present. See section 9.7.3 for a description of how to see hydrogen bonds and nearby atoms for ligands "visiting" a **Molecule Project** from a table.

9.13 Screen ligands

In virtual screenings, a large general library of drug-like compounds are docked one by one to the target protein. The top-scoring ligands are expected to present with ligands that will either be candidates for strong binders in their own right, or give insight into molecule elements that are particular important for strong binding to the particular binding pocket. The Screen Ligands tool use the same docking algorithms (see section 9.12.4) as the Dock Ligands tool (section 9.12.2). The Screen Ligands wizard (figure 9.61) thus needs the same input and the parameters in step 2 are the same, and have the same meaning, as for the Dock Ligands tool (section 9.12.2), with a few exceptions listed below.

Percentage of top ranked docking results to keep. A virtual screening can involve **Molecule Table(s)** with a very large number of compounds (> 100,000). The **Docking Results Table** will in that case take up a considerable amount of disk space - equivalent to the input table. This can also make the results table slow to work with when e.g. sorting the entries based on a particular column. In general, it will only be the top ranked ligands that are interesting to look at, and this option therefore discards all but the top ranked docking results. When the parameter is set to 10, only the 10 % best scored ligands will be included in the results table.

The option to **calculate RMSDs between input structure and results** has been removed, as this does not make sense to do for a screening library.

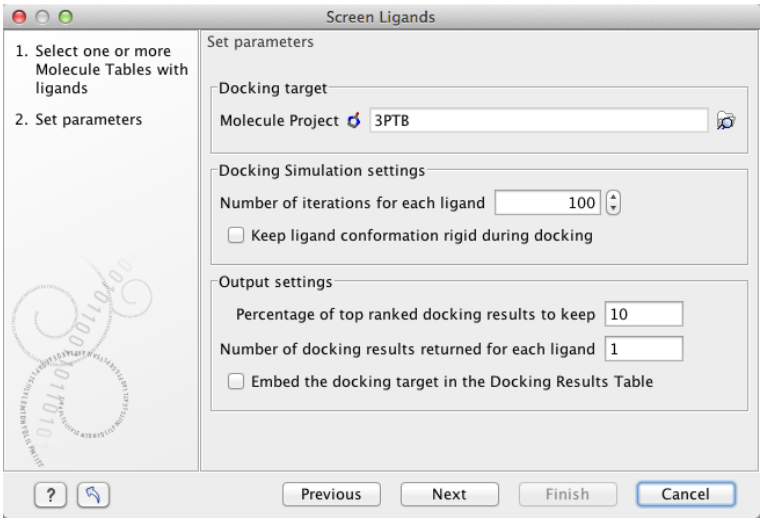


Figure 9.61: *Parameter settings for virtual screening.*

9.14 Improving docking and screening accuracy

This section gives some advice on how to do a successful docking or screening.

9.14.1 Docking

A ligand docking simulation sample how the ligand can fit into the binding site specified by the Binding Site Setup. When the simulation has ended, the best scoring binding mode will be returned as a result, together with the score (section 9.12).

It is always a good idea to check that experimental knowledge about protein-ligand binding can be reproduced or supported, by docking a known ligand into the binding site setup.

Examples of experimental knowledge:

- Ligand binding mode known from crystal structure
- Binding mode of a very similar ligand is known from crystal structure
- Some information about protein-ligand interaction is known, such as which amino acids are interacting with the ligand

To improve the docking accuracy, a number of points can be considered relating to the protein target (section 9.14.3), the ligand (section 9.14.4) and the docking simulation (section 9.14.6) itself.

9.14.2 Screening

The Screen Ligands tool found in the Toolbox does the same as the Dock Ligands tool, except a few options are different in the wizard. This means that many of the aspects to consider for improving a docking also apply to screenings.

The results obtained with a virtual screening should enrich the molecule library by ranking the molecules such that the top-ranked set have a higher percentage of hits than the overall library.

Just as for *in-vitro* screenings, a number of false positives and false negatives are to be expected for a virtual screening.

To improve the accuracy of the screening, a number of points can be considered relating to the protein target (section 9.14.3), the molecules in the screening library (section 9.14.5) and the screening simulation (section 9.14.7) itself.

9.14.3 Protein Target

Selecting the optimal protein target structure

The optimal protein structure to use for a docking or screening simulation is identical to the protein used in the experiments, binds to a ligand very similar to the one you are trying to dock or expect to see as top-scoring binders in a screening, and comes from a high quality crystal structure.

The BLAST tool can be used to search for alternative protein structures in the PDB, using the sequence corresponding to the protein used in the experiments as input (see section 6.2.5).

The protein structure and the protein in the experiments should ideally be from the same species

If the protein structure is from the PDB, the species information can be found on pdb.org.

If the protein structure is not from the same species, or if there are introduced mutations into the protein in the experiments, it should be checked that the binding site residues are at least conserved. To do that, find the protein sequence corresponding to the protein used in the experiments, e.g. using the "Search for Sequences in UniProt" tool (see section 11.1). Extract the protein sequence of your protein structure using the "Show Sequence" option in the Side Panel of the Molecule Project (section 9.4.1). Make a sequence alignment of the two sequences using the "Create Alignment" tool (see section 14.1). If the sequences have high sequence identity, and all amino acids found in the binding site are conserved, the structure should be fine to use.

The protein structure resolution should be as high as possible

If the protein structure is from the PDB, the resolution information can be found on pdb.org, or using the "Search for PDB Structures at NCBI" tool in the workbench.

The general uncertainty in crystal structures with lower resolution (e.g. >2.5 Å) makes it hard to reproduce a ligand binding mode to a high accuracy (<2 Å RMSD). For some low resolution structures, atoms and whole side chains are missing. If this results in an incomplete structure in the binding site, then the protein structure is not fit to be used for docking. For other low resolution structures, the side chains have been modeled fully, even though their positions are not well determined. That may result in a poor representation of the binding site. Try representing the protein or just the binding pocket residues using one of the Atom and Bonds visualizations, and select the "Color by Temperature" color scheme (section 9.2). Atom positions not well defined in the protein structure will show up in clear red color, while very well determined atom positions will be blue.

The protein structure should be in complex with a ligand in the binding site

A functioning protein can in some cases take up different overall conformations, depending on whether it binds to a substrate or not. The local conformation of side chains in the binding pocket

can also depend on which type of molecule is binding. The protein structure should therefore ideally be interacting with something in the binding pocket similar to the ligand being docked. If the binding site is empty in the structure, it may be that side chain conformations are not optimal for ligand interaction.

9.14.4 Ligand

Ligand structure and representation

Except for rotations around single bonds, the 3D structure of the ligand is not changed during the docking simulation.

Bond lengths and angles, as well as planarity around atoms participating in double and delocalized bonds and ring structures, should therefore be correct and sensible in the initial structure.

Some problems with the ligand structure and representation will give rise to Issues being raised on import. Right-click in the Molecule Table or Molecule Project and select:

Show | Issues

and sort the issues table by Molecule, to see if any issues are raised for the ligand.

Ligand structures can be imported to the workbench from PDB, Mol2, or SDF files holding 3D information about the molecules, or generated in the workbench from 2D information or from SMILES representation (see section 6.2). Many ligand 3D structures are available from http://zinc.docking.org/. If there are problems with the ligand structure, try one of the alternative options for importing or generating a ligand structure, and see if the situation improves.

Based on bond orders, atom hybridization and hydrogen atoms, all heavy atoms are assigned a type prior to the docking simulation. This type determines what target interactions are favorable. It is therefore important that the bond orders, atom hybridization and number of hydrogen atoms on the heavy atoms correspond to the expected chemical state of the ligand when interacting with the target.

The representation can be adjusted from the right-click context menu on individual atoms. This will not change the structure. Alternatively, the Ligand Optimizer (section 9.11) can be used to change representation and structure of the ligand.

Rotatable bonds

The docking simulation samples different positions of the ligand in the binding site as well as different combinations of rotations around single bonds in the ligand. The more rotatable/flexible bonds a ligand have, the more different combinations to test. If a ligand has more than around eight rotatable (flexible) bonds, it can require much more sample iterations to find the optimal binding mode.

The number of rotatable (flexible) bonds can be seen in the *Property viewer* when selecting a *Docking result* in the Project Tree, or from the *Flexible bonds* column in *Docking results tables*.

If the ligand has many flexible bonds, make sure that all single bonds in the molecule can actually be expected to have low energy barriers for rotation, otherwise, change the single bond into a delocalized bond as for delocalized pi-systems (using options from the right-click context menu

on one of the atoms participating in the bond). In this way the initial geometry for this part of the molecule will be kept during the docking simulation.

To increase the chance of observing the correct binding mode for a ligand with many rotatable bonds, you can try increasing the number of docking simulation iterations for each ligand to e.g. 500 (default is 100). This can be done using the Dock Ligands tool from the Toolbox, having the ligand in a Molecule Table (use the Extract Ligands tool to extract ligands from a Molecule Project and into a Molecule Tably).

9.14.5 Screening library

The screening library consists of ligands in one or more Molecule Tables. Each ligand in the library should have sensible structure and representation, and not too many rotatable bonds (see section 9.14.4). However, it will typically not be feasible to check molecules in a library one by one in detail. You should check at least some of the molecules, to convince yourself of the general quality of the library. To do that, connect the Molecule Table to a 3D view from the Side Panel (see section 9.7.3), and select a couple of entries to visualize the structure and representation of the molecules. Browse through the issues raised on import or generation of the library (right-click on table and Show | Issues). The list of issues will also be connected to the 3D view, so that the relevant molecule will be displayed, and the implicated atom(s) selected.

If only a subset of the molecules in the library are of poor quality, you can just delete those entries from the Molecule Table, to avoid spending time on screening them.

If there are molecules in the library that you expect to see binding with a high score, make sure to check their structure and representation explicitly (see section 9.14.4).

9.14.6 Docking simulation

The Dock Ligand option found in the Side Panel of Molecule Projects carries out a docking simulation the same way as the Dock Ligands tool in the Toolbox, using the default settings. Using the Dock Ligands tool from the Toolbox, a couple of parameters can be adjusted that might improve the outcome of the docking. This requires that the ligands are found in a Molecule Table (can be extracted from a Molecule Project using the Extract Ligands tool).

Sufficient sampling

The number of docking simulation iterations is sufficient when the returned binding mode is consistent from one docking simulation to another. The default of 100 iterations will often be fine, but otherwise try increasing to 500 or even more iterations. If the ligand has very many rotatable bonds and/or the binding site is too flat, a proper sampling may not be feasible.

Limitations in the scoring function

The scoring of binding modes is not perfect. It can therefore happen, that the correct binding mode is not the best scoring, even though it is still recognized with a good score. To see more than just the best scoring binding mode, set the "Number of docking results returned for each ligand" to more than one.

9.14.7 Screening simulation

While increasing the sampling for some of the molecules in the library could improve the screening accuracy, just as for a regular docking simulation, it will in general not be of interest to do a thorough sampling for each molecule in a screening, as the size of the library will make it too time consuming.

Ranking the binding of different molecules is a hard challenge for all scoring functions. A larger molecule will tend to get a higher score, due to its ability to form more interactions with the protein. It is good to keep this in mind, when studying the top scoring compounds.

9.15 Find potential binding pockets

A molecular docking is aimed at a specific region of the target protein, expected to be the binding site. You may know the location of this site based on the position of ligands or cofactors co-crystallized with the protein structure, or positions of amino acids known to participate in the binding. If you don't know the binding site, you can use the "Find Binding Pockets" tool to search for cavities on the protein surface.

To run the "Find Binding Pockets" tool:

Toolbox | **Drug Design** ( ) | **Find Binding Pockets** ( )

This will open up the wizard that allows you to specify the **Molecule Project** holding the target protein that should be used as input.

Note! All protein chains in the **Molecule Project** will be considered in the evaluation, and protein chains, which are not biologically relevant (e.g. crystal packing units), should therefore be deleted before the search for binding pockets is initiated.

Click on the button labeled **Next**.

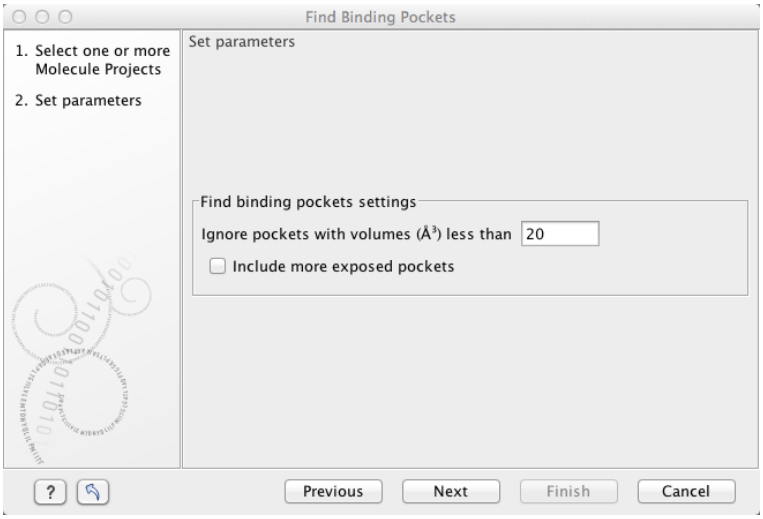


Figure 9.62: *Parameter settings in the Find Binding Pockets tool.*

In the wizard step 2 (figure 9.62), the following parameters can be adjusted:

- **Find binding pockets settings**

- **Ignore pockets with volumes (Å^3) less than**. Pockets with volume below the given threshold are ignored. Analysis of 5600 protein-ligand structures from the PDB has revealed that 95 % of binding sites are within one of the three largest solvent-accessible pockets found on the protein [Li et al., 2008]. Typically, it is therefore only relevant to look at the largest pockets, and the smaller pockets can safely be ignored.
- **Include more exposed pockets**. Per default, only compact pockets will be found. Enabling this option will include more exposed pockets in the search. Pockets with good drug-binding properties are typically compact, but not always [Cheng et al., 2007a], and it can therefore sometimes be relevant to look for the more exposed pockets too.

Click on the button labeled Finish. After a while, binding pockets appear in the **Project Tree** of the **Molecule Project** that was used as input. You can use the check boxes in the **Project Tree** to display the identified binding pockets one by one. This is shown in figure 9.63.

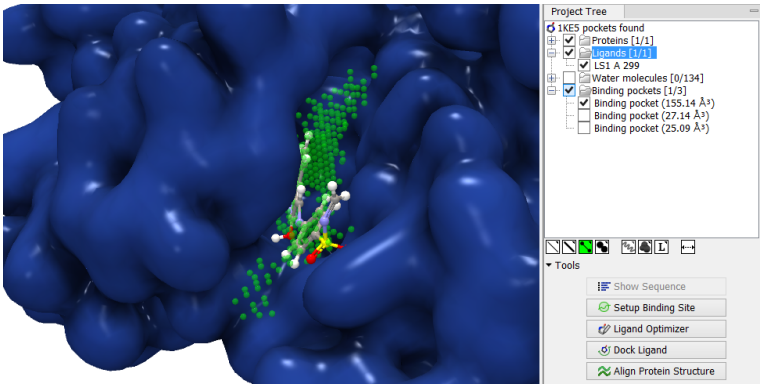


Figure 9.63: *Binding pocket indicated with green spheres.*

Note! If no pockets are found, a notification will appear in the workbench, and nothing will be added to the **Project Tree**. You can increase the chances of finding pockets by decreasing the volume limit for when a pocket is ignored, or include more exposed pockets.

9.15.1 The Find Binding Pockets algorithm

The same algorithm is used to find binding pockets in the **Find Binding Pockets** tool and in the **Setup Binding Site** dialog box. In the case of the **Find Binding Pockets** tool, the target is all protein chains in the **Molecule Project** combined. In the case of the **Setup Binding Site** dialog box, the target is all molecules included in the binding site setup. First, a discrete 3D grid with a resolution of 0.8 Å, covering all target molecules, is created. At every grid point, a sphere is placed with a radius of 1.4 Å. If this sphere overlaps with any of the target atoms (represented by their van der Waals radii), the grid point is part of the inaccessible volume. All other points are referred to as accessible.

Second, each accessible grid point is checked to see whether it is part of a pocket, using the following procedure. 16 directions are selected uniformly, and these are followed away from the point (see the red and orange lines in the 2D example depicted in figure 9.64). If an inaccessible grid point is hit, before hitting the grid boundary, it is registered. If more than 12 of the directions hit inaccessible volume (or only 8, if the option "Include more exposed pockets" is selected in the "Find Binding Pockets" tool wizard), then the point is marked as being part of a pocket (visualized as green or blue dots in the 3D view).

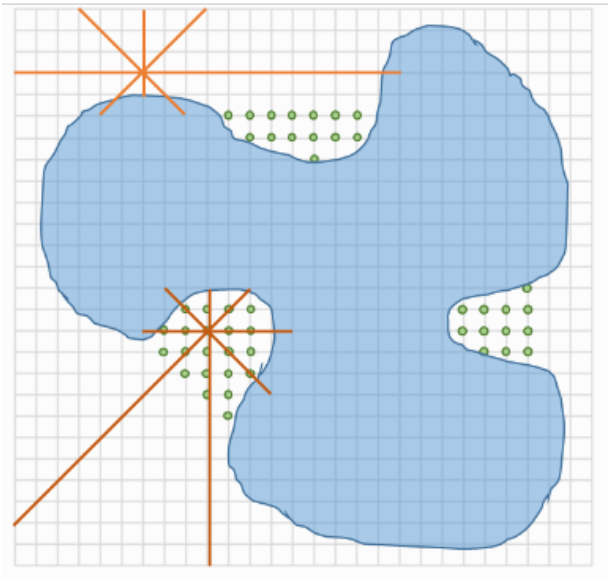


Figure 9.64: *2D schematic of the* Find Binding Pockets *algorithm. The blue area represents the target, and all grid points covered in blue are inaccessible. The probe directions are indicated for both a pocket point (red lines) and an exposed point (orange lines). Green dots indicate pocket points.*

The final step is to group the neighboring grid points to form the individual pockets. The volume of a pocket is then estimated as the number of grid points belonging to the pocket times the volume of a unit grid cell. It is typically only relevant to see the largest pockets, and the smaller pockets can safely be ignored. The "Ignore pockets with volumes (Å^3) less than" option in the "Find Binding Pockets" wizard, thus specifies the limit for discarding the smallest pockets found. When binding pockets are found in the Binding Site Setup, all pockets with volume less than 10 Å^3 are discarded.

9.16 Calculate molecular properties

Using the Calculate Molecular Properties tool you can calculate commonly used properties of small molecules, such as Lipinski's rule of five or log P. These properties can be useful for identifying potential drug-like molecules, or for removing non drug-like molecules from a compound library before starting a large virtual screening experiment.

To run the "Calculate Molecular Properties" tool:

Toolbox | **Drug Design** ( ) | **Calculate Molecular Properties** ( )

The Calculate Molecular Properties tool takes one or more **Molecule Tables** as input (figure 9.65).

Note! Molecular properties will only be calculated for small molecules, i.e. ligands and docking results. All other molecules present in the tables will be ignored.

Click on the button labeled **Next**.

In the wizard step 2 (figure 9.66), the following molecular properties can be selected:

- **Log P**. The partition-coefficient log P is a measure of lipophilicity and is one of the criteria used to assess the druglikeness of a given molecule. See section 9.16.1 for a short

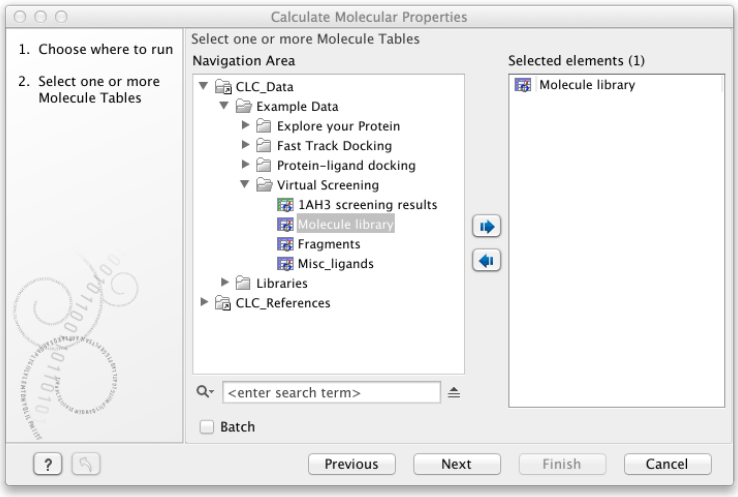


Figure 9.65: *Select Molecule Tables containing small molecules (ligands or docking results).*

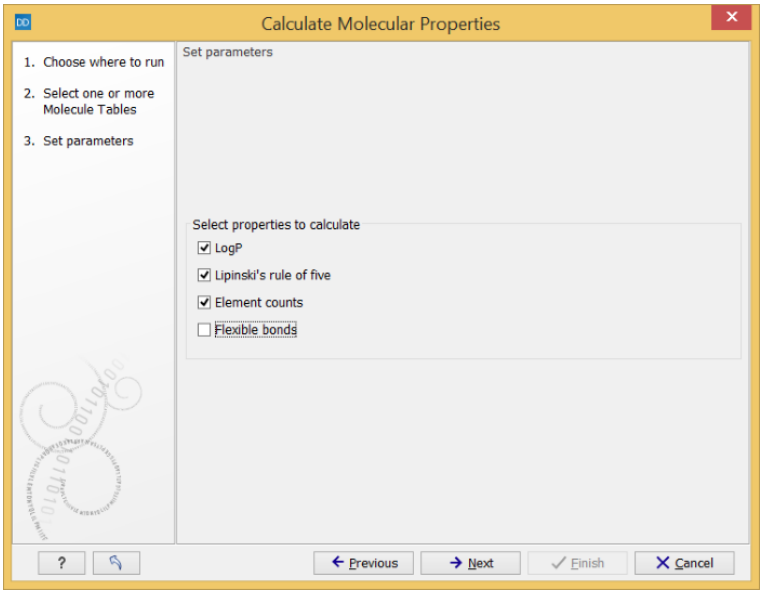


Figure 9.66: *Molecular properties available in the Calculate Molecular Properties tool.*

description of the log P prediction algorithm used in *CLC Drug Discovery Workbench*.

- **Lipinski's rule of five**. Druglikeness of a given molecule is an important property. Lipinski formulated a simple set of rules that can be used to evaluate if a molecule has properties that would make it a likely orally active drug in humans [Lipinski et al., 2001]:
  - Not more than 5 hydrogen bond donors (counted as the total number of nitrogen-hydrogen and oxygen-hydrogen bonds)
  - Not more than 10 hydrogen bond acceptors (counted as the total number of nitrogen and oxygen atoms)
  - Molecular weight $<$ 500 daltons
  - Log P (octanol-water partition coefficient) $\leq$ 5

  The number of violations of the Lipinski rules gives an indication of how drug-like a given molecule is. In general, orally active drugs have fewer than two violations. When calculating

the Lipinski's rule of five property, the number of hydrogen bond donors/acceptors, molecular weight, log P, and total number of violations will be listed as properties in the output Molecule Table.

- **Element counts**. This option counts the number of H, C, N, O, P, and S atoms in the molecule. All other atoms are summarized in the 'Other atoms' column. These values can be useful if you are looking for the presence of a particular element or a certain composition of element types (the filtering can be done using the filter tool in the **Molecule Table**).
- **Flexible bonds**. This option reports the number of rotatable covalent bonds (single order, non-terminal bonds, which are not part of a ring system).

In the wizard step 3, choose to "Open" or "Save" the results and click on the button labeled **Finish**.

For each **Molecule Table** given as input, a corresponding **Molecule Table** containing calculated molecular properties is provided as output. The calculated molecular properties will appear as new column entries. The settings in the Side Panel can be used to toggle each property on or off (Show Columns palette). An example of calculated properties is shown in figure 9.67.

Note! The calculated molecular properties will overwrite existing information in the **Molecule Tables** if the column names are identical.

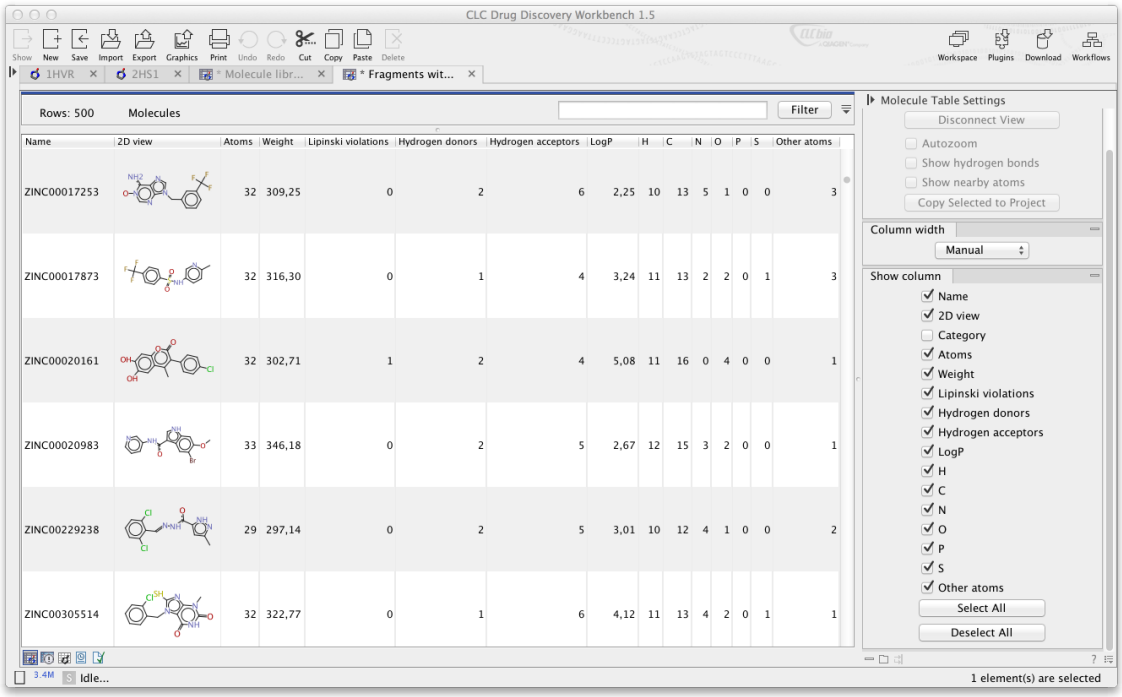


Figure 9.67: *Ligand molecules with calculated molecular properties.*

9.16.1 The log P algorithm

The calculation of the octanol-water partition coefficient (log P) is based on the XLOGP3-AA method [Cheng et al., 2007b].

XLOGP3-AA is an atom-additive method that calculates log P by adding up contributions from each atom in the given molecule. Each atom is categorized into one of the 83 basic atom types or

the four terminal groups suggested by Cheng et al. The terminal groups are included to take the effects of the common cyano-, diazo-, nitro-, and nitro oxide-groups into account. In addition, two correction factors are included to account for intermolecular interactions: 1) internal hydrogen bonding, which can give rise to an increase in the hydrophobicity of a molecule, and 2) molecules containing an amino acid moiety to compensate for overestimation of log P in these cases.

The contribution of each atom type, terminal group, and correction factor has been found by multivariate regression analysis of molecules with known experimental log P values (see [Cheng et al., 2007b] for details).

Note! The XLOGP3-AA atom typing and assignment of terminal groups / correction factors depends on the chemical preparation of the molecule. In particular, aromaticity plays a big role in the assignments, and incorrect preparation can lead to deviations in the log P estimates.

9.17 Extract ligands

With the "Extract Ligands" tool it is possible to extract all ligands from one or more **Molecule Projects**. To run the "Extract Ligands" tool:

Toolbox | **Drug Design** ( ) | **Extract Ligands** ( )

The Extract Ligands tool takes one or more **Molecule Projects** as input (figure 9.68). The output is a **Molecule Table** listing all molecules found in the Ligands category in each of the **Molecule Projects**. The ligand properties can then be easily compared in the table, and the table can be used as input for the Dock Ligands or Screen Ligands tools.

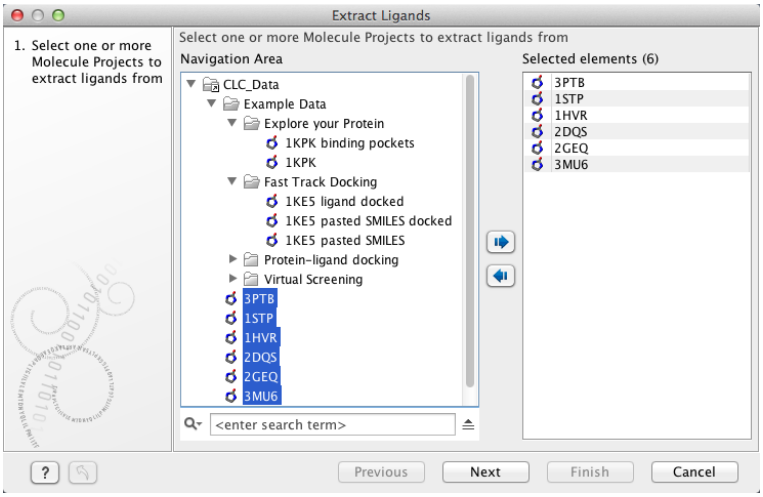


Figure 9.68: *Select Molecule Projects with ligands to extract.*

Chapter 10

Viewing and editing sequences

Contents

CLC Drug Discovery Workbench offers five different ways of viewing and editing single sequences as described in the first five sections of this chapter. Furthermore, this chapter also explains how to create a new sequence and how to gather several sequences in a sequence list.

10.1 View sequence

When you double-click a sequence in the **Navigation Area**, the sequence will open automatically, and you will see the nucleotides or amino acids. The zoom options described in section 2.2 allow you to e.g. zoom out in order to see more of the sequence in one view. There are a number of options for viewing and editing the sequence which are all described in this section.

All the options described in this section also apply to alignments (further described in section 14.2).

10.1.1 Sequence settings in Side Panel

Each view of a sequence has a **Side Panel** located at the right side of the view (see figure 10.1.

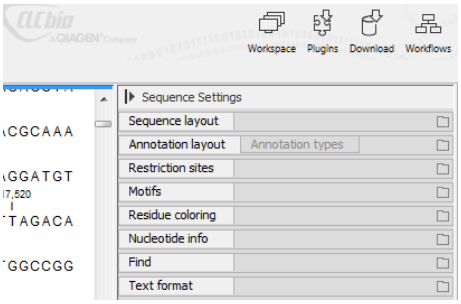


Figure 10.1: *Overview of the Side Panel which is always shown to the right of a view.*

When you make changes in the **Side Panel** the view of the sequence is instantly updated. To show or hide the **Side Panel**:

select the View | Ctrl + U

or **Click the ( ) at the top right corner of the Side Panel to hide | Click the ( ) to the right to show**

Below, each group of settings will be explained. Some of the preferences are not the same for nucleotide and protein sequences, but the differences will be explained for each group of settings.

Note! When you make changes to the settings in the **Side Panel**, they are not automatically saved when you save the sequence. Click **Save/restore Settings** ( ) to save the settings (see section 4.5 for more information).

Sequence Layout

These preferences determine the overall layout of the sequence:

- **Spacing.** Inserts a space at a specified interval:
  - **No spacing.** The sequence is shown with no spaces.
  - **Every 10 residues.** There is a space every 10 residues, starting from the beginning of the sequence.
  - **Every 3 residues, frame 1.** There is a space every 3 residues, corresponding to the reading frame starting at the first residue.
  - **Every 3 residues, frame 2.** There is a space every 3 residues, corresponding to the reading frame starting at the second residue.
  - **Every 3 residues, frame 3.** There is a space every 3 residues, corresponding to the reading frame starting at the third residue.
- **Wrap sequences.** Shows the sequence on more than one line.

  - **No wrap.** The sequence is displayed on one line.
  - **Auto wrap.** Wraps the sequence to fit the width of the view, not matter if it is zoomed in our out (displays minimum 10 nucleotides on each line).
  - **Fixed wrap.** Makes it possible to specify when the sequence should be wrapped. In the text field below, you can choose the number of residues to display on each line.
- **Double stranded.** Shows both strands of a sequence (only applies to DNA sequences).
- **Numbers on sequences.** Shows residue positions along the sequence. The starting point can be changed by setting the number in the field below. If you set it to e.g. 101, the first residue will have the position of -100. This can also be done by right-clicking an annotation and choosing **Set Numbers Relative to This Annotation**.
- **Numbers on plus strand.** Whether to set the numbers relative to the positive or the negative strand in a nucleotide sequence (only applies to DNA sequences).
- **Lock numbers.** When you scroll vertically, the position numbers remain visible. (Only possible when the sequence is not wrapped.)
- **Lock labels.** When you scroll horizontally, the label of the sequence remains visible.
- **Sequence label.** Defines the label to the left of the sequence.
  - Name (this is the default information to be shown).
  - Accession (sequences downloaded from databases like GenBank have an accession number).
  - Latin name.
  - Latin name (accession).
  - Common name.
  - Common name (accession).
- **Matching residues as dots** Residues in aligned sequences identical to residues in the first (reference) sequence will be presented as dots. An option that is only available for "Alignments" and "Read mappings".

Annotation Layout and Annotation Types See section 10.3.1.

Motifs

See section 13.7.1.

Residue coloring

These preferences make it possible to color both the residue letter and set a background color for the residue.

- **Non-standard residues.** For nucleotide sequences this will color the residues that are not C, G, A, T or U. For amino acids only B, Z, and X are colored as non-standard residues.

  - **Foreground color.** Sets the color of the letter. Click the color box to change the color.
  - **Background color.** Sets the background color of the residues. Click the color box to change the color.

- **Rasmol colors.** Colors the residues according to the Rasmol color scheme.
  See http://www.openrasmol.org/doc/rasmol.html
  - **Foreground color.** Sets the color of the letter. Click the color box to change the color.
  - **Background color.** Sets the background color of the residues. Click the color box to change the color.

- **Polarity colors (only protein).** Colors the residues according to the following categories:
  - **Green** neutral, polar
  - **Black** neutral, nonpolar
  - **Red** acidic, polar
  - **Blue** basic ,polar
  - As with other options, you can choose to set or change the coloring for either the residue letter or its background:
    * **Foreground color.** Sets the color of the letter. Click the color box to change the color.
    * **Background color.** Sets the background color of the residues. Click the color box to change the color.

- **Trace colors (only DNA).** Colors the residues according to the color conventions of chromatogram traces: A=green, C=blue, G=black, and T=red.
  - **Foreground color.** Sets the color of the letter.
  - **Background color.** Sets the background color of the residues.

Nucleotide info

These preferences only apply to nucleotide sequences.

- **Color space encoding.** Lets you define a few settings for how the colors should appear.

  Infer encoding This is used if you want to display the colors for non-color space sequence (e.g. a reference sequence). The colors are then simply inferred from the sequence.

  Show corrections This is only relevant for mapping results - it will show where the mapping process has detected color errors.

  Hide unaligned This option determines whether color for the unaligned ends of reads should be displayed. It also controls whether colors should be shown for gaps. The idea behind this is that these color dots will interfere with the color alignment, so it is possible to turn them off.

- **Translation.** Displays a translation into protein just below the nucleotide sequence. Depending on the zoom level, the amino acids are displayed with three letters or one letter. In cases where variants are present in the reads, synonymous variants are shown in orange in the translated sequence whereas non-synonymous are shown in red.

  - **Frame.** Determines where to start the translation.
    * **ORF/CDS**. If the sequence is annotated, the translation will follow the CDS or ORF annotations. If annotations overlap, only one translation will be shown. If only one annotation is visible, the Workbench will attempt to use this annotation to mark the start and stop for the translation. In cases where this is not possible, the first annotation will be used (i.e. the one closest to the 5' end of the sequence).
    * **Selection.** This option will only take effect when you make a selection on the sequence. The translation will start from the first nucleotide selected. Making a new selection will automatically display the corresponding translation. Read more about selecting in section 10.1.2.
    * **+1 to -1.** Select one of the six reading frames.
    * **All forward/All reverse.** Shows either all forward or all reverse reading frames.
    * **All.** Select all reading frames at once. The translations will be displayed on top of each other.
  - **Table.** The translation table to use in the translation.
  - **Only AUG start codons.** For most genetic codes, a number of codons can be start codons (TTG, CTG, or ATG). These will be colored green, unless selecting the "Only AUG start codons" option, which will result in only the AUG codons colored in green.
  - **Single letter codes.** Choose to represent the amino acids with a single letter instead of three letters.

- **G/C content.** Calculates the G/C content of a part of the sequence and shows it as a gradient of colors or as a graph below the sequence.
  - **Window length.** Determines the length of the part of the sequence to calculate. A window length of 9 will calculate the G/C content for the nucleotide in question plus the 4 nucleotides to the left and the 4 nucleotides to the right. A narrow window will focus on small fluctuations in the G/C content level, whereas a wider window will show fluctuations between larger parts of the sequence.
  - **Foreground color.** Colors the letter using a gradient, where the left side color is used for low levels of G/C content and the right side color is used for high levels of G/C content. The sliders just above the gradient color box can be dragged to highlight relevant levels of G/C content. The colors can be changed by clicking the box. This will show a list of gradients to choose from.
  - **Background color.** Sets a background color of the residues using a gradient in the same way as described above.
  - **Graph.** The G/C content level is displayed on a graph (Learn how to export the data behind the graph in section 6.5).
    * **Height.** Specifies the height of the graph.
    * **Type.** The graph can be displayed as Line plot, Bar plot or as a Color bar.
    * **Color box.** For Line and Bar plots, the color of the plot can be set by clicking the color box. For Colors, the color box is replaced by a gradient color box as described under Foreground color.

- **Secondary structure.** Allows you to choose how to display a symbolic representation of the secondary structure along the sequence. See section **??** for a detailed description of the settings.

Protein info

These preferences only apply to proteins. The first nine items are different hydrophobicity scales. These are described in section 13.11.2.

- **Kyte-Doolittle.** The Kyte-Doolittle scale is widely used for detecting hydrophobic regions in proteins. Regions with a positive value are hydrophobic. This scale can be used for identifying both surface-exposed regions as well as transmembrane regions, depending on the window size used. Short window sizes of 5-7 generally work well for predicting putative surface-exposed regions. Large window sizes of 19-21 are well suited for finding transmembrane domains if the values calculated are above 1.6 [Kyte and Doolittle, 1982]. These values should be used as a rule of thumb and deviations from the rule may occur.
- **Cornette.** Cornette *et al.* computed an optimal hydrophobicity scale based on 28 published scales [Cornette et al., 1987]. This optimized scale is also suitable for prediction of alpha-helices in proteins.
- **Engelman.** The Engelman hydrophobicity scale, also known as the GES-scale, is another scale which can be used for prediction of protein hydrophobicity [Engelman et al., 1986]. As the Kyte-Doolittle scale, this scale is useful for predicting transmembrane regions in proteins.
- **Eisenberg.** The Eisenberg scale is a normalized consensus hydrophobicity scale which shares many features with the other hydrophobicity scales [Eisenberg et al., 1984].
- **Rose.** The hydrophobicity scale by Rose *et al.* is correlated to the average area of buried amino acids in globular proteins [Rose et al., 1985]. This results in a scale which is not showing the helices of a protein, but rather the surface accessibility.
- **Janin.** This scale also provides information about the accessible and buried amino acid residues of globular proteins [Janin, 1979].
- **Hopp-Woods.** Hopp and Woods developed their hydrophobicity scale for identification of potentially antigenic sites in proteins. This scale is basically a hydrophilic index where apolar residues have been assigned negative values. Antigenic sites are likely to be predicted when using a window size of 7 [Hopp and Woods, 1983].
- **Welling.** [Welling et al., 1985] Welling et al. used information on the relative occurrence of amino acids in antigenic regions to make a scale which is useful for prediction of antigenic regions. This method is better than the Hopp-Woods scale of hydrophobicity which is also used to identify antigenic regions.
- **Kolaskar-Tongaonkar.** A semi-empirical method for prediction of antigenic regions has been developed [Kolaskar and Tongaonkar, 1990]. This method also includes information of surface accessibility and flexibility and at the time of publication the method was able to predict antigenic determinants with an accuracy of 75%.
- **Surface Probability.** Display of surface probability based on the algorithm by [Emini et al., 1985]. This algorithm has been used to identify antigenic determinants on the surface of proteins.
- **Chain Flexibility.** Display of backbone chain flexibility based on the algorithm by [Karplus and Schulz, 1985]. It is known that chain flexibility is an indication of a putative antigenic determinant.

Find

The Find function can be used for searching the sequence and is invoked by pressing Ctrl + Shift + F (⌘ + Shift + F on Mac). Initially, specify the 'search term' to be found, select the type of search (see various options in the following) and finally click on the Find button. The first occurrence of the search term will then be highlighted. Clicking the find button again will find the next occurrence and so on. If the search string is found, the corresponding part of the sequence will be selected.

- **Search term.** Enter the text or number to search for. The search function does not discriminate between lower and upper case characters.
- **Sequence search.** Search the nucleotides or amino acids. For amino acids, the single letter abbreviations should be used for searching. The sequence search also has a set of advanced search parameters:
  - Include negative strand. This will search on the negative strand as well.
  - Treat ambiguous characters as wildcards in search term. If you search for e.g. ATN, you will find both ATG and ATC. If you wish to find literally exact matches for ATN (i.e. only find ATN - not ATG), this option should not be selected.
  - Treat ambiguous characters as wildcards in sequence. If you search for e.g. ATG, you will find both ATG and ATN. If you have large regions of Ns, this option should not be selected.

  Note that if you enter a position instead of a sequence, it will automatically switch to position search.
- **Annotation search.** Searches the annotations on the sequence. The search is performed both on the labels of the annotations, but also on the text appearing in the tooltip that you see when you keep the mouse cursor fixed. If the search term is found, the part of the sequence corresponding to the matching annotation is selected. Below this option you can choose to search for translations as well. Sequences annotated with coding regions often have the translation specified which can lead to undesired results.
- **Position search.** Finds a specific position on the sequence. In order to find an interval, e.g. from position 500 to 570, enter "500..570" in the search field. This will make a selection from position 500 to 570 (both included). Notice the two periods (..) between the start an end number. If you enter positions including thousands separators like 123,345, the comma will just be ignored and it would be equivalent to entering 123345.
- **Include negative strand.** When searching the sequence for nucleotides or amino acids, you can search on both strands.
- **Name search.** Searches for sequence names. This is useful for searching sequence lists and mapping results for example.

This concludes the description of the **View Preferences**. Next, the options for selecting and editing sequences are described.

Text format

These preferences allow you to adjust the format of all the text in the view (both residue letters, sequence name and translations if they are shown).

- **Text size.** Five different sizes.
- **Font.** Shows a list of Fonts available on your computer.
- **Bold residues.** Makes the residues bold.

10.1.2 Selecting parts of the sequence

You can select parts of a sequence:

> **Click Selection ( ) in Toolbar | Press and hold down the mouse button on the sequence where you want the selection to start | move the mouse to the end of the selection while holding the button | release the mouse button**

Alternatively, you can search for a specific interval using the find function described above.

If you have made a selection and wish to adjust it:

> **drag the edge of the selection (you can see the mouse cursor change to a horizontal arrow**
>
> or **press and hold the Shift key while using the right and left arrow keys to adjust the right side of the selection.**

If you wish to select the entire sequence:

> **double-click the sequence name to the left**

Selecting several parts at the same time (multiselect)

You can select several parts of sequence by holding down the **Ctrl** button while making selections. Holding down the **Shift** button lets you extend or reduce an existing selection to the position you clicked.

To select a part of a sequence covered by an annotation:

> **right-click the annotation | Select annotation**
>
> or **double-click the annotation**

Open a selection in a new view

A selection can be opened in a new view and saved as a new sequence:

> **right-click the selection | Open selection in New View ( )**

This opens the annotated part of the sequence in a new view. The new sequence can be saved by dragging the tab of the sequence view into the **Navigation Area**.

A selection can also be copied to the clipboard and pasted into another program:

> **make a selection | Ctrl + C (⌘ + C on Mac)**

Note! The annotations covering the selection will not be copied.

A selection of a sequence can be edited as described in the following section.

10.1.3 Editing the sequence

When you make a selection, it can be edited by:

right-click the selection | Edit Selection ( )

A dialog appears displaying the sequence. You can add, remove or change the text and click **OK**. The original selected part of the sequence is now replaced by the sequence entered in the dialog. This dialog also allows you to paste text into the sequence using Ctrl + V (⌘ + V on Mac).

If you delete the text in the dialog and press **OK**, the selected text on the sequence will also be deleted. Another way to delete a part of the sequence is to:

right-click the selection | Delete Selection ( )

If you wish to correct only one residue, this is possible by simply making the selection cover only one residue and then type the new residue.

Note When editing annotated nucleotide sequences, the annotation content is not updated automatically (but its position is). Please refer to section 10.3.3 for details on annotation editing.

10.1.4 Sequence region types

The various annotations on sequences cover parts of the sequence. Some cover an interval, some cover intervals with unknown endpoints, some cover more than one interval etc. In the following, all of these will be referred to as *regions*. Regions are generally illustrated by markings (often arrows) on the sequences. An arrow pointing to the right indicates that the corresponding region is located on the positive strand of the sequence. Figure 10.2 is an example of three regions with separate colors.

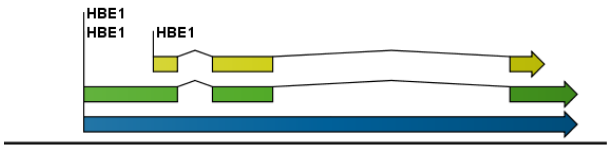


Figure 10.2: *Three regions on a human beta globin DNA sequence (HUMHBB).*

Figure 10.3 shows an artificial sequence with all the different kinds of regions.

10.2 Circular DNA

A sequence can be shown as a circular molecule:

Select a sequence in the Navigation Area and right-click on the file name | Hold the mouse over "Show" to enable a list of options | Select "Circular View" ( )

or **If the sequence is already open | Click "Show Circular View" ( ) at the lower left part of the view**

This will open a view of the molecule similar to the one in figure 10.4.

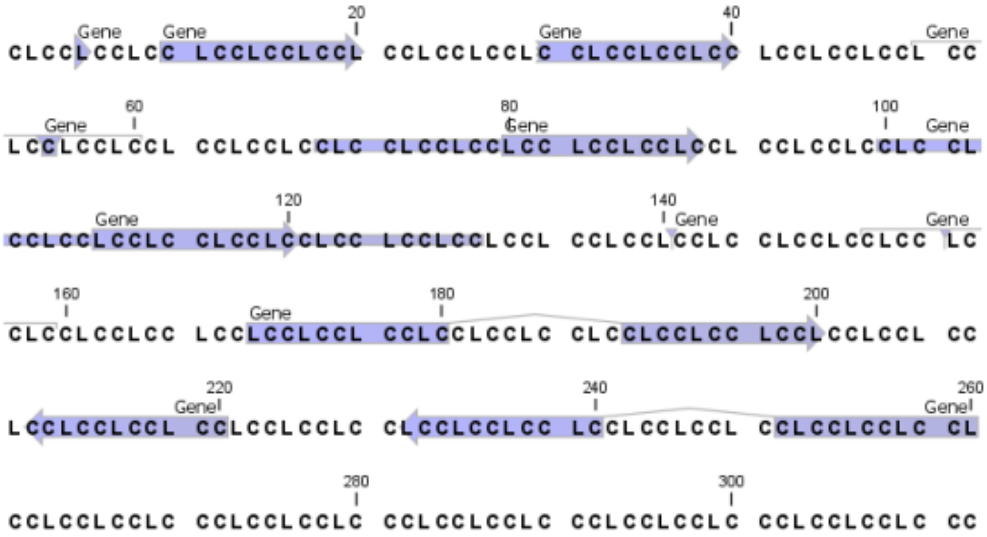


Figure 10.3: *Region #1: A single residue, Region #2: A range of residues including both endpoints, Region #3: A range of residues starting somewhere before 30 and continuing up to and including 40, Region #4: A single residue somewhere between 50 and 60 inclusive, Region #5: A range of residues beginning somewhere between 70 and 80 inclusive and ending at 90 inclusive, Region #6: A range of residues beginning somewhere between 100 and 110 inclusive and ending somewhere between 120 and 130 inclusive, Region #7: A site between residues 140 and 141, Region #8: A site between two residues somewhere between 150 and 160 inclusive, Region #9: A region that covers ranges from 170 to 180 inclusive and 190 to 200 inclusive, Region #10: A region on negative strand that covers ranges from 210 to 220 inclusive, Region #11: A region on negative strand that covers ranges from 230 to 240 inclusive and 250 to 260 inclusive.*

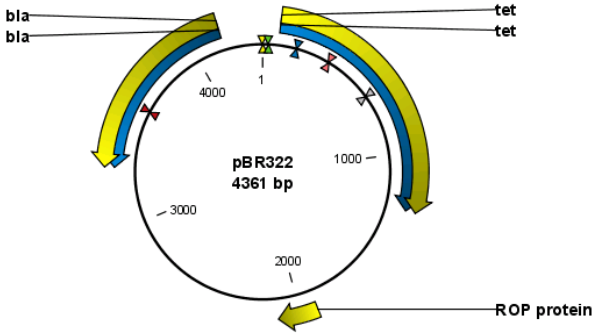


Figure 10.4: *A molecule shown in a circular view.*

This view of the sequence shares some of the properties of the linear view of sequences as described in section 10.1, but there are some differences. The similarities and differences are listed below:

- **Similarities**:
  - The editing options.
  - Options for adding, editing and removing annotations.
  - **Restriction Sites**, **Annotation Types**, **Find** and **Text Format** preferences groups.
- **Differences**:

- In the **Sequence Layout** preferences, only the following options are available in the circular view: **Numbers on plus strand**, **Numbers on sequence** and **Sequence label**.
- You cannot zoom in to see the residues in the circular molecule. If you wish to see these details, split the view with a linear view of the sequence
- In the **Annotation Layout**, you also have the option of showing the labels as **Stacked**. This means that there are no overlapping labels and that all labels of both annotations and restriction sites are adjusted along the left and right edges of the view.

10.2.1 Using split views to see details of the circular molecule

In order to see the nucleotides of a circular molecule you can open a new view displaying a circular view of the molecule:

Press and hold the Ctrl button (⌘ on Mac) | click Show Sequence (ACT) at the bottom of the view

This will open a linear view of the sequence below the circular view. When you zoom in on the linear view you can see the residues as shown in figure 10.5.

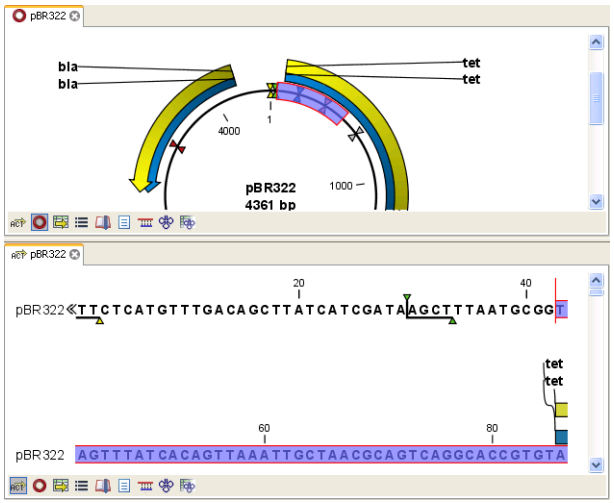


Figure 10.5: *Two views showing the same sequence. The bottom view is zoomed in.*

Note! If you make a selection in one of the views, the other view will also make the corresponding selection, providing an easy way for you to focus on the same region in both views.

10.2.2 Mark molecule as circular and specify starting point

You can mark a DNA molecule as circular by right-clicking its name in either the sequence view or the circular view. In the right-click menu you can also make a circular molecule linear. A circular molecule displayed in the normal sequence view, will have the sequence ends marked with a ï¿½.

The starting point of a circular sequence can be changed by:

make a selection starting at the position that you want to be the new starting point | right-click the selection | Move Starting Point to Selection Start

Note! This can only be done for sequence that have been marked as circular.

10.3 Working with annotations

Annotations provide information about specific regions of a sequence.

Annotations derive from different sources:

- Sequences downloaded from databases like UniProt are annotated.
- In some of the data formats that can be imported into *CLC Drug Discovery Workbench*, sequences can have annotations (e.g. the Swiss-Prot format).
- The result of a number of analyses in *CLC Drug Discovery Workbench* are annotations on the sequence (e.g. Pfam Domain Search and Motif Search).
- A protein structure can be linked with a sequence (section 9.4.2), and atom groups defined on the structure transferred to sequence annotations or vica versa (section 9.4.3).
- You can manually add annotations to a sequence (described in the section 10.3.2).

If you would like to extract parts of a sequence (or several sequences) based on its annotations, you can find a description of how to do this in section **??**.

Note! Annotations are included if you export the sequence in CLC format. When exporting in other formats, annotations are not preserved in the exported file.

10.3.1 Viewing annotations

Annotations can be viewed in a number of different ways:

- As arrows or boxes in all views displaying sequences (sequence lists, alignments etc)
- In the table of annotations ().
- In the text view of sequences ()

In the following sections, these view options will be described in more detail. In all the views except the text view (), annotations can be added, modified and deleted. This is described in the following sections.

View Annotations in sequence views

Figure 10.6 shows an annotation displayed on a sequence.

The various sequence views listed in section 10.3.1 have different default settings for showing annotations. However, they all have two groups in the **Side Panel** in common:

- **Annotation Layout**
- **Annotation Types**

The two groups are shown in figure 10.7.

In the **Annotation layout** group, you can specify how the annotations should be displayed (notice that there are some minor differences between the different sequence views):

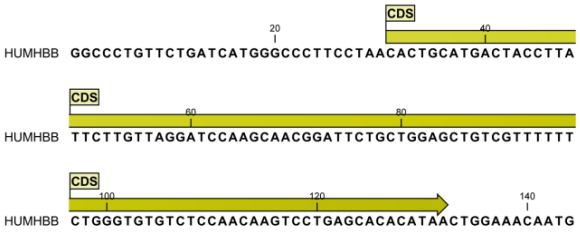


Figure 10.6: *An annotation showing a coding region on a genomic dna sequence.*

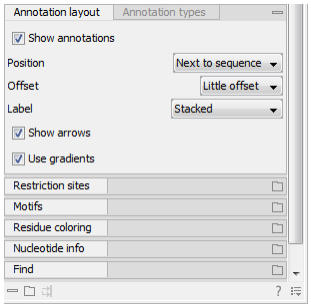


Figure 10.7: *The annotation layout in the Side Panel. The annotation types can be shown by clicking on the "Annotation types" tab.*

- **Show annotations.** Determines whether the annotations are shown.
- **Position.**
  - **On sequence.** The annotations are placed on the sequence. The residues are visible through the annotations (if you have zoomed in to 100%).
  - **Next to sequence.** The annotations are placed above the sequence.
  - **Separate layer.** The annotations are placed above the sequence and above restriction sites (only applicable for nucleotide sequences).
- **Offset.** If several annotations cover the same part of a sequence, they can be spread out.
  - **Piled.** The annotations are piled on top of each other. Only the one at front is visible.
  - **Little offset.** The annotations are piled on top of each other, but they have been offset a little.
  - **More offset.** Same as above, but with more spreading.
  - **Most offset.** The annotations are placed above each other with a little space between. This can take up a lot of space on the screen.
- **Label.** The name of the annotation can shown as a label. Additional information about the sequence is shown if you place the mouse cursor on the annotation and keep it still.
  - **No labels.** No labels are displayed.
  - **On annotation.** The labels are displayed in the annotation's box.
  - **Over annotation.** The labels are displayed above the annotations.
  - **Before annotation.** The labels are placed just to the left of the annotation.
  - **Flag.** The labels are displayed as flags at the beginning of the annotation.

  - **Stacked.** The labels are offset so that the text of all labels is visible. This means that there is varying distance between each sequence line to make room for the labels.

- **Show arrows.** Displays the end of the annotation as an arrow. This can be useful to see the orientation of the annotation (for DNA sequences). Annotations on the negative strand will have an arrow pointing to the left.
- **Use gradients.** Fills the boxes with gradient color.

In the **Annotation types** group, you can choose which kinds of annotations that should be displayed. This group lists all the types of annotations that are attached to the sequence(s) in the view. For sequences with many annotations, it can be easier to get an overview if you deselect the annotation types that are not relevant.

Unchecking the checkboxes in the **Annotation layout** will not remove this type of annotations them from the sequence - it will just hide them from the view.

Besides selecting which types of annotations that should be displayed, the **Annotation types** group is also used to change the color of the annotations on the sequence. Click the colored square next to the relevant annotation type to change the color.

This will display a dialog with five tabs: Swatches, HSB, HSI, RGB, and CMYK. They represent five different ways of specifying colors. Apply your settings and click **OK**. When you click **OK**, the color settings cannot be reset. The **Reset** function only works for changes made before pressing **OK**.

Furthermore, the **Annotation types** can be used to easily browse the annotations by clicking the small button ( ) next to the type. This will display a list of the annotations of that type (see figure 10.8).

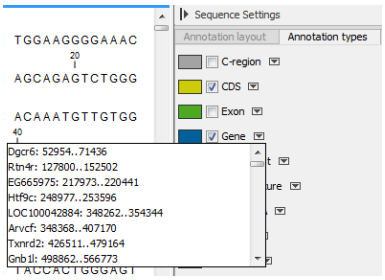


Figure 10.8: *Browsing the gene annotations on a sequence.*

Clicking an annotation in the list will select this region on the sequence. In this way, you can quickly find a specific annotation on a long sequence.

Note: A waved end on an annotation (figure 10.9) means that the annotation is torn, i.e., it extends beyond the sequence displayed. An annotation can be torn when a new, smaller sequence has been created from a larger sequence. A common example of this situation is when you select a section of a stand alone sequence and open it in a new view. If there are annotations present within this selected region that extend beyond the selection, then the selected sequence shown in the new view will exhibit these torn annotations.

View Annotations in a table

Annotations can also be viewed in a table:

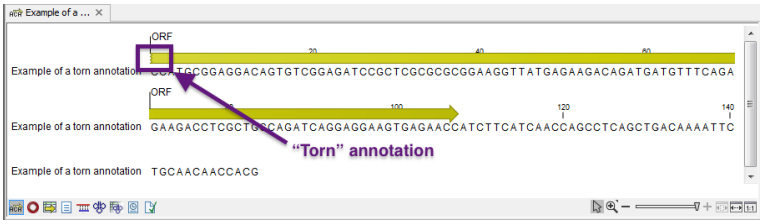


Figure 10.9: *Example of a torn annotation on a sequence.*

Select a sequence in the Navigation Area and right-click on the file name | Hold the mouse over "Show" to enable a list of options | Annotation Table ( )

or **If the sequence is already open | Click Show Annotation Table ( ) at the lower left part of the view**

This will open a view similar to the one in figure 10.10).

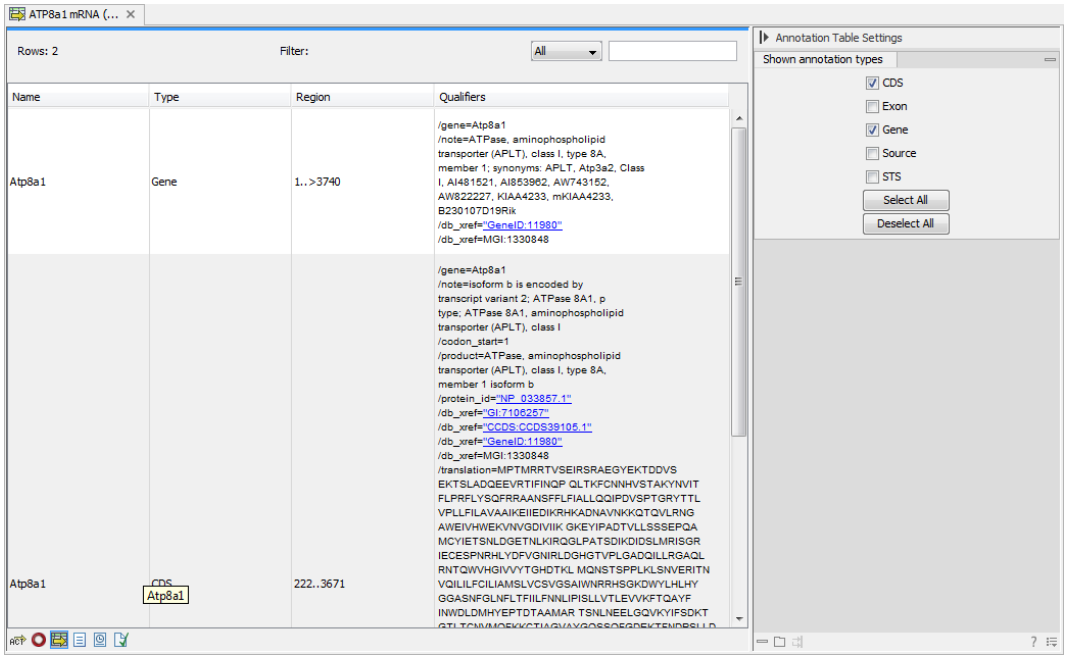


Figure 10.10: *A table showing annotations on the sequence.*

In the **Side Panel** you can show or hide individual annotation types in the table. E.g. if you only wish to see "gene" annotations, de-select the other annotation types so that only "gene" is selected.

Each row in the table is an annotation which is represented with the following information:

- **Name.**
- **Type.**
- **Region.**
- **Qualifiers.**

The Name, Type and Region for each annotation can be edited simply by double-clicking, typing the change directly, and pressing **Enter**.

This information corresponds to the information in the dialog when you edit and add annotations (see section 10.3.2).

You can benefit from this table in several ways:

- It provides an intelligible overview of all the annotations on the sequence.
- You can use the filter at the top to search the annotations. Type e.g. "UCP" into the filter and you will find all annotations which have "UCP" in either the name, the type, the region or the qualifiers. Combined with showing or hiding the annotation types in the **Side Panel**, this makes it easy to find annotations or a subset of annotations.
- You can copy and paste annotations, e.g. from one sequence to another.
- If you wish to edit many annotations consecutively, the double-click editing makes this very fast (see section 10.3.2).

10.3.2 Adding annotations

Adding annotations to a sequence can be done in two ways:

Open the sequence in a sequence view (double-click in the Navigation Area) | make a selection covering the part of the sequence you want to annotate[1] **| right-click the selection | Add Annotation ()**

or **Select a sequence in the Navigation Area and right-click on the file name | Hold the mouse over "Show" to enable a list of options | Annotation table () | right click anywhere in the annotation table | select Add Annotation ()**

This will display a dialog like the one in figure 10.11.

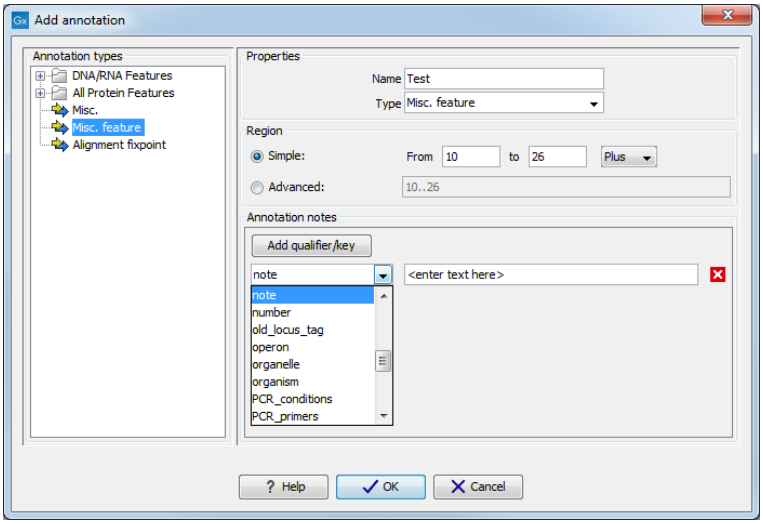


Figure 10.11: *The Add Annotation dialog.*

The left-hand part of the dialog lists a number of **Annotation types**. When you have selected an annotation type, it appears in **Type** to the right. You can also select an annotation directly in this list. Choosing an annotation type is mandatory. If you wish to use an annotation type which is

not present in the list, simply enter this type into the **Type** field [2].

The right-hand part of the dialog contains the following text fields:

- **Name.** The name of the annotation which can be shown on the label in the sequence views. (Whether the name is actually shown depends on the **Annotation Layout** preferences, see section 10.3.1).
- **Type.** Reflects the left-hand part of the dialog as described above. You can also choose directly in this list or type your own annotation type.
- **Region.** If you have already made a selection, this field will show the positions of the selection. You can modify the region further using the conventions of DDBJ, EMBL and GenBank. The following are examples of how to use the syntax (based on http://www.ncbi.nlm.nih.gov/collab/FT/):
  - **467**. Points to a single residue in the presented sequence.
  - **340..565**. Points to a continuous range of residues bounded by and including the starting and ending residues.
  - **<345..500**. Indicates that the exact lower boundary point of a region is unknown. The location begins at some residue previous to the first residue specified (which is not necessarily contained in the presented sequence) and continues up to and including the ending residue.
  - **<1..888**. The region starts before the first sequenced residue and continues up to and including residue 888.
  - **1..>888**. The region starts at the first sequenced residue and continues beyond residue 888.
  - **(102.110)**. Indicates that the exact location is unknown, but that it is one of the residues between residues 102 and 110, inclusive.
  - **123^124**. Points to a site between residues 123 and 124.
  - **join(12..78,134..202)**. Regions 12 to 78 and 134 to 202 should be joined to form one contiguous sequence.
  - **complement(34..126)** Start at the residue complementary to 126 and finish at the residue complementary to residue 34 (the region is on the strand complementary to the presented strand).
  - **complement(join(2691..4571,4918..5163))**. Joins regions 2691 to 4571 and 4918 to 5163, then complements the joined segments (the region is on the strand complementary to the presented strand).
  - **join(complement(4918..5163),complement(2691..4571))**. Complements regions 4918 to 5163 and 2691 to 4571, then joins the complemented segments (the region is on the strand complementary to the presented strand).
- **Annotations.** In this field, you can add more information about the annotation like comments and links. Click the **Add qualifier/key** button to enter information. Select a qualifier which describes the kind of information you wish to add. If an appropriate qualifier is not present

[2] Note that your own annotation types will be converted to "unsure" when exporting in GenBank format. As long as you use the sequence in CLC format, you own annotation type will be preserved

in the list, you can type your own qualifier. The pre-defined qualifiers are derived from the GenBank format. You can add as many qualifier/key lines as you wish by clicking the button. Redundant lines can be removed by clicking the delete icon ( ). The information entered on these lines is shown in the annotation table (see section 10.3.1) and in the yellow box which appears when you place the mouse cursor on the annotation. If you write a hyperlink in the **Key** text field, like e.g. "www.clcbio.com", it will be recognized as a hyperlink. Clicking the link in the annotation table will open a web browser.

Click **OK** to add the annotation.

Note! The annotation will be included if you export the sequence in GenBank, Swiss-Prot or CLC format. When exporting in other formats, annotations are not preserved in the exported file.

10.3.3 Edit annotations

To edit an existing annotation from within a sequence view:

right-click the annotation | **Edit Annotation ( )**

This will show the same dialog as in figure 10.11, with the exception that some of the fields are filled out depending on how much information the annotation contains.

There is another way of quickly editing annotations which is particularly useful when you wish to edit several annotations.

To edit the information, simply double-click and you will be able to edit e.g. the name or the annotation type. If you wish to edit the qualifiers and double-click in this column, you will see the dialog for editing annotations.

Advanced editing of annotations

Sometimes you end up with annotations which do not have a meaningful name. In that case there is an advanced batch rename functionality:

Open the Annotation Table ( ) | **select the annotations that you want to rename** | **right-click the selection** | **Advanced Rename**

This will bring up the dialog shown in figure 10.12.

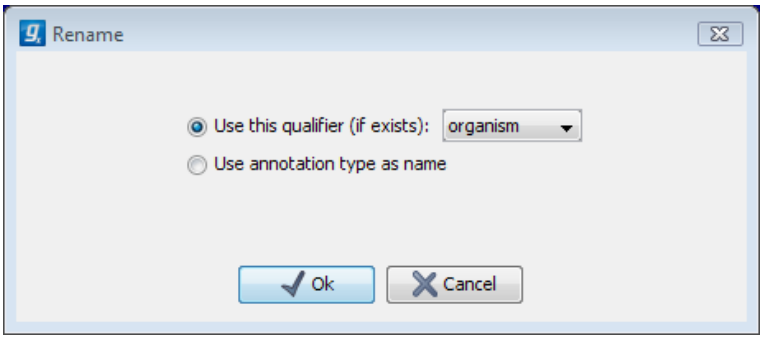


Figure 10.12: *The Advanced Rename dialog.*

In this dialog, you have two options:

- **Use this qualifier.** Use one of the qualifiers as name. A list of all qualifiers of all the selected annotations is shown. Note that if one of the annotations do not have the qualifier you have chosen, it will not be renamed. If an annotation has multiple qualifiers of the same type, the first is used for naming.
- **Use annotation type as name.** The annotation's type will be used as name (e.g. if you have an annotation of type "Promoter", it will get "Promoter" as its name by using this option).

A similar functionality for batch re-typing annotations is available in the right-click menu as well, in case your annotations are not typed correctly:

> **Open the Annotation Table ( ) | select the annotations that you want to retype | right-click the selection | Advanced Retype**

This will bring up the dialog shown in figure 10.13.

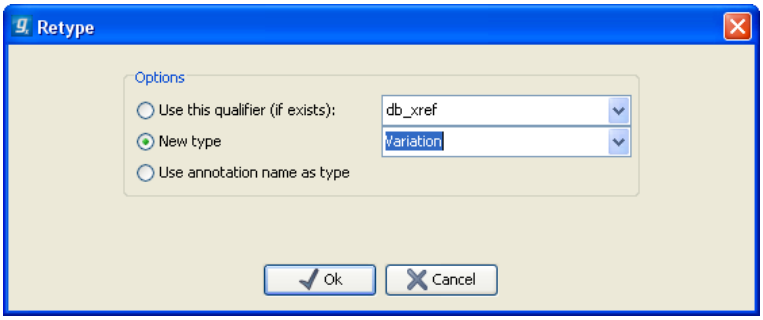


Figure 10.13: *The Advanced Retype dialog.*

In this dialog, you have two options:

- **Use this qualifier.** Use one of the qualifiers as type. A list of all qualifiers of all the selected annotations is shown. Note that if one of the annotations do not have the qualifier you have chosen, it will not be retyped. If an annotation has multiple qualifiers of the same type, the first is used for the new type.
- **New type**. You can select from a list of all the pre-defined types as well as enter your own annotation type. All the selected annotations will then get this type.
- **Use annotation name as type.** The annotation's name will be used as type (e.g. if you have an annotation named "Promoter", it will get "Promoter" as its type by using this option).

10.3.4 Removing annotations

Annotations can be hidden using the **Annotation Types** preferences in the **Side Panel** to the right of the view (see section 10.3.1). In order to completely remove the annotation:

> **right-click the annotation | Delete Annotation ( )**

If you want to remove all annotations of one type:

> **right-click an annotation of the type you want to remove | Delete | Delete Annotations of Type "type"**

If you want to remove all annotations from a sequence:

right-click an annotation | Delete | Delete All Annotations

The removal of annotations can be undone using Ctrl + Z or Undo ( ) in the Toolbar.

If you have more sequences (e.g. in a sequence list, alignment or contig), you have two additional options:

right-click an annotation | Delete | Delete All Annotations from All Sequences

right-click an annotation | Delete | Delete Annotations of Type "type" from All Sequences

10.4 Element information

The normal view of a sequence (by double-clicking) shows the annotations as boxes along the sequence, but often there is more information available about sequences. This information is available through the **Element info** view.

To view the sequence information:

Select a sequence in the Navigation Area and right-click on the file name | Hold the mouse over "Show" to enable a list of options | Element Info ( )

Another way to show the text view is to open the sequence in the **View Area** and click on the "Show Element Info" icon ( ) found at the bottom of the window.

This will display a view similar to fig 10.14.

Figure 10.14: *The initial display of sequence info for the HUMHBB DNA sequence from the Example data.*

All the lines in the view are headings, and the corresponding text can be shown by clicking the text.

- **Name.** The name of the sequence which is also shown in sequence views and in the **Navigation Area**.
- **Description.** A description of the sequence.
- **Comments.** The author's comments about the sequence.
- **Keywords.** Keywords describing the sequence.

- **Db source.** Accession numbers in other databases concerning the same sequence.
- **Gb Division.** Abbreviation of GenBank divisions. See section 3.3 in the GenBank release notes for a full list of GenBank divisions.
- **Length.** The length of the sequence.
- **Modification date.** Modification date from the database. This means that this date does not reflect your own changes to the sequence. See the history (section 2.1.2) for information about the latest changes to the sequence after it was downloaded from the database.
- **Latin name.** Latin name of the organism.
- **Common name.** Scientific name of the organism.
- **Taxonomy name.** Taxonomic classification levels.

The information available depends on the origin of the sequence. Sequences downloaded from database like NCBI and UniProt (see section 11) have this information. On the other hand, some sequence formats like fasta format do not contain this information.

Some of the information can be edited by clicking the blue **Edit** text. This means that you can add your own information to sequences that do not derive from databases.

Note that for other kinds of data, the **Element info** will only have **Name** and **Description**.

10.5 View as text

A sequence can be viewed as text without any layout and text formatting. This displays all the information about the sequence in the GenBank file format. To view a sequence as text:

> **Select a sequence in the Navigation Area and right-click on the file name | Hold the mouse over "Show" to enable a list of options | Select "Text View" ()**

Another way to show the text view is to open the sequence in the **View Area** and click on the "Show Text View" icon () found at the bottom of the window.

This makes it possible to see background information about e.g. the authors and the origin of DNA and protein sequences. Selections or the entire text of the **Sequence Text View** can be copied and pasted into other programs:

Much of the information is also displayed in the **Sequence info**, where it is easier to get an overview (see section 10.4.)

In the **Side Panel**, you find a search field for searching the text in the view.

10.6 Sequence Lists

The **Sequence List** shows a number of sequences in a tabular format or it can show the sequences together in a normal sequence view.

Having sequences in a sequence list can help organizing sequence data.

Sequence lists are generated automatically when you import files containing more than one sequence. Sequence lists may also be created as the output from particular Workbench tool including database searches.

Sequence List can also be created from single sequences or by merging already existing sequence lists with the Workbench. To do this:

> **select two or more sequences or sequence lists** | **right-click the elements** | **New** | **Sequence List ( )**

Alternatively, you can launch this took via the menu system:

> **File** | **New** | **Sequence List ( )**

This opens the **Sequence List** Wizard:

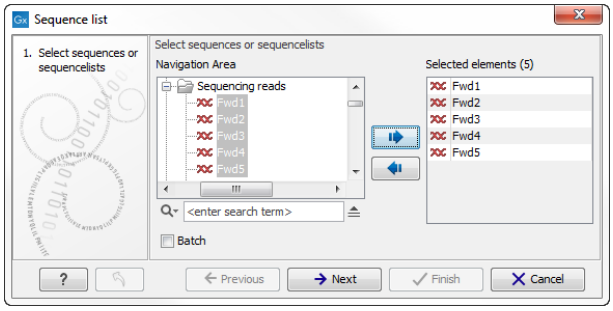


Figure 10.15: *A Sequence List dialog.*

The dialog allows you to select more sequences to include in the list, or to remove already chosen sequences from the list.

Clicking **Finish** opens the sequence list. It can be saved by clicking **Save** ( ) or by dragging the tab of the view into the **Navigation Area**.

Opening a Sequence list is done by:

> **right-click the sequence list in the Navigation Area** | **Show ( )** | **Graphical Sequence List ( ) OR Table ( )**

The two different views of the same sequence list are shown in split screen in figure 10.16.

10.6.1 Graphical view of sequence lists

The graphical view of sequence lists is almost identical to the view of single sequences (see section 10.1). The main difference is that you now can see more than one sequence in the same view.

However, you also have a few extra options for sorting, deleting and adding sequences:

- To add extra sequences to the list, right-click an empty (white) space in the view, and select **Add Sequences**.
- To delete a sequence from the list, right-click the sequence's name and select **Delete Sequence**.

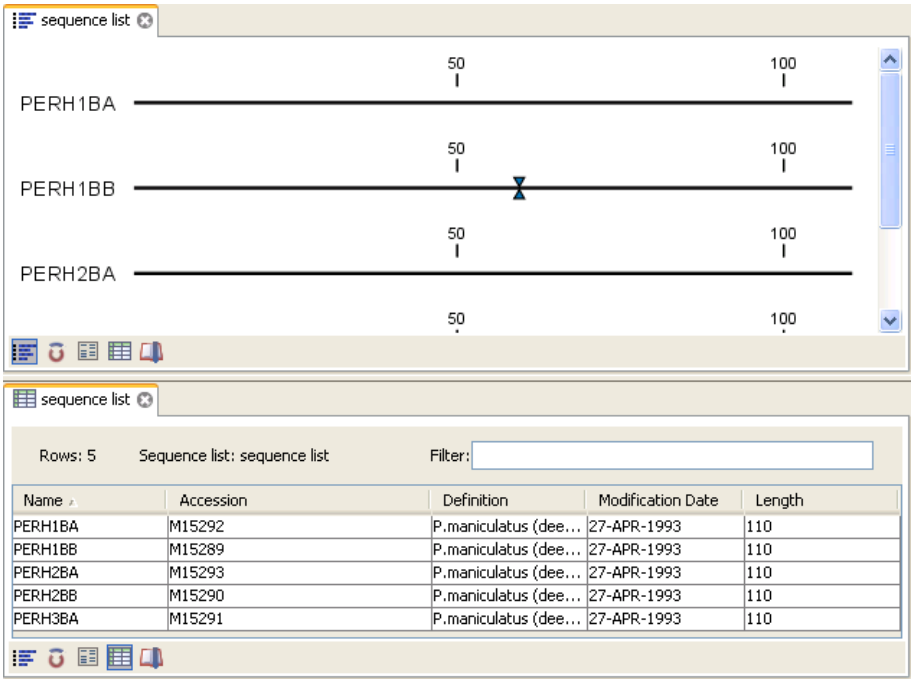


Figure 10.16: *A sequence list containing multiple sequences can be viewed in either a table or in a graphical sequence list. The graphical view is useful for viewing annotations and the sequence itself, while the table view provides other information like sequence lengths, and the number of sequences in the list (number of Rows reported).*

- To sort the sequences in the list, right-click the name of one of the sequences and select **Sort Sequence List by Name** or **Sort Sequence List by Length**.
- To rename a sequence, right-click the name of the sequence and select **Rename Sequence**.

10.6.2 Sequence list table

Each sequence in the table sequence list is displayed with:

- Name.
- Accession.
- Description.
- Modification date.
- Length.
- First 50 residues.

The number of sequences in the list is reported as the number of Rows at the top of the table view.

Adding and removing sequences from the list is easy: adding is done by dragging the sequence from another list or from the **Navigation Area** and drop it in the table. To delete sequences, simply select them and press **Delete** ( ).

You can also create a subset of the sequence list:

select the relevant sequences | **right-click** | **Create New Sequence List**

This will create a new sequence list, which only includes the selected sequences.

Learn more about tables in Appendix 3.3.

10.6.3 Extract sequences from sequence list

Sequences can be extracted from a sequence list when the sequence list is opened in tabular view. One or more sequences can be dragged (with the mouse) directly from the table into the **Navigation Area**. This allows you to extract specific sequences from the entire list. Another option is to extract all sequences found in the list. This can be done with the **Extract Sequences** tool:

Toolbox | **Sequence Analysis (**  **)**| **Extract Sequences (**  **)** .

A description of how to use the **Extract Sequences** tool can be found in section 13.3.

Click **Next** if you wish to adjust how to handle the results (see section 7.2). If not, click **Finish**.

Chapter 11

Data download

Contents

CLC Drug Discovery Workbench offers different ways of searching and downloading online data. You must be online when initiating and performing the following searches:

11.1 UniProt (Swiss-Prot/TrEMBL) search

This section describes searches in UniProt and the handling of search results. UniProt is a global database of protein sequences.

The UniProt search view (figure 11.1) is opened in this way:

Download | Search for Sequences in UniProt ()

11.1.1 UniProt search options

Conducting a search in **UniProt** from *CLC Drug Discovery Workbench* corresponds to conducting the search on UniProt's website. When conducting the search from *CLC Drug Discovery Workbench*, the results are available and ready to work with straight away.

Above the search fields, you can choose which database to search:

- **Swiss-Prot** This is believed to be the most accurate and best quality protein database available. All entries in the database has been currated manually and data are entered according to the original research paper.

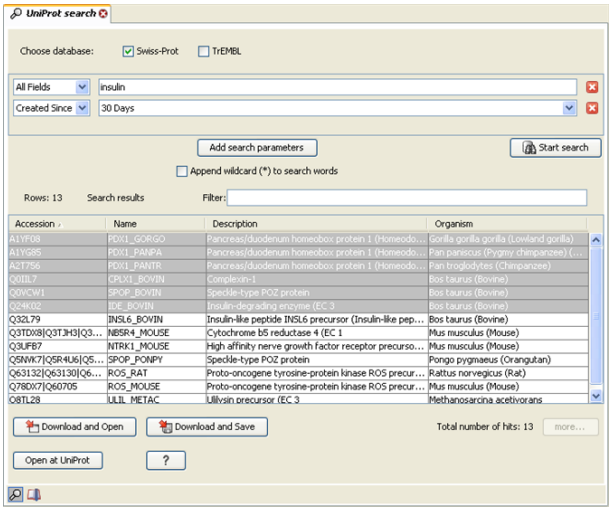


Figure 11.1: *The UniProt search view.*

- **TrEMBL** This database contain computer annotated protein sequences, thus the quality of the annotations is not as good as the Swiss-Prot database.

As default, *CLC Drug Discovery Workbench* offers one text field where the search parameters can be entered. Click **Add search parameters** to add more parameters to your search.
Note! The search is a "and" search, meaning that when adding search parameters to your search, you search for both (or all) text strings rather than "any" of the text strings.

You can append a wildcard character by checking the checkbox at the bottom. This means that you only have to enter the first part of the search text, e.g. searching for "genom" will find both "genomic" and "genome".

The following parameters can be added to the search:

- **All fields**. Text, searches in all parameters in the UniProt database at the same time.
- **Organism**. Text.
- **Description**. Text.
- **Created Since**. Between 30 days and 10 years.
- **Feature**. Text.

The search parameters listed in the dialog are the most recently used. The **All fields** allows searches in all parameters in the UniProt database at the same time.

When you are satisfied with the parameters you have entered, click **Start search**.

Note! When conducting a search, no files are downloaded. Instead, the program produces a list of links to the files in the UniProt database. This ensures a much faster search.

11.1.2 Handling of UniProt search results

The search result is presented as a list of links to the files in the UniProt database. The **View** displays 50 hits at a time (can be changed in the **Preferences** (see chapter 4). More hits can

be displayed by clicking the **More...** button at the bottom right of the **View**. More hits can be displayed by clicking the **More...** button at the bottom left of the **View**.

Each sequence hit is represented by text in three columns:

- Accession
- Name
- Description
- Organism
- Length.

It is possible to exclude one or more of these columns by adjust the View preferences for the database search view. Furthermore, your changes in the View preferences can be saved. See section 4.5.

Several sequences can be selected, and by clicking the buttons in the bottom of the search view, you can do the following:

- Download and open, does not save the sequence.
- Download and save, lets you choose location for saving sequence.
- Open at UniProt, searches the sequence at UniProt's web page.

Double-clicking a hit will download and open the sequence. The hits can also be copied into the **View Area** or the **Navigation Area** from the search results by drag and drop, copy/paste or by using the right-click menu as described below.

Drag and drop from UniProt search results

The sequences from the search results can be opened by dragging them into a position in the **View Area**.

Note! A sequence is not saved until the **View** displaying the sequence is closed. When that happens, a dialog opens: Save changes of sequence x? (Yes or No).

The sequence can also be saved by dragging it into the **Navigation Area**. It is possible to select more sequences and drag all of them into the **Navigation Area** at the same time.

Download UniProt search results using right-click menu

You may also select one or more sequences from the list and download using the right-click menu (see figure **??**). Choosing **Download and Save** lets you select a folder or location where the sequences are saved when they are downloaded. Choosing **Download and Open** opens a new view for each of the selected sequences.

Copy/paste from UniProt search results

When using copy/paste to bring the search results into the **Navigation Area**, the actual files are downloaded from UniProt.

To copy/paste files into the **Navigation Area**:

select one or more of the search results | **Ctrl + C (⌘ + C on Mac)** | **select location or folder in the Navigation Area** | **Ctrl + V**

Note! Search results are downloaded before they are saved. Downloading and saving several files may take some time. However, since the process runs in the background (displayed in the **Toolbox** under the **Processes** tab) it is possible to continue other tasks in the program. Like the search process, the download process can be stopped, paused, and resumed.

11.1.3 Save UniProt search parameters

The search view can be saved either using dragging the search tab and and dropping it in the **Navigation Area** or by clicking **Save** (). When saving the search, only the parameters are saved - not the results of the search. This is useful if you have a special search that you perform from time to time.

Even if you don't save the search, the next time you open the search view, it will remember the parameters from the last time you did a search.

11.2 Search for structures at NCBI

This section describes searches for three dimensional structures from the NCBI structure database http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml. For manipulating and visualization of the downloaded structures see section 9.1.

The NCBI search view is opened in this way:

Download | **Search for structures at NCBI** ()

or **Ctrl + B (⌘ + B on Mac)**

This opens the view shown in figure 11.2:

11.2.1 Structure search options

Conducting a search in the **NCBI Database** from *CLC Drug Discovery Workbench* corresponds to conducting search for structures on the NCBI's Entrez website. When conducting the search from *CLC Drug Discovery Workbench*, the results are available and ready to work with straight away.

As default, *CLC Drug Discovery Workbench* offers one text field where the search parameters can be entered. Click **Add search parameters** to add more parameters to your search.

Note! The search is a "AND" search, meaning that when adding search parameters to your search, you search for both (or all) text strings rather than "any" of the text strings.

You can append a wildcard character by clicking the checkbox at the bottom. This means that you only have to enter the first part of the search text, e.g. searching for "prot" will find both

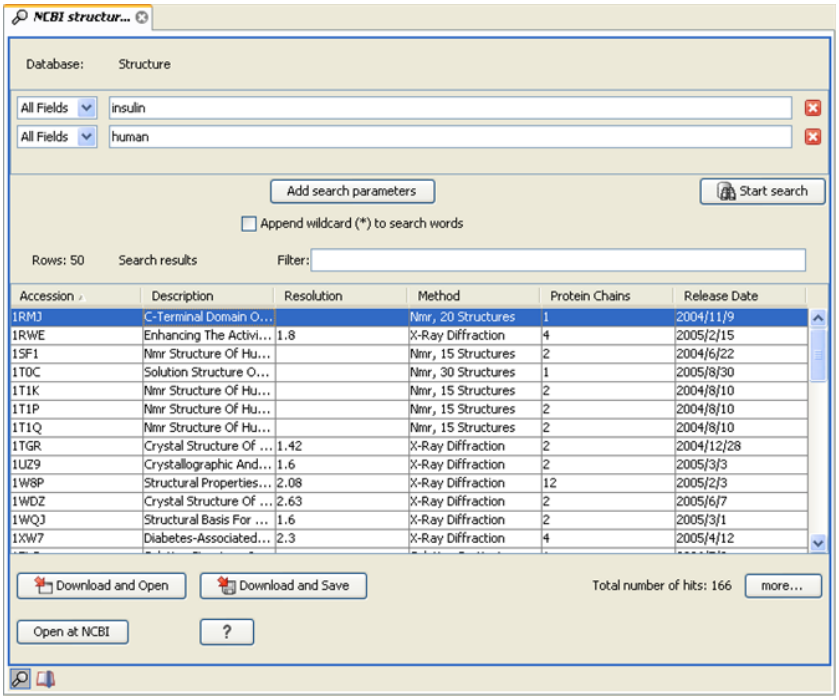


Figure 11.2: *The structure search view.*

"protein" and "protease".

The following parameters can be added to the search:

- **All fields**. Text, searches in all parameters in the NCBI structure database at the same time.
- **Organism**. Text.
- **Author**. Text.
- **PdbAcc**. The accession number of the structure in the PDB database.

The search parameters are the most recently used. The **All fields** allows searches in all parameters in the database at the same time.

All fields also provide an opportunity to restrict a search to parameters which are not listed in the dialog. E.g. writing 'gene[Feature key] AND mouse' in **All fields** generates hits in the GenBank database which contains one or more genes and where 'mouse' appears somewhere in GenBank file. NB: the 'Feature Key' option is only available in GenBank when searching for nucleotide structures. For more information about how to use this syntax, see http://www.ncbi.nlm.nih.gov/entrez/query/static/help/Summary_Matrices.html#Search_Fields_and_Qualifiers

When you are satisfied with the parameters you have entered click **Start search**.

Note! When conducting a search, no files are downloaded. Instead, the program produces a list of links to the files in the NCBI database. This ensures a much faster search.

11.2.2 Handling of NCBI structure search results

The search result is presented as a list of links to the files in the NCBI database. The **View** displays 50 hits at a time (can be changed in the **Preferences** (see chapter 4). More hits can be displayed by clicking the **More...** button at the bottom right of the **View**.

Each structure hit is represented by text in three columns:

- Accession.
- Description.
- Resolution.
- Method.
- Protein chains
- Release date.

It is possible to exclude one or more of these columns by adjust the View preferences for the database search view. Furthermore, your changes in the View preferences can be saved. See section 4.5.

Several structures can be selected, and by clicking the buttons in the bottom of the search view, you can do the following:

- **Download and open.** Download and open immediately.
- **Download and save.** Download and save lets you choose location for saving structure.
- **Open at NCBI.** Open additional information on the selected structure at NCBI's web page.

Double-clicking a hit will download and open the structure. The hits can also be copied into the **View Area** or the **Navigation Area** from the search results by drag and drop, copy/paste or by using the right-click menu as described below.

Drag and drop from structure search results

The structures from the search results can be opened by dragging them into a position in the **View Area**.

Note! A structure is not saved until the **View** displaying the structure is closed. When that happens, a dialog opens: Save changes of structure x? (Yes or No).

The structure can also be saved by dragging it into the **Navigation Area**. It is possible to select more structures and drag all of them into the **Navigation Area** at the same time.

Download structure search results using right-click menu

You may also select one or more structures from the list and download using the right-click menu (see figure 11.3). Choosing **Download and Save** lets you select a folder or location where the

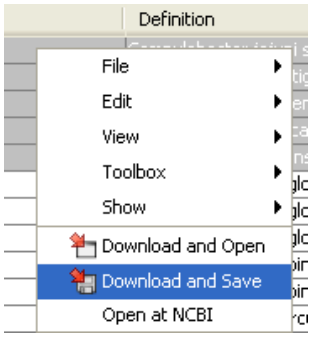


Figure 11.3: *By right-clicking a search result, it is possible to choose how to handle the relevant structure.*

structures are saved when they are downloaded. Choosing **Download and Open** opens a new view for each of the selected structures.

The selected structures are not downloaded from the NCBI website but is downloaded from the RCSB Protein Data Bank `http://www.rcsb.org/pdb/home/home.do` in PDB format.

Copy/paste from structure search results

When using copy/paste to bring the search results into the **Navigation Area**, the actual files are downloaded.

To copy/paste files into the **Navigation Area**:

> **select one or more of the search results** | **Ctrl + C (⌘ + C on Mac)** | **select location or folder in the Navigation Area** | **Ctrl + V**

Note! Search results are downloaded before they are saved. Downloading and saving several files may take some time. However, since the process runs in the background (displayed in the **Status bar**) it is possible to continue other tasks in the program. Like the search process, the download process can be stopped. This is done in the **Toolbox** in the **Processes** tab.

11.2.3 Save structure search parameters

The search view can be saved either using dragging the search tab and and dropping it in the **Navigation Area** or by clicking **Save** ( ). When saving the search, only the parameters are saved - not the results of the search. This is useful if you have a special search that you perform from time to time.

Even if you don't save the search, the next time you open the search view, it will remember the parameters from the last time you did a search.

11.3 Sequence web info

CLC Drug Discovery Workbench provides direct access to web-based search in various databases and on the Internet using your computer's default browser. You can look up a sequence in the databases of NCBI and UniProt, search for a sequence on the Internet using Google and search for Pubmed references at NCBI. This is useful for quickly obtaining updated and additional information about a sequence.

The functionality of these search functions depends on the information that the sequence contains. You can see this information by viewing the sequence as text (see section 10.5). In the following sections, we will explain this in further detail.

The procedure for searching is identical for all four search options (see also figure 11.4):

> **Open a sequence or a sequence list | Right-click the name of the sequence | Web Info ( ) | select the desired search function**

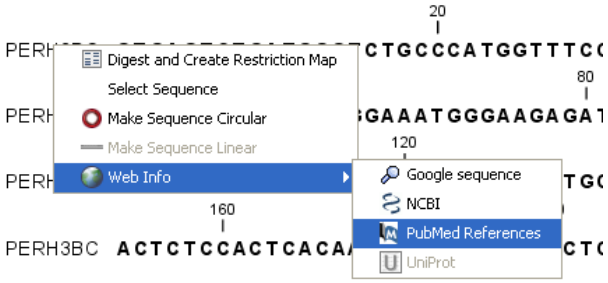


Figure 11.4: *Open webpages with information about this sequence.*

This will open your computer's default browser searching for the sequence that you selected.

Google sequence The Google search function uses the accession number of the sequence which is used as search term on `http://www.google.com`. The resulting web page is equivalent to typing the accession number of the sequence into the search field on `http://www.google.com`.

PubMed References The PubMed references search option lets you look up Pubmed articles based on references contained in the sequence file (when you view the sequence as text it contains a number of "PUBMED" lines). Not all sequence have these PubMed references, but in this case you will se a dialog and the browser will not open.

UniProt The UniProt search function searches in the UniProt database (`http://www.ebi.uniprot.org`) using the accession number. Furthermore, it checks whether the sequence was indeed downloaded from UniProt.

Additional annotation information When sequences are downloaded from GenBank they often link to additional information on taxonomy, conserved domains etc. If such information is available for a sequence it is possible to access additional accurate online information. If the db_xref identifier line is found as part of the annotation information in the downloaded GenBank file, it is possible to easily look up additional information on the NCBI web-site.

To access this feature, simply right click an annotation and see which databases are available.

Chapter 12

BLAST search

Contents

CLC Drug Discovery Workbench offers to conduct BLAST searches on protein and DNA sequences. In short, a BLAST search identifies homologous sequences between your input (query) query sequence and a database of sequences [McGinnis and Madden, 2004]. BLAST (Basic Local Alignment Search Tool), identifies homologous sequences using a heuristic method which finds short matches between two sequences. After initial match BLAST attempts to start local alignments from these initial matches.

If you are interested in the bioinformatics behind BLAST, there is an easy-to-read explanation of this in section 12.6.

Figure 12.9 shows an example of a BLAST result in the *CLC Drug Discovery Workbench*.

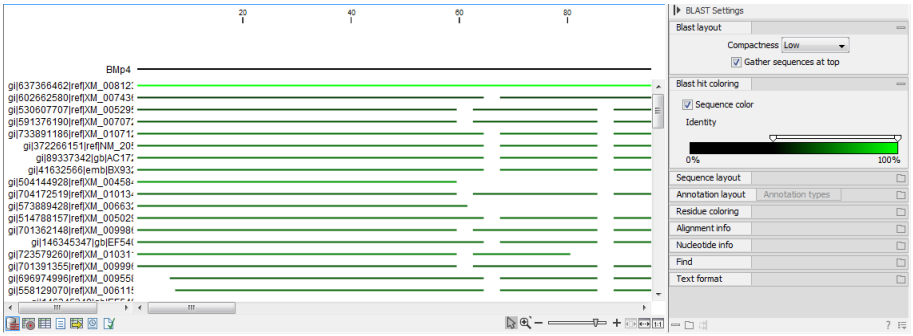


Figure 12.1: *Display of the output of a BLAST search. At the top is there a graphical representation of BLAST hits with tool-tips showing additional information on individual hits. Below is a tabular form of the BLAST results.*

12.1 Running BLAST searches

With the *CLC Drug Discovery Workbench* there are two ways of performing BLAST searches: You can either have the BLAST process run on NCBI's BLAST servers (`http://www.ncbi.nlm.nih.gov/`) or you can perform the BLAST search on your own computer.

The advantage of running the BLAST search on NCBI servers is that you have readily access to the popular, and often very large, BLAST databases without having to download them to your own computer. The advantages of running BLAST on your own computer include that you can use your own sequence collections as blast databases, and that running big batch BLAST jobs can be faster and more reliable when done locally.

12.1.1 BLAST at NCBI

When running a BLAST search at the NCBI, the Workbench sends the sequences you select to the NCBI's BLAST servers. When the results are ready, they will be automatically downloaded and displayed in the Workbench. When you enter a large number of sequences for searching with BLAST, the Workbench automatically splits the sequences up into smaller subsets and sends one subset at the time to NCBI. This is to avoid exceeding any internal limits the NCBI places on the number of sequences that can be submitted to them for BLAST searching. The size of the subset created in the CLC software depends both on the number and size of the sequences.

To start a BLAST job to search your sequences against databases held at the NCBI, go to:

Toolbox | **Sequence Analysis** ( ) | **BLAST** ( )| **BLAST at NCBI** ( )

Alternatively, use the keyboard shortcut: Ctrl+Shift+B for Windows and ⌘ +Shift+B on Mac OS.

This opens the dialog seen in figure 12.2

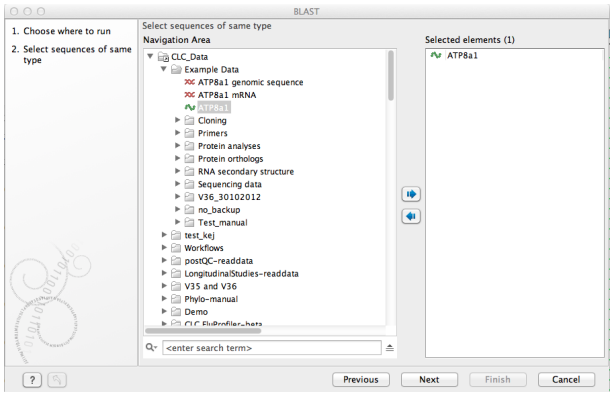


Figure 12.2: *Choose one or more sequences to conduct a BLAST search with.*

Select one or more sequences of the same type (either DNA or protein) and click **Next**.

In this dialog, you choose which type of BLAST search to conduct, and which database to search against (figure 12.3). The databases at the NCBI listed in the dropdown box will correspond to the query sequence type you have, DNA or protein, and the type of blast search you can chose among to run. A complete list of these databases can be found in Appendix C. Here you can also read how to add additional databases available the NCBI to the list provided in the dropdown menu.

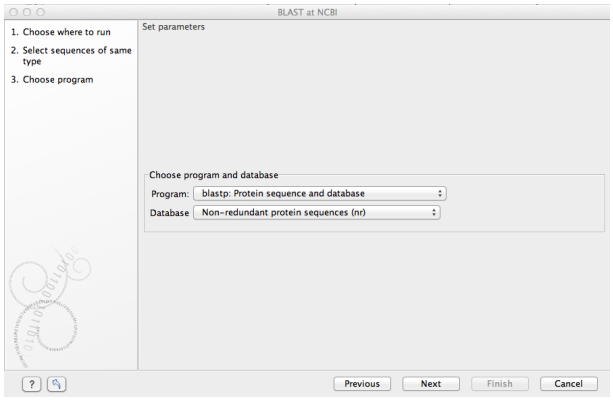


Figure 12.3: *Choose a BLAST Program and a database for the search.*

BLAST programs for DNA query sequences:

- **blastn: DNA sequence against a DNA database.** Searches for DNA sequences with homologous regions to your nucleotide query sequence.
- **blastx: Translated DNA sequence against a Protein database.** Automatic translation of your DNA query sequence in six frames; these translated sequences are then used to search a protein database.
- **tblastx: Translated DNA sequence against a Translated DNA database.** Automatic translation of your DNA query sequence and the DNA database, in six frames. The resulting

peptide query sequences are used to search the resulting peptide database. Note that this type of search is computationally intensive.

BLAST programs for protein query sequences:

- **blastp: Protein sequence against Protein database.** Used to look for peptide sequences with homologous regions to your peptide query sequence.
- **tblastn: Protein sequence against Translated DNA database.** Peptide query sequences are searched against an automatically translated, in six frames, DNA database.

If you search against the **Protein Data Bank protein** database homologous sequences are found to the query sequence, these can be downloaded and opened with the 3D view.

Click **Next**.

This window, see figure 12.4, allows you to choose parameters to tune your BLAST search, to meet your requirements.

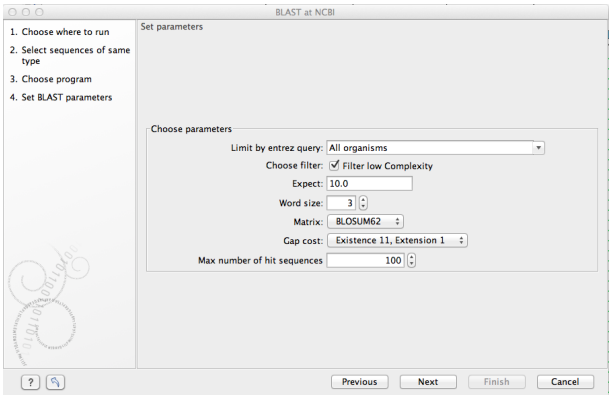


Figure 12.4: *Parameters that can be set before submitting a BLAST search.*

When choosing blastx or tblastx to conduct a search, you get the option of selecting a translation table for the genetic code. The standard genetic code is set as default. This setting is particularly useful when working with organisms or organelles that have a genetic code different from the standard genetic code.

The following description of BLAST search parameters is based on information from http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml.

- **Limit by Entrez query.** BLAST searches can be limited to the results of an Entrez query against the database chosen. This can be used to limit searches to subsets of entries in the BLAST databases. Any terms can be entered that would normally be allowed in an Entrez search session. More information about Entrez queries can be found at http://www.ncbi.nlm.nih.gov/books/NBK3837/#EntrezHelp.Entrez_Searching_Options. The syntax described there is the same as would be accepted in the CLC interface. Some commonly used Entrez queries are pre-entered and can be chosen in the drop down menu.
- **Choose filter.** You can choose to apply **Low-complexity**. Mask off segments of the query sequence that have low compositional complexity. Filtering can eliminate statistically

significant, but biologically uninteresting reports from the BLAST output (e.g. hits against common acidic-, basic- or proline-rich regions), leaving the more biologically interesting regions of the query sequence available for specific matching against database sequences.

- **Expect.** The threshold for reporting matches against database sequences: the default value is 10, meaning that under the circumstances of this search, 10 matches are expected to be found merely by chance according to the stochastic model of Karlin and Altschul (1990). Details of how E-values are calculated can be found at the NCBI: http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html If the E-value ascribed to a match is greater than the EXPECT threshold, the match will not be reported. Lower EXPECT thresholds are more stringent, leading to fewer chance matches being reported. Increasing the threshold results in more matches being reported, but many may just matching by chance, not due to any biological similarity. Values of E less than one can be entered as decimals, or in scientific notiation. For example, 0.001, 1e-3 and 10e-4 would be equivalent and acceptable values.

- **Word Size.** BLAST is a heuristic that works by finding word-matches between the query and database sequences. You may think of this process as finding "hot-spots" that BLAST can then use to initiate extensions that might lead to full-blown alignments. For nucleotide-nucleotide searches (i.e. "BLASTn") an exact match of the entire word is required before an extension is initiated, so that you normally regulate the sensitivity and speed of the search by increasing or decreasing the wordsize. For other BLAST searches non-exact word matches are taken into account based upon the similarity between words. The amount of similarity can be varied so that you normally uses just the wordsizes 2 and 3 for these searches.

- **Match/mismatch** A key element in evaluating the quality of a pairwise sequence alignment is the "substitution matrix", which assigns a score for aligning any possible pair of residues. The matrix used in a BLAST search can be changed depending on the type of sequences you are searching with (see the BLAST Frequently Asked Questions). Only applicable for protein sequences or translated DNA sequences.

- **Gap Cost.** The pull down menu shows the Gap Costs (Penalty to open Gap and penalty to extend Gap). Increasing the Gap Costs and Lambda ratio will result in alignments which decrease the number of Gaps introduced.

- **Max number of hit sequences.** The maximum number of database sequences, where BLAST found matches to your query sequence, to be included in the BLAST report.

The parameters you choose will affect how long BLAST takes to run. A search of a small database, requesting only hits that meet stringent criteria will generally be quite quick. Searching large databases, or allowing for very remote matches, will of course take longer.

Click **Next** if you wish to adjust how to handle the results (see section 7.2). If not, click **Finish**.

12.1.2 BLAST a partial sequence against NCBI

You can search a database using only a part of a sequence directly from the sequence view:

select the sequence region to send to BLAST | right-click the selection | BLAST Selection Against NCBI ()

This will go directly to the dialog shown in figure 12.3 and the rest of the options are the same as when performing a BLAST search with a full sequence.

12.1.3 BLAST against local data

Running BLAST searches on your local machine can have several advantages over running the searches remotely at the NCBI:

- It can be faster.
- It does not rely on having a stable internet connection.
- It does not depend on the availability of the NCBI BLAST servers.
- You can use longer query sequences.
- You use your own data sets to search against.

On a technical level, the *CLC Drug Discovery Workbench* uses the NCBI's blast+ software (see `ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/`). Thus, the results of using a particular data set to search the same database, with the same search parameters, would give the same results, whether run locally or at the NCBI.

There are a number of options for what you can search against:

- You can create a database based on data already imported into your Workbench (see section 12.4.3)
- You can add pre-formatted databases (see section 12.4.1)
- You can use sequence data from the **Navigation Area** directly, without creating a database first.

To conduct a local BLAST search, go to:

Toolbox | **Sequence Analysis** () | **BLAST** ()| **BLAST** ()

This opens the dialog seen in figure 12.5:

Select one or more sequences of the same type (DNA or protein) and click **Next**.

This opens the dialog seen in figure 12.6:

At the top, you can choose between different BLAST programs.

BLAST programs for DNA query sequences:

- **blastn: DNA sequence against a DNA database.** Searches for DNA sequences with homologous regions to your nucleotide query sequence.
- **blastx: Translated DNA sequence against a Protein database.** Automatic translation of your DNA query sequence in six frames; these translated sequences are then used to search a protein database.

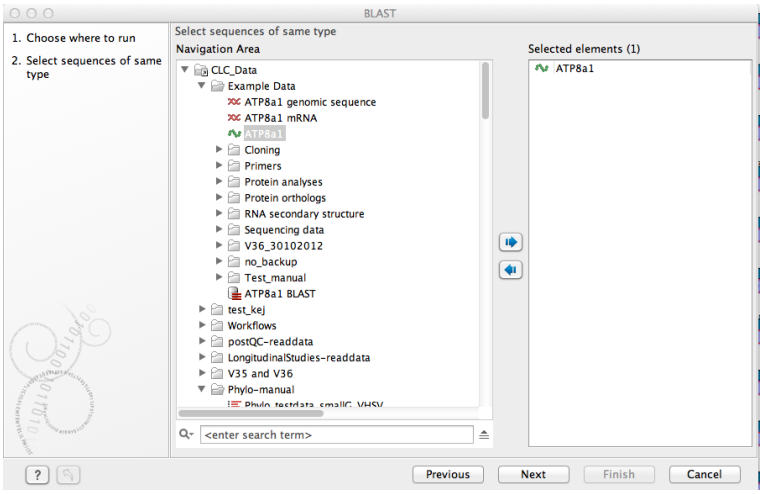


Figure 12.5: *Choose one or more sequences to conduct a BLAST search.*

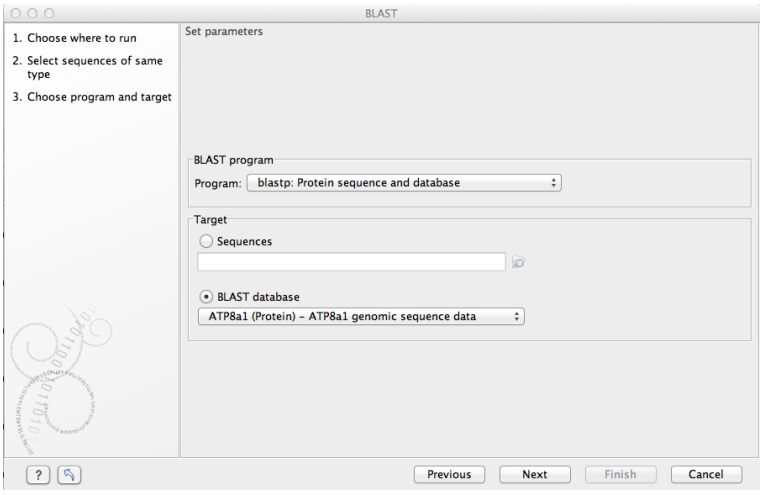


Figure 12.6: *Choose a BLAST program and a target database.*

- **tblastx: Translated DNA sequence against a Translated DNA database.** Automatic translation of your DNA query sequence and the DNA database, in six frames. The resulting peptide query sequences are used to search the resulting peptide database. Note that this type of search is computationally intensive.

BLAST programs for protein query sequences:

- **blastp: Protein sequence against Protein database.** Used to look for peptide sequences with homologous regions to your peptide query sequence.
- **tblastn: Protein sequence against Translated DNA database.** Peptide query sequences are searched against an automatically translated, in six frames, DNA database.

In cases where you have selected blastx or tblastx to conduct a search, you will get the option of selecting a translation table for the genetic code. The standard genetic code is set as default. This setting is particularly useful when working with organisms or organelles that have a genetic code that differs from the standard genetic code.

If you search against the **Protein Data Bank** database and homologous sequences are found to the query sequence, these can be downloaded and opened with the **3D Molecule Viewer** (see section 6.2.5).

Click **Next**.

This dialog allows you to adjust the parameters to meet the requirements of your BLAST search (figure 12.7).

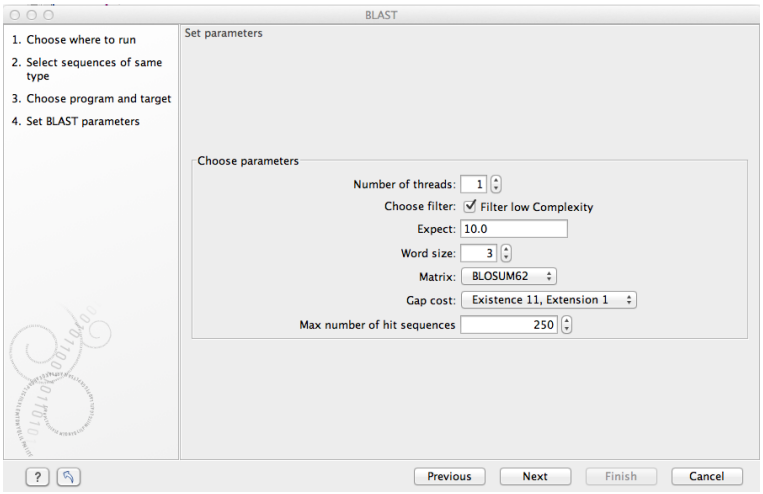


Figure 12.7: *Parameters that can be set before submitting a local BLAST search.*

- **Number of threads.** You can specify the number of threads, which should be used if your Workbench is installed on a multi-threaded system.
- **Choose filter.** You can choose to apply **Low-complexity**. Mask off segments of the query sequence that have low compositional complexity. Filtering can eliminate statistically significant, but biologically uninteresting reports from the BLAST output (e.g. hits against common acidic-, basic- or proline-rich regions), leaving the more biologically interesting regions of the query sequence available for specific matching against database sequences.
- **Expect.** The threshold for reporting matches against database sequences: the default value is 10, meaning that under the circumstances of this search, 10 matches are expected to be found merely by chance according to the stochastic model of Karlin and Altschul (1990). Details of how E-values are calculated can be found at the NCBI: `http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html` If the E-value ascribed to a match is greater than the EXPECT threshold, the match will not be reported. Lower EXPECT thresholds are more stringent, leading to fewer chance matches being reported. Increasing the threshold results in more matches being reported, but many may just matching by chance, not due to any biological similarity. Values of E less than one can be entered as decimals, or in scientific notiation. For example, 0.001, 1e-3 and 10e-4 would be equivalent and acceptable values.
- **Word Size.** BLAST is a heuristic that works by finding word-matches between the query and database sequences. You may think of this process as finding "hot-spots" that BLAST can then use to initiate extensions that might lead to full-blown alignments. For nucleotide-nucleotide searches (i.e. "BLASTn") an exact match of the entire word is required before an extension is initiated, so that you normally regulate the sensitivity and speed of the

search by increasing or decreasing the wordsize. For other BLAST searches non-exact word matches are taken into account based upon the similarity between words. The amount of similarity can be varied so that you normally uses just the wordsizes 2 and 3 for these searches.

- **Match/mismatch** A key element in evaluating the quality of a pairwise sequence alignment is the "substitution matrix", which assigns a score for aligning any possible pair of residues. The matrix used in a BLAST search can be changed depending on the type of sequences you are searching with (see the BLAST Frequently Asked Questions). Only applicable for protein sequences or translated DNA sequences.
- **Gap Cost.** The pull down menu shows the Gap Costs (Penalty to open Gap and penalty to extend Gap). Increasing the Gap Costs and Lambda ratio will result in alignments which decrease the number of Gaps introduced.
- **Max number of hit sequences.** The maximum number of database sequences, where BLAST found matches to your query sequence, to be included in the BLAST report.

You then specify the target database to use:

- **Sequences**. When you choose this option, you can use sequence data from the **Navigation Area** as database by clicking the **Browse and select** icon ( ). A temporary BLAST database will be created from these sequences and used for the BLAST search. It is deleted afterwards. If you want to be able to click in the BLAST result to retrieve the hit sequences from the BLAST database at a later point, you should *not* use this option; create a create a BLAST database first, see section 12.4.3.
- **BLAST Database**. Select a database already available in one of your designated BLAST database folders. Read more in section 12.5.

When a database or a set of sequences has been selected, click **Next**.

This opens the dialog seen in figure 12.8:

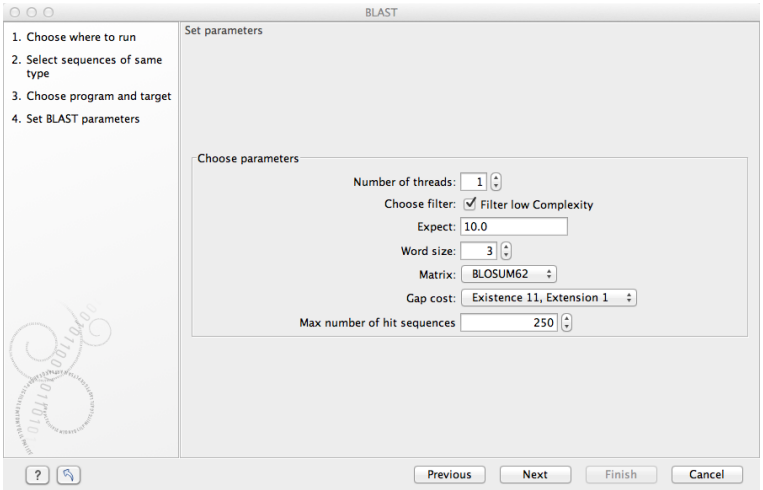


Figure 12.8: *Examples of parameters that can be set before submitting a BLAST search.*

See section 12.1.1 for information about these limitations.

12.1.4 BLAST a partial sequence against a local database

You can search a database using only a part of a sequence directly from the sequence view:

select the region that you wish to BLAST | right-click the selection | BLAST Selection Against Local Database ( )

This will go directly to the dialog shown in figure 12.6 and the rest of the options are the same as when performing a BLAST search with a full sequence.

12.2 Output from BLAST searches

The output of a BLAST search is similar whether you have chosen to run your search locally or at the NCBI.

If a **single query** sequence was used, then the results will show the hits and High-Scoring Segment Pairs (HSPs) found in that database with that single sequence. If **more than one query** sequence was used, the default view of the results is a summary table, where the description of the top match found for each query sequence and the number of matches found is reported. The summary table is described in detail in section 12.2.2.

12.2.1 Graphical overview for each query sequence

Double clicking on a given row of a tabular blast table opens a graphical overview of the blast results for a particular query sequence, as shown in figure figure 12.9. In cases where only one sequence was entered into a BLAST search, such a graphical overview is the default output.

Figure 12.9 shows an example of a BLAST result for an individual query sequence in the *CLC Drug Discovery Workbench*.

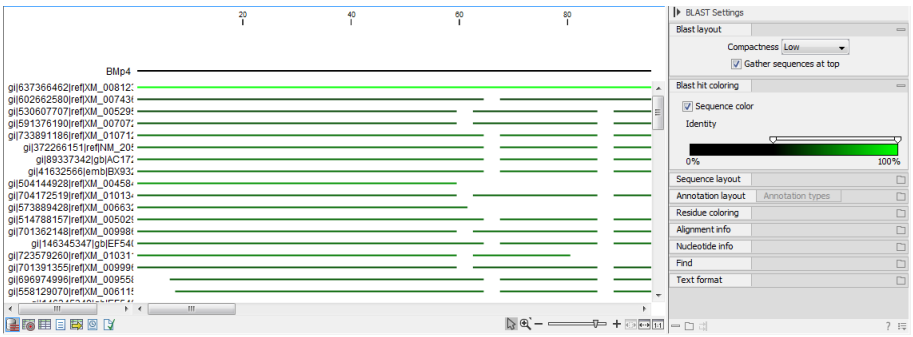


Figure 12.9: *Default display of the output of a BLAST search for one query sequence. At the top is there a graphical representation of BLAST hits with tooltips showing additional information on individual hits.*

Detailed descriptions of the overview BLAST table and the graphical BLAST results view are described below.

12.2.2 Overview BLAST table

In the overview BLAST table for a multi-sequence blast search, as shown in figure 12.10, there is one row for each query sequence. Each row represents the BLAST result for this query sequence.

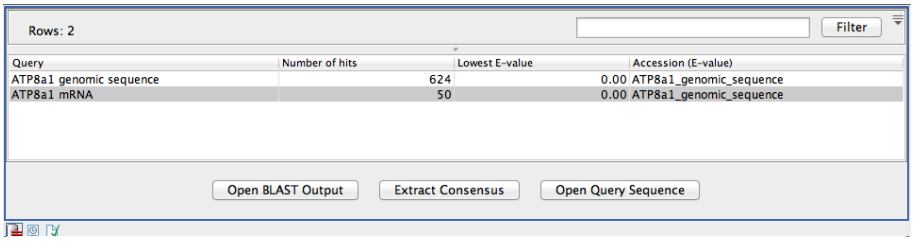


Figure 12.10: *An overview BLAST table summarizing the results for a number of query sequences.*

Double-clicking a row will open the BLAST result for this query sequence, allowing more detailed investigation of the result. You can also select one or more rows and click the **Open BLAST Output** button at the bottom of the view. Consensus sequence can be extracted by clicking the **Extract Consensus** button at the bottom. Clicking the **Open Query Sequence** will open a sequence list with the selected query sequences. This can be useful in work flows where BLAST is used as a filtering mechanism where you can filter the table to include e.g. sequences that have a certain top hit and then extract those.

In the overview table, the following information is shown:

- Query: Since this table displays information about several query sequences, the first column is the name of the query sequence.
- Number of HSPs: The number of High-scoring Segment Pairs (HSPs) for this query sequence.
- For the following list, the value of the best HSP is displayed together with accession number and description of this HSP, with respect to E-value, identity or positive value, hit length or bit score.
  - Lowest E-value
  - Accession (E-value)
  - Description (E-value)
  - Greatest identity %
  - Accession (identity %)
  - Description (identity %)
  - Greatest positive %
  - Accession (positive %)
  - Description (positive %)
  - Greatest HSPs length
  - Accession (HSP length)
  - Description (HSP length)
  - Greatest bit score
  - Accession (bit score)
  - Description (bit score)

If you wish to save some of the BLAST results as individual elements in the **Navigation Area**, open them and click **Save As** in the **File** menu.

12.2.3 BLAST graphics

The **BLAST editor** shows the sequences hits which were found in the BLAST search. The hit sequences are represented by colored horizontal lines, and when hovering the mouse pointer over a BLAST hit sequence, a tooltip appears, listing the characteristics of the sequence. As default, the query sequence is fitted to the window width, but it is possible to zoom in the windows and see the actual sequence alignments returned from the BLAST server.

There are several settings available in the **BLAST Settings** side panel.

- **Blast layout.** You can control the level of **Compactness** for displaying sequences:
  - **Not compact.** Full detail and spaces between the sequences.
  - **Low.** The normal settings where the residues are visible (when zoomed in) but with no extra spaces between.
  - **Medium.** The sequences are represented as lines and the residues are not visible. There is some space between the sequences.
  - **Compact.** Even less space between the sequences.

  You can also choose to **Gather sequences at top**. Enabling this option affects the view that is shown when scrolling horizontally along a BLAST result. If selected, the sequence hits which did not contribute to the visible part of the BLAST graphics will be omitted whereas the found BLAST hits will automatically be placed right below the query sequence.
- **BLAST hit coloring.** You can choose whether to color hit sequences and adjust the coloring scale for visualisation of identity level.

The remaining View preferences for BLAST Graphics are the same as those of alignments. See section 10.1.

Some of the information available in the tooltips when hovering over a particular hit sequence is:

- **Name of sequence.** Here is shown some additional information of the sequence which was found. This line corresponds to the description line in GenBank (if the search was conducted on the nr database).
- **Score.** This shows the bit score of the local alignment generated through the BLAST search.
- **Expect.** Also known as the E-value. A low value indicates a homologous sequence. Higher E-values indicate that BLAST found a less homologous sequence.
- **Identities.** This number shows the number of identical residues or nucleotides in the obtained alignment.
- **Gaps.** This number shows whether the alignment has gaps or not.
- **Strand.** This is only valid for nucleotide sequences and show the direction of the aligned strands. Minus indicate a complementary strand.

The numbers of the query and subject sequences refer to the sequence positions in the submitted and found sequences. If the subject sequence has number 59 in front of the sequence, this

means that 58 residues are found upstream of this position, but these are not included in the alignment.

By right clicking the sequence name in the Graphical BLAST output it is possible to download the full hits sequence from NCBI with accompanying annotations and information. It is also possible to just open the actual hit sequence in a new view.

12.2.4 BLAST HSP table

In addition to the graphical display of a BLAST result, it is possible to view the BLAST results in a tabular view. In the tabular view, one can get a quick and fast overview of the results. Here you can also select multiple sequences and download or open all of these in one single step. Moreover, there is a link from each sequence to the sequence at NCBI. These possibilities are either available through a right-click with the mouse or by using the buttons below the table.

The **BLAST table** view can be shown in the following way:

Click the Show BLAST HSP Table button () at the bottom of the view

Figure 12.11 is an example of a BLAST HSP Table.

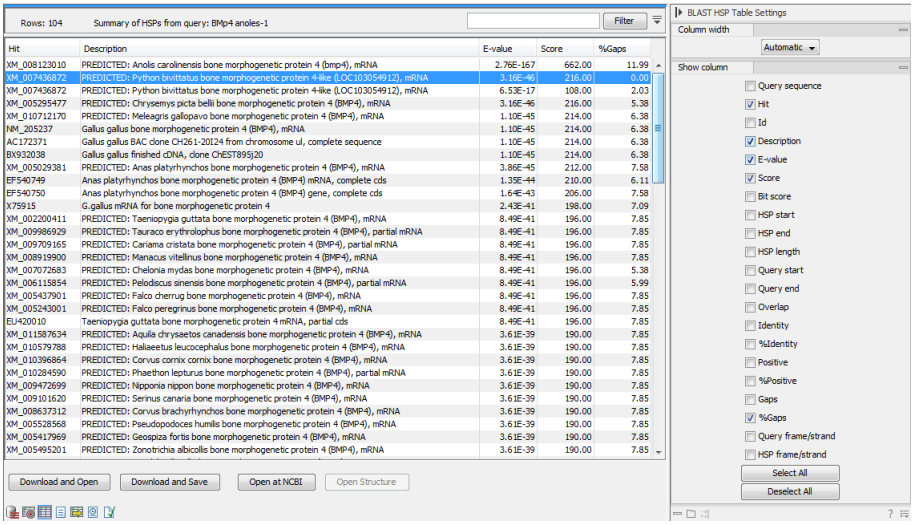


Figure 12.11: *BLAST HSP Table. The HSPs can be sorted by the different columns, simply by clicking the column heading.*

The BLAST HSP Table includes the following information:

- **Query sequence.** The sequence which was used for the search.
- **HSP.** The Name of the sequences found in the BLAST search.
- **Id.** GenBank ID.
- **Description.** Text from NCBI describing the sequence.
- **E-value.** Measure of quality of the match. Higher E-values indicate that BLAST found a less homologous sequence.
- **Score.** This shows the score of the local alignment generated through the BLAST search.

- **Bit score.** This shows the bit score of the local alignment generated through the BLAST search. Bit scores are normalized, which means that the bit scores from different alignments can be compared, even if different scoring matrices have been used.
- **HSP start.** Shows the start position in the HSP sequence.
- **HSP end.** Shows the end position in the HSP sequence.
- **HSP length.** The length of the HSP.
- **Query start.** Shows the start position in the query sequence.
- **Query end.** Shows the end position in the query sequence.
- **Overlap.** Display a percentage value for the overlap of the query sequence and HSP sequence. Only the length of the local alignment is taken into account and not the full length query sequence.
- **Identity.** Shows the number of identical residues in the query and HSP sequence.
- **%Identity.** Shows the percentage of identical residues in the query and HSP sequence.
- **Positive.** Shows the number of similar but not necessarily identical residues in the query and HSP sequence.
- **%Positive.** Shows the percentage of similar but not necessarily identical residues in the query and HSP sequence.
- **Gaps.** Shows the number of gaps in the query and HSP sequence.
- **%Gaps.** Shows the percentage of gaps in the query and HSP sequence.
- **Query Frame/Strand.** Shows the frame or strand of the query sequence.
- **HSP Frame/Strand.** Shows the frame or strand of the HSP sequence.

In the **BLAST table** view you can handle the HSP sequences. Select one or more sequences from the table, and apply one of the following functions.

- **Download and Open.** Download the full sequence from NCBI and opens it. If multiple sequences are selected, they will all open (if the same sequence is listed several times, only one copy of the sequence is downloaded and opened).
- **Download and Save.** Download the full sequence from NCBI and save it. When you click the button, there will be a save dialog letting you specify a folder to save the sequences. If multiple sequences are selected, they will all open (if the same sequence is listed several times, only one copy of the sequence is downloaded and opened).
- **Open at NCBI.** Opens the corresponding sequence(s) at GenBank at NCBI. Here is stored additional information regarding the selected sequence(s). The default Internet browser is used for this purpose.
- **Open structure.** If the HSP sequence contain structure information, the sequence is opened in a text view or a 3D view (3D view in *CLC Main Workbench* or *CLC Genomics Workbench*).

The HSPs can be sorted by the different columns, simply by clicking the column heading. In cases where individual rows have been selected in the table, the selected rows will still be selected after sorting the data.

You can do a text-based search in the information in the BLAST table by using the filter at the upper right part of the view. In this way you can search for e.g. species or other information which is typically included in the "Description" field.

The table is integrated with the graphical view described in section 12.2.3 so that selecting a HSP in the table will make a selection on the corresponding sequence in the graphical view.

12.2.5 BLAST hit table

The **BLAST Hit table** view can be shown in the following way:

Click the Show BLAST Hit Table button () at the bottom of the view

Figure 12.12 is an example of a BLAST Hit Table.

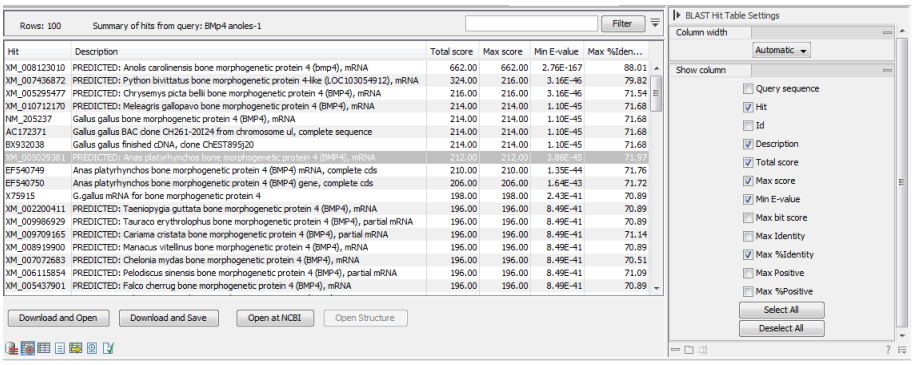


Figure 12.12: *BLAST Hit Table. The hits can be sorted by the different columns, simply by clicking the column heading.*

The BLAST Hit Table includes the following information:

- **Query sequence.** The sequence which was used for the search.
- **Hit.** The Name of the sequences found in the BLAST search.
- **Id.** GenBank ID.
- **Description.** Text from NCBI describing the sequence.
- **Total Score.** Total score for all HSPs.
- **Max Score.** Maximum score of all HSPs.
- **Min E-value.** Minimum e-value of all HSPs.
- **Max Bit score.** Maximum Bit score of all HSPs.
- **Max Identity.** Shows the maximum number of identical residues in the query and Hit sequence.

- **Max %Identity.** Shows the percentage of maximum identical residues in the query and Hit sequence.
- **Max Positive.** Shows the maximum number of similar but not necessarily identical residues in the query and Hit sequence.
- **Max %Positive.** Shows the percentage of maximum similar but not necessarily identical residues in the query and Hit sequence.

12.3 Extract consensus sequence

You can extract a consensus sequence from a BLAST result. Clicking on the button Extract Consensus Sequence opens a dialog where you can decide how to handle regions with low coverage. The first step is to define a **threshold for when coverage is considered low**. The default value is 0, which means that low coverage is defined as no coverage (i.e. no reads align to the reference at this position). That means if you have one read covering a given position, it will only be that read that determines the consensus sequence. If you need more confidence that the consensus sequence is correct, we advise raising this value. Setting a higher low coverage threshold will require more mapped reads to construct the consensus sequence.

A consensus based on mapped reads cannot be generated in regions that meet or are below the value set for the low coverage threshold, there are several options for handling these low coverage regions:

- **Remove regions with low coverage**. When using this option, no consensus sequence is created for the low coverage regions. There are two ways of creating the consensus sequence from the remaining contiguous stretches of high coverage: either the consensus sequence is **split** into separate sequence when there is a low coverage region, or the low coverage region is simply ignored, and the high-coverage regions are directly **joined** (in this case, an annotation is added at the position where a low coverage region is removed in the consensus sequence produced, see below).
- **Insert 'N' ambiguity symbols**. This will simply add Ns for each base in the low coverage region. An annotation is added for the low coverage region in the consensus sequence produced (see below).
- **Fill from reference sequence**. This option will use the sequence from the reference to construct the consensus sequence for low coverage regions. An annotation is added for the low coverage region in the consensus sequence produced (see below).

In addition to deciding how to handle low coverage regions, you can also decide how to handle conflicts or disagreement between the reads when building a consensus sequence in regions above the low coverage threshold:

- **Vote**. Whenever the reads disagree on the base at a given position, the vote resolution will let the majority of the reads decide which base is correct. In addition, you can specify to let the voting use the base calling **quality scores** from the reads. This is done be simply adding all quality scores for each base and let the sum determine which one is correct. The base with the highest total quality scores will be chosen. If there are two bases that end up summing to the same total quality score for all reads at that location, A is preferred before

C, C before G, and G before T. An annotation with the complete information that was used to resolve the conflict will be added.

- **Insert ambiguity codes**. When this options is selected, read conflicts are addressed by using an ambiguity code representing all read bases represented at the reference location. The problem with the voting option is that it will not be able to represent true biological heterozygous variation in the data. For a diploid genome, if two different alleles are present in an almost even number of reads, only one will be represented in the consensus sequence. With the option to insert ambiguity codes, this can be solved. However, if an ambiguity code would always be inserted if just one read had a different base, there would be an ambiguity code whenever there was a sequencing error. In high-coverage NGS data that would be a big problem, because sequencing errors would be abundant. To solve this problem, you can specify a **Noise threshold**. The default value for this is 0.1 which means that for a base to contribute to the ambiguity code, it must be in at least 10 % of the reads at a given position. The **Minimum nucleotide count** specifies the minimum number of reads that are required before a nucleotide is included. Nucleotides below this limit are considered noise.

- **Use quality score**. The "Use quality score" checkbox option is available for conflicts regardless of whether "Vote" or "Insert ambiguity codes" has been selected. The "Use quality score" checkbox option allows you to use the base calling **quality scores** from the reads. This is done by simply adding all the quality scores for each base and let the sum determine which bases to consider. In other words, if quality scores are used, we will sum the quality score (instead of amount of reads) for each base on each position before applying the noise filters and finally call the consensus symbol.

Click **Next** to set the output option as shown in figure 12.13).

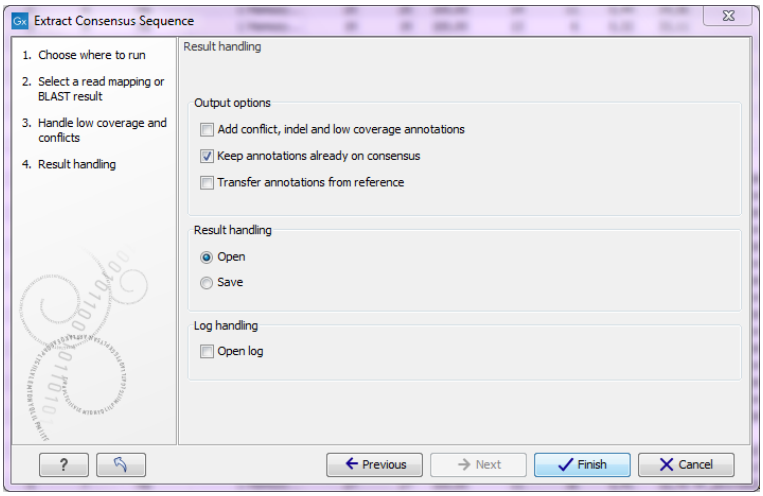


Figure 12.13: *Choose to add annotations to the consensus sequence.*

The annotations that can be added to the consensus sequence produced by this tool show both conflicts that have been resolved and low coverage regions (unless you have chosen to split the consensus sequence). Please note that for large data sets, this can amount to a very high number of annotations, which will cause the tool to take longer to complete, and the result will take up much more disk space.

It is also possible to transfer existing annotations to the consensus sequence produced. Please note that since the consensus sequence produced may be broken up, the annotations will also be broken up, and you cannot expect them to have the same length as before. In some cases, gaps and low-coverage regions will lead to differences in the sequence coordinates between the input data and the new consensus sequence. The annotations copied will be placed in the region on the consensus that corresponds to the region on the input data, but the actual coordinates might have changed.

Copied/transferred annotations will contain the same qualifier text as the original. That is, the text is not updated. As an example, if the annotation contains 'translation' as qualifier text this translation will be copied to the new sequence and will thus reflect the translation of the original sequence, not the new sequence, which may differ.

The resulting consensus sequence (or sequences) will have quality scores assigned if quality scores were found in the reads used to call the consensus. For a given consensus symbol X we compute its quality score from the "column" in the read mapping. Let Y be the sum of all quality scores corresponding to the "column" below X, and let Z be the sum of all quality scores from that column that supported X[1]. Let $Q = Z - (Y - Z)$, then we will assign X the quality score of q where

$$q = \begin{cases} 64 & \text{if } Q > 64 \\ 0 & \text{if } Q < 0 \\ Q & \text{otherwise} \end{cases}$$

12.4 Local BLAST databases

BLAST databases on your local system can be made available for searches via your *CLC Drug Discovery Workbench*, (section 12.4.1). To make adding databases even easier, you can download pre-formatted BLAST databases from the NCBI from within your *CLC Drug Discovery Workbench*, (section 12.4.2). You can also easily create your own local blast databases from sequences within your *CLC Drug Discovery Workbench*, (section 12.4.3).

12.4.1 Make pre-formatted BLAST databases available

To use databases that have been downloaded or created outside the Workbench, you can either:

- Put the database files in one of the locations defined in the BLAST database manager (see section 12.5). All the files that comprise a given BLAST database must be included. This may be as few as three files, but can be more (figure 12.14).
- Add the location where your BLAST databases are stored using the BLAST database manager (see section 12.5) (figure 12.18).

[1] By supporting a consensus symbol, we understand the following: when conflicts are resolved using voting, then only the reads having the symbol that is eventually called are said to support the consensus. When ambiguity codes are used instead, all reads contribute to the called consensus and thus $Y = Z$.

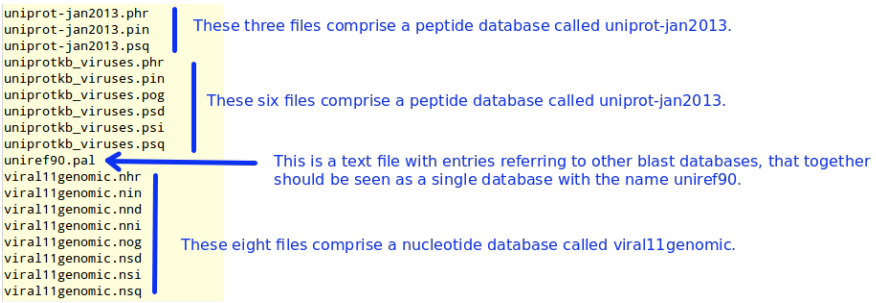


Figure 12.14: *BLAST databases are made up of several files. The exact number varies and depends on the tool used to build the databases as well as how large the database is. Large databases will be split into the number of volumes and there will be several files per volume. If you have made your BLAST database, or downloaded BLAST database files, outside the Workbench, you will need to ensure that all the files associated with that BLAST database are available in a CLC Blast database location.*

12.4.2 Download NCBI pre-formatted BLAST databases

Many popular pre-formatted databases are available for download from the NCBI. You can download any of the databases available from the list at `ftp://ftp.ncbi.nlm.nih.gov/blast/db/` from within your *CLC Drug Discovery Workbench*.

You must be connected to the internet to use this tool.

To download a database, go to:

Toolbox | **Sequence Analysis (**  **)** | **BLAST (**  **)**| **Download BLAST Databases (**  **)**

A window like the one in figure 12.15 pops up showing you the list of databases available for download.

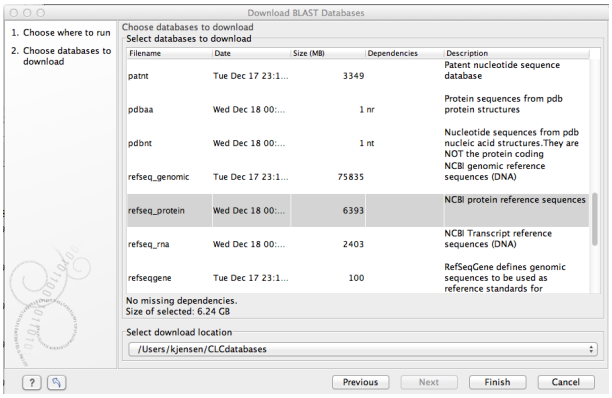


Figure 12.15: *Choose from pre-formatted BLAST databases at the NCBI available for download.*

In this window, you can see the names of the databases, the date they were made available for download on the NCBI site, the size of the files associated with that database, and a brief description of each database. You can also see whether the database has any dependencies. This aspect is described below.

You can also specify which of your database locations you would like to store the files in. Please see the **Manage BLAST Databases** section for more on this (section 12.5).

There are two very important things to note if you wish to take advantage of this tool.

- Many of the databases listed are very large. Please make sure you have space for them. If you are working on a shared system, we recommend you discuss your plans with your system administrator and fellow users.
- Some of the databases listed are dependent on others. This will be listed in the **Dependencies** column of the **Download BLAST Databases** window. This means that while the database your are interested in may seem very small, it may require that you also download a very big database on which it depends.

An example of the second item above is *Swissprot*. To download a database from the NCBI that would allow you to search just Swissprot entries, you need to download the whole *nr* database in addition to the entry for Swissprot.

12.4.3 Create local BLAST databases

In the *CLC Drug Discovery Workbench* you can create a local database that you can use for local BLAST searches. You can specify a location on your computer to save the BLAST database files to. The Workbench will list the BLAST databases found in these locations when you set up a local BLAST search (see section 12.1.3).

DNA, RNA, and protein sequences located in the **Navigation Area** can be used to create BLAST databases from. Any given BLAST database can only include one molecule type. If you wish to use a pre-formatted BLAST database instead, see section 12.4.1.

To create a BLAST database, go to:

Toolbox | **Sequence Analysis** ( ) | **BLAST** ( )| **Create BLAST Database** ( )

This opens the dialog seen in figure 12.16.

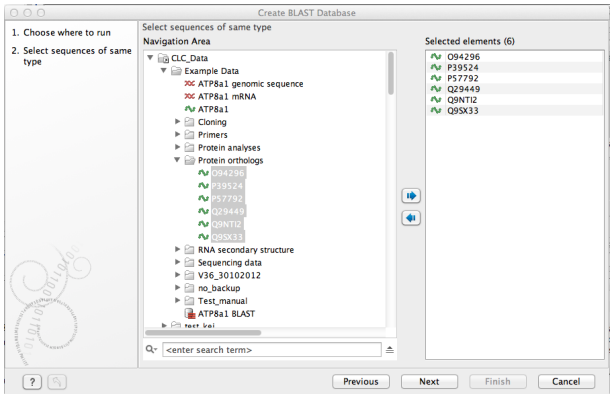


Figure 12.16: *Add sequences for the BLAST database.*

Select sequences or sequence lists you wish to include in your database and click **Next**.

In the next dialog, shown in figure 12.17, you provide the following information:

- **Name.** The name of the BLAST database. This name will be used when running BLAST searches and also as the base file name for the BLAST database files.
- **Description.** You can add more details to describe the contents of the database.

- **Location.** You can select the location to save the BLAST database files to. You can add or change the locations in this list using the **Manage BLAST Databases** tool, see section 12.5.

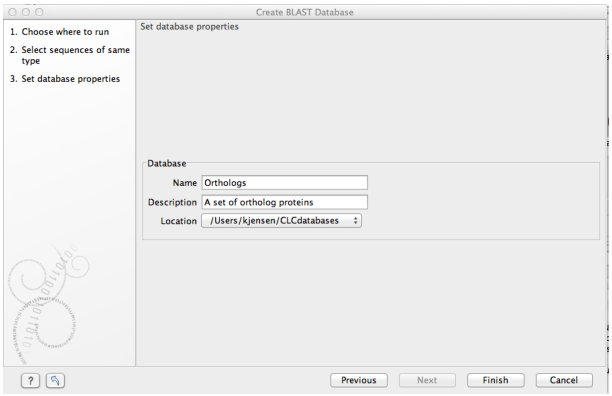


Figure 12.17: *Providing a name and description for the database, and the location to save the files to.*

Click **Finish** to create the BLAST database. Once the process is complete, the new database will be available in the **Manage BLAST Databases** dialog, see section 12.5, and when running local BLAST (see section 12.1.3).

12.5 Manage BLAST databases

The BLAST databases available as targets for running local BLAST searches (see section 12.1.3) can be managed through the Manage BLAST Databases dialog (see figure 12.18):

Toolbox | **Sequence Analysis (** **)** | **BLAST (** **)**| **Manage BLAST Databases (** **)**

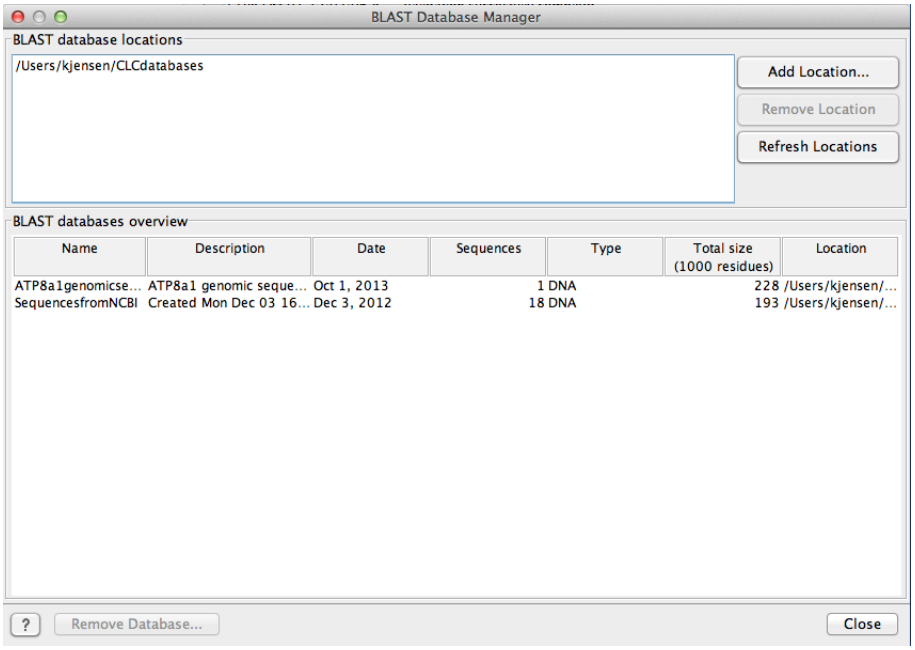


Figure 12.18: *Overview of available BLAST databases.*

At the top of the dialog, there is a list of the **BLAST database locations**. These locations are

folders where the Workbench will look for valid BLAST databases. These can either be created from within the Workbench using the **Create BLAST Database tool**, see section 12.4.3, or they can be pre-formatted BLAST databases.

The list of locations can be modified using the **Add Location** and **Remove Location** buttons. Once the Workbench has scanned the locations, it will keep a cache of the databases (in order to improve performance). If you have added new databases that are not listed, you can press **Refresh Locations** to clear the cache and search the database locations again.

Note: The BLAST database location and all folders in its path should **not** have any spaces in their names on Linux or Mac systems.

By default a BLAST database location will be added under your home area in a folder called CLCdatabases. This folder is scanned recursively, through all subfolders, to look for valid databases. All other folder locations are scanned only at the top level.

Below the list of locations, all the BLAST databases are listed with the following information:

- **Name.** The name of the BLAST database.
- **Description.** Detailed description of the contents of the database.
- **Date.** The date the database was created.
- **Sequences.** The number of sequences in the database.
- **Type.** The type can be either nucleotide (DNA) or protein.
- **Total size (1000 residues).** The number of residues in the database, either bases or amino acid.
- **Location.** The location of the database.

Below the list of BLAST databases, there is a button to **Remove Database**. This option will delete the database files belonging to the database selected.

12.6 Bioinformatics explained: BLAST

BLAST (Basic Local Alignment Search Tool) has become the *defacto* standard in search and alignment tools [Altschul et al., 1990]. The BLAST algorithm is still actively being developed and is one of the most cited papers ever written in this field of biology. Many researchers use BLAST as an initial screening of their sequence data from the laboratory and to get an idea of what they are working on. BLAST is far from being basic as the name indicates; it is a highly advanced algorithm which has become very popular due to availability, speed, and accuracy. In short, a BLAST search identifies homologous sequences by searching one or more databases usually hosted by NCBI (`http://www.ncbi.nlm.nih.gov/`), on the query sequence of interest [McGinnis and Madden, 2004].

BLAST is an open source program and anyone can download and change the program code. This has also given rise to a number of BLAST derivatives; WU-BLAST is probably the most commonly used [Altschul and Gish, 1996].

BLAST is highly scalable and comes in a number of different computer platform configurations which makes usage on both small desktop computers and large computer clusters possible.

12.6.1 Examples of BLAST usage

BLAST can be used for a lot of different purposes. A few of them are mentioned below.

- **Looking for species.** If you are sequencing DNA from unknown species, BLAST may help identify the correct species or homologous species.
- **Looking for domains.** If you BLAST a protein sequence (or a translated nucleotide sequence) BLAST will look for known domains in the query sequence.
- **Looking at phylogeny.** You can use the BLAST web pages to generate a phylogenetic tree of the BLAST result.
- **Mapping DNA to a known chromosome.** If you are sequencing a gene from a known species but have no idea of the chromosome location, BLAST can help you. BLAST will show you the position of the query sequence in relation to the hit sequences.
- **Annotations.** BLAST can also be used to map annotations from one organism to another or look for common genes in two related species.

12.6.2 Searching for homology

Most research projects involving sequencing of either DNA or protein have a requirement for obtaining biological information of the newly sequenced and maybe unknown sequence. If the researchers have no prior information of the sequence and biological content, valuable information can often be obtained using BLAST. The BLAST algorithm will search for homologous sequences in predefined and annotated databases of the users choice.

In an easy and fast way the researcher can gain knowledge of gene or protein function and find evolutionary relations between the newly sequenced DNA and well established data.

After the BLAST search the user will receive a report specifying found homologous sequences and their local alignments to the query sequence.

12.6.3 How does BLAST work?

BLAST identifies homologous sequences using a heuristic method which initially finds short matches between two sequences; thus, the method does not take the entire sequence space into account. After initial match, BLAST attempts to start local alignments from these initial matches. This also means that BLAST does not guarantee the optimal alignment, thus some sequence hits may be missed. In order to find optimal alignments, the Smith-Waterman algorithm should be used (see below). In the following, the BLAST algorithm is described in more detail.

Seeding

When finding a match between a query sequence and a hit sequence, the starting point is the *words* that the two sequences have in common. A word is simply defined as a number of letters. For blastp the default word size is 3 *W=3*. If a query sequence has a QWRTG, the searched words are QWR, WRT, RTG. See figure 12.19 for an illustration of words in a protein sequence.

Figure 12.19: *Generation of exact BLAST words with a word size of W=3.*

During the initial BLAST seeding, the algorithm finds all common words between the query sequence and the hit sequence(s). Only regions with a word hit will be used to build on an alignment.

BLAST will start out by making words for the entire query sequence (see figure 12.19). For each word in the query sequence, a compilation of neighborhood words, which exceed the threshold of *T*, is also generated.

A neighborhood word is a word obtaining a score of at least *T* when comparing, using a selected scoring matrix (see figure 12.20). The default scoring matrix for blastp is BLOSUM62 (for explanation of scoring matrices, see www.clcbio.com/be). The compilation of exact words and neighborhood words is then used to match against the database sequences.

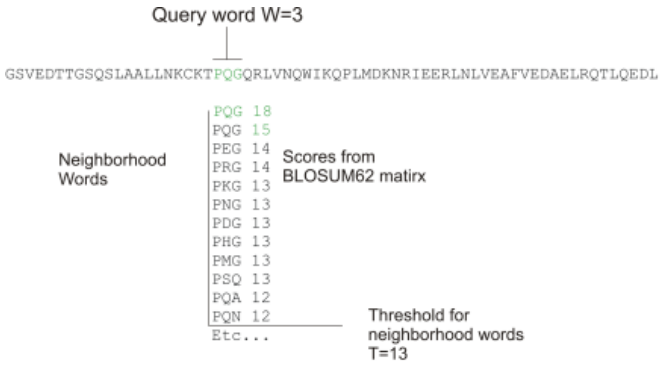


Figure 12.20: *Neighborhood BLAST words based on the BLOSUM62 matrix. Only words where the threshold* T *exceeds 13 are included in the initial seeding.*

After initial finding of words (seeding), the BLAST algorithm will extend the (only 3 residues long) alignment in both directions (see figure 12.21). Each time the alignment is extended, an alignment score is increases/decreased. When the alignment score drops below a predefined threshold, the extension of the alignment stops. This ensures that the alignment is not extended to regions where only very poor alignment between the query and hit sequence is possible. If the obtained alignment receives a score above a certain threshold, it will be included in the final BLAST result.

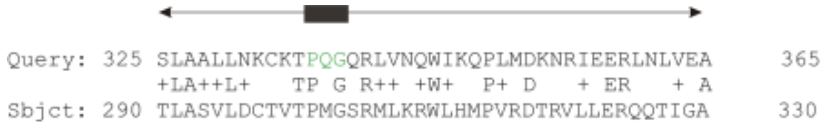


Figure 12.21: *Blast aligning in both directions. The initial word match is marked green.*

By tweaking the word size *W* and the neighborhood word threshold *T*, it is possible to limit the

search space. E.g. by increasing *T*, the number of neighboring words will drop and thus limit the search space as shown in figure 12.22.

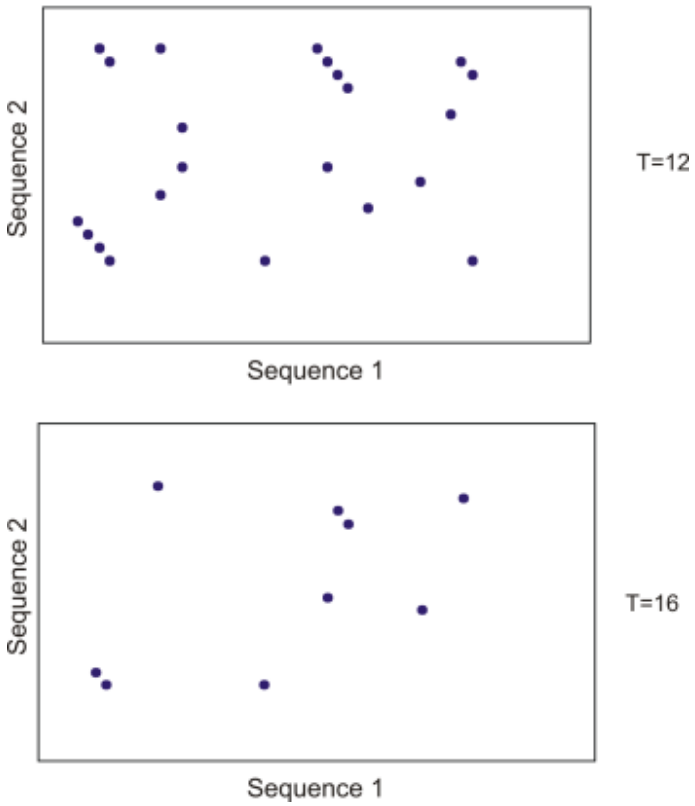


Figure 12.22: *Each dot represents a word match. Increasing the threshold of* T *limits the search space significantly.*

This will increase the speed of BLAST significantly but may result in loss of sensitivity. Increasing the word size *W* will also increase the speed but again with a loss of sensitivity.

12.6.4 Which BLAST program should I use?

Depending on the nature of the sequence it is possible to use different BLAST programs for the database search. There are five versions of the BLAST program, blastn, blastp, blastx, tblastn, tblastx:

Option	Query Type	DB Type	Comparison	Note
blastn	Nucleotide	Nucleotide	Nucleotide-Nucleotide	
blastp	Protein	Protein	Protein-Protein	
tblastn	Protein	Nucleotide	Protein-Protein	The database is translated into protein
blastx	Nucleotide	Protein	Protein-Protein	The queries are translated into protein
tblastx	Nucleotide	Nucleotide	Protein-Protein	The queries and database are translated into protein

The most commonly used method is to BLAST a nucleotide sequence against a nucleotide database (blastn) or a protein sequence against a protein database (blastp). But often another BLAST program will produce more interesting hits. E.g. if a nucleotide sequence is translated

before the search, it is more likely to find better and more accurate hits than just a blastn search. One of the reasons for this is that protein sequences are evolutionarily more conserved than nucleotide sequences. Another good reason for translating the query sequence before the search is that you get protein hits which are likely to be annotated. Thus you can directly see the protein function of the sequenced gene.

12.6.5 Which BLAST options should I change?

The NCBI BLAST web pages and the BLAST command line tool offer a number of different options which can be changed in order to obtain the best possible result. Changing these parameters can have a great impact on the search result. It is not the scope of this document to comment on all of the options available but merely the options which can be changed with a direct impact on the search result.

The E-value

The *expect value* (E-value) can be changed in order to limit the number of hits to the most significant ones. The lower the E-value, the better the hit. The E-value is dependent on the length of the query sequence and the size of the database. For example, an alignment obtaining an E-value of 0.05 means that there is a 5 in 100 chance of occurring by chance alone.

E-values are very dependent on the query sequence length and the database size. Short identical sequence may have a high E-value and may be regarded as "false positive" hits. This is often seen if one searches for short primer regions, small domain regions etc. The default threshold for the E-value on the BLAST web page is 10. Increasing this value will most likely generate more hits. Below are some rules of thumb which can be used as a guide but should be considered with common sense.

- **E-value < 10e-100** Identical sequences. You will get long alignments across the entire query and hit sequence.
- **10e-100 < E-value < 10e-50** Almost identical sequences. A long stretch of the query protein is matched to the database.
- **10e-50 < E-value < 10e-10** Closely related sequences, could be a domain match or similar.
- **10e-10 < E-value < 1** Could be a true homologue but it is a gray area.
- **E-value > 1** Proteins are most likely not related
- **E-value > 10** Hits are most likely junk unless the query sequence is very short.

Gap costs

For blastp it is possible to specify gap cost for the chosen substitution matrix. There is only a limited number of options for these parameters. The *open gap cost* is the price of introducing gaps in the alignment, and *extension gap cost* is the price of every extension past the initial opening gap. Increasing the gap costs will result in alignments with fewer gaps.

Filters

It is possible to set different filter options before running the BLAST search. Low-complexity regions have a very simple composition compared to the rest of the sequence and may result in problems during the BLAST search [Wootton and Federhen, 1993]. A low complexity region of a protein can for example look like this 'fftfffllsss', which in this case is a region as part of a signal peptide. In the output of the BLAST search, low-complexity regions will be marked in lowercase gray characters (default setting). The low complexity region cannot be thought of as a significant match; thus, disabling the low complexity filter is likely to generate more hits to sequences which are not truly related.

Word size

Change of the word size has a great impact on the seeded sequence space as described above. But one can change the word size to find sequence matches which would otherwise not be found using the default parameters. For instance the word size can be decreased when searching for primers or short nucleotides. For blastn a suitable setting would be to decrease the default word size of 11 to 7, increase the E-value significantly (1000) and turn off the complexity filtering.

For blastp a similar approach can be used. Decrease the word size to 2, increase the E-value and use a more stringent substitution matrix, e.g. a PAM30 matrix.

Fortunately, the optimal search options for finding short, nearly exact matches can already be found on the BLAST web pages http://www.ncbi.nlm.nih.gov/BLAST/.

Substitution matrix

For protein BLAST searches, a default substitution matrix is provided. If you are looking at distantly related proteins, you should either choose a high-numbered PAM matrix or a low-numbered BLOSUM matrix. See *Bioinformatics Explained* on scoring matrices on http://www.clcbio.com/be/. The default scoring matrix for blastp is BLOSUM62.

12.6.6 Explanation of the BLAST output

The BLAST output comes in different flavors. On the NCBI web page the default output is html, and the following description will use the html output as example. Ordinary text and xml output for easy computational parsing is also available.

The default layout of the NCBI BLAST result is a graphical representation of the hits found, a table of sequence identifiers of the hits together with scoring information, and alignments of the query sequence and the hits.

The graphical output (shown in figure 12.23) gives a quick overview of the query sequence and the resulting hit sequences. The hits are colored according to the obtained alignment scores.

The table view (shown in figure 12.24) provides more detailed information on each hit and furthermore acts as a hyperlink to the corresponding sequence in GenBank.

In the alignment view one can manually inspect the individual alignments generated by the BLAST algorithm. This is particularly useful for detailed inspection of the sequence hit found(sbjct) and the corresponding alignment. In the alignment view, all scores are described for each alignment,

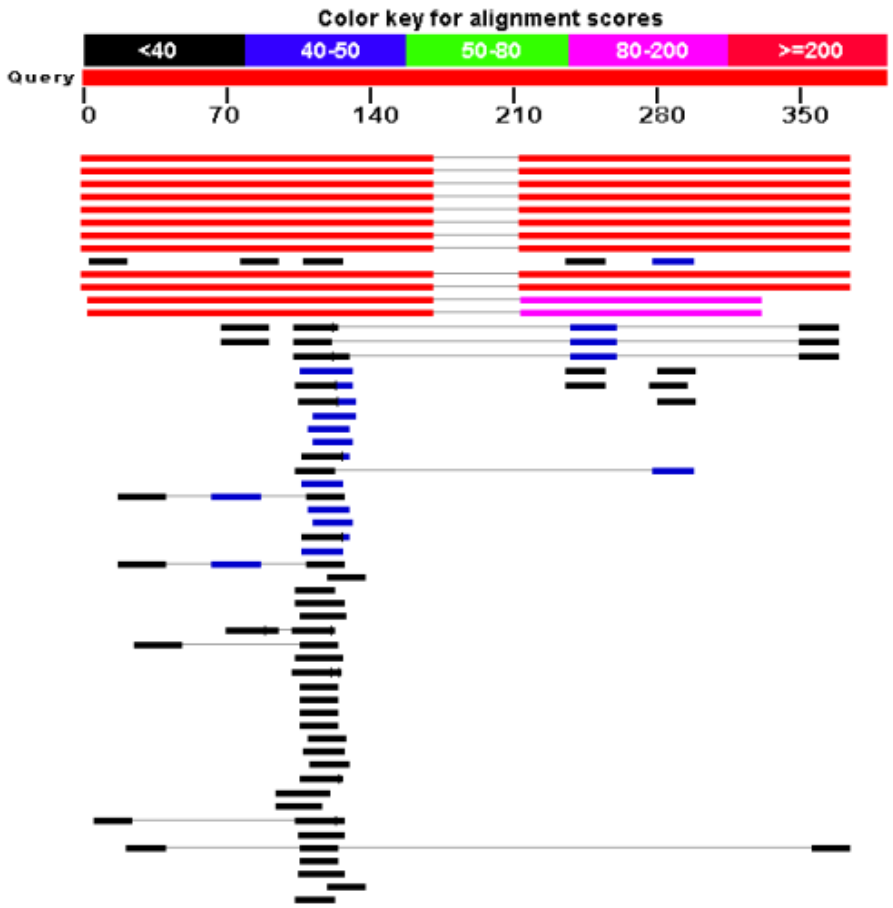


Figure 12.23: *BLAST graphical view. A simple graphical overview of the hits found aligned to the query sequence. The alignments are color coded ranging from black to red as indicated in the color label at the top.*

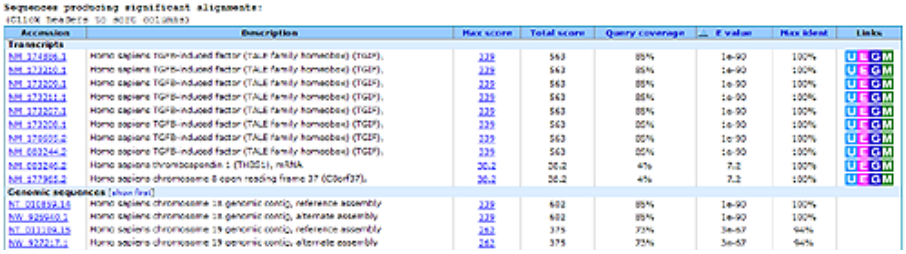


Figure 12.24: *BLAST table view. A table view with one row per hit, showing the accession number and description field from the sequence file together with BLAST output scores.*

and the start and stop positions for the query and hit sequence are listed. The strand and orientation for query sequence and hits are also found here.

In most cases, the table view of the results will be easier to interpret than tens of sequence alignments.

Figure 12.25: *Alignment view of BLAST results. Individual alignments are represented together with BLAST scores and more.*

12.6.7 I want to BLAST against my own sequence database, is this possible?

It is possible to download the entire BLAST program package and use it on your own computer, institution computer cluster or similar. This is preferred if you want to search in proprietary sequences or sequences unavailable in the public databases stored at NCBI. The downloadable BLAST package can either be installed as a web-based tool or as a command line tool. It is available for a wide range of different operating systems.

The BLAST package can be downloaded free of charge from the following location `http://www.ncbi.nlm.nih.gov/BLAST/download.shtml`

Pre-formatted databases are available from a dedicated BLAST ftp site `ftp://ftp.ncbi.nlm.nih.gov/blast/db/`. Moreover, it is possible to download programs/scripts from the same site enabling automatic download of changed BLAST databases. Thus it is possible to schedule a nightly update of changed databases and have the updated BLAST database stored locally or on a shared network drive at all times. Most BLAST databases on the NCBI site are updated on a daily basis to include all recent sequence submissions to GenBank.

A few commercial software packages are available for searching your own data. The advantage of using a commercial program is obvious when BLAST is integrated with the existing tools of these programs. Furthermore, they let you perform BLAST searches and retain annotations on the query sequence (see figure 12.26). It is also much easier to batch download a selection of hit sequences for further inspection.

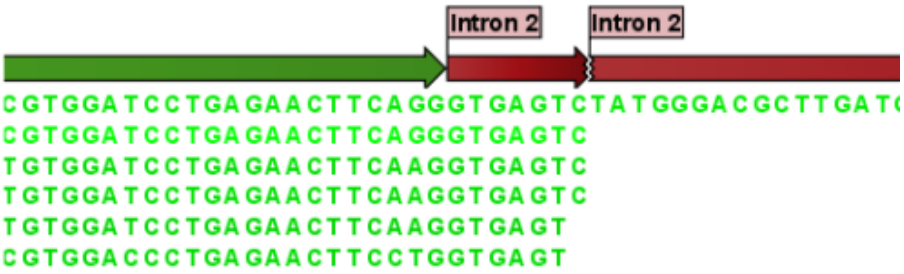


Figure 12.26: *Snippet of alignment view of BLAST results from CLC Main Workbench. Individual alignments are represented directly in a graphical view. The top sequence is the query sequence and is shown with a selection of annotations.*

12.6.8 What you cannot get out of BLAST

Don't expect BLAST to produce the best available alignment. BLAST is a heuristic method which does not guarantee the best results, and therefore you cannot rely on BLAST if you wish to find *all* the hits in the database.

Instead, use the Smith-Waterman algorithm for obtaining the best possible local alignments [Smith and Waterman, 1981].

BLAST only makes local alignments. This means that a great but short hit in another sequence may not at all be related to the query sequence even though the sequences align well in a small region. It may be a domain or similar.

It is always a good idea to be cautious of the material in the database. For instance, the sequences may be wrongly annotated; hypothetical proteins are often simple translations of a found ORF on a sequenced nucleotide sequence and may not represent a true protein.

Don't expect to see the best result using the default settings. As described above, the settings should be adjusted according to the what kind of query sequence is used, and what kind of results you want. It is a good idea to perform the same BLAST search with different settings to get an idea of how they work. There is not a final answer on how to adjust the settings for your particular sequence.

12.6.9 Other useful resources

The BLAST web page hosted at NCBI
`http://www.ncbi.nlm.nih.gov/BLAST`

Download pages for the BLAST programs
`http://www.ncbi.nlm.nih.gov/BLAST/download.shtml`

Download pages for pre-formatted BLAST databases
`ftp://ftp.ncbi.nlm.nih.gov/blast/db/`

O'Reilly book on BLAST
`http://www.oreilly.com/catalog/blast/`

Explanation of scoring/substitution matrices and more
`http://www.clcbio.com/be/`

Chapter 13

Sequence analyses

Contents

CLC Drug Discovery Workbench offers different kinds of sequence analyses The analyses are described in this chapter.

13.1 Find and Model Structure

This tool is used to find suitable protein structures for representing a given protein sequence. From the resulting table, a structure model (homology model) of the sequence can be created by one click, using one of the found protein structures as template.

To run the *Find and Model Structure* tool:

Toolbox | **Sequence Analysis** () | **Find and Model Structure** ()

Note: Before running the tool, a protein structure sequence database must be downloaded and installed using the 'Download Find Structure Database' tool (see section 13.2).

In the tool wizard step 1, select the amino acid sequence to use as query from the Navigation Area.

In step 2, specify if the output table should be opened or saved.

The Find and Model Structure tool carries out the following steps, to find and rank available structures representing the query sequence:

Input: Query protein sequence

1. BLAST against protein structure sequence database
2. Filter away low quality hits
3. Rank the available structures

Output: Table listing available structures

In the output table (figure 13.1), the column named "Available Structures" contains links that will invoke a menu with the options to either create a structure model of the query sequence or just download the structure. This is further described in section 13.1.1. The remaining columns contain additional information originating from the PDB file or from the BLAST search.

The three steps carried out by the *Find and Model Structure* tool are described in short below.

BLAST against protein structure sequence database

A local BLAST search is carried out for the query sequence against the protein structure sequence database (see section 13.2).

BLAST hits with E-value > 0.0001 are rejected and a maximum of 2500 BLAST hits are retrieved. Read more about BLAST in section 12.6.

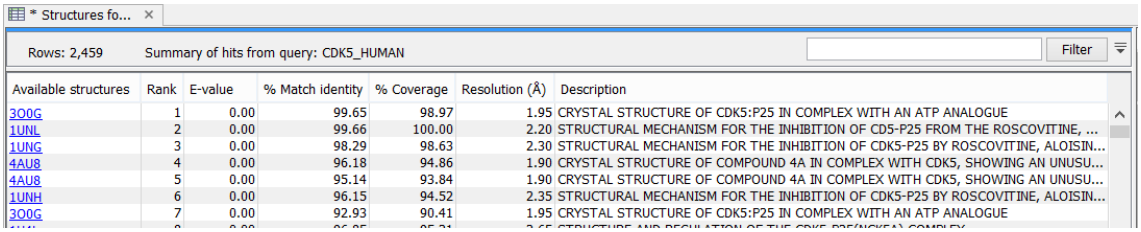


Figure 13.1: *Table output from Find and Model Structure.*

Filter away low quality hits

From the list of BLAST hits, entries are rejected based on the following rules:

- PDB structures with a resolution lower than 4 Å are removed since they cannot be expected to represent a trustworthy atomistic model.
- BLAST hits with an identity to the query sequence lower than 20 % are removed since they most likely would result in inaccurate models.

Rank the available structures

For the resulting list of available structures, each structure is scored based on its homology to the query sequence, and the quality of the structure itself. The *Template quality score* is used to rank the structures in the table, and the rank of each structure is shown in the "Rank" column (see figure 13.1). Read more about the *Template quality score* in section 13.1.2.

13.1.1 Create structure model

Clicking on a link in the "Available structures" column will show a menu with three options:

- Download and Open
- Download and Create Model
- Help

The "Download and Open" option will do the following:

1. **Download and import** the PDB file containing the structure.
2. **Create an alignment** between the query and structure sequences.
3. **Open a 3D view** (Molecule Project) with the molecules from the PDB file and open the created sequence alignment. The sequence originating from the structure will be linked to the structure in the 3D view, so that selections on the sequence will show up on the structure (see section 9.4).

The "Download and Create Model" option will do the following:

1. **Download and import** the PDB file containing the structure.

2. **Generate a biomolecule** involving the protein chain to be modeled. Biomolecule information available in the template PDB file is used (see section 9.6). If several biomolecules involving the chain are available, the first one is applied.

3. **Create an alignment** between the query and structure sequences.

4. **Create a model structure** by mapping the query sequence onto the structure based on the sequence alignment (see section 13.1.1). If multiple copies of the template protein chain have been made to generate a biomolecule, all copies are modeled at the same time.

5. **Open a 3D view** (a Molecule Project) with the structure model shown in both backbone and wireframe representation. The model is colored by temperature (see figure 13.2), to indicate local model uncertainty (see section 13.1.1). Other molecules from the template PDB file are shown in orange or yellow coloring. The created sequence alignment is also opened and linked with the 3D views so that selections on the model sequence will show up on the model structure (see section 9.4).

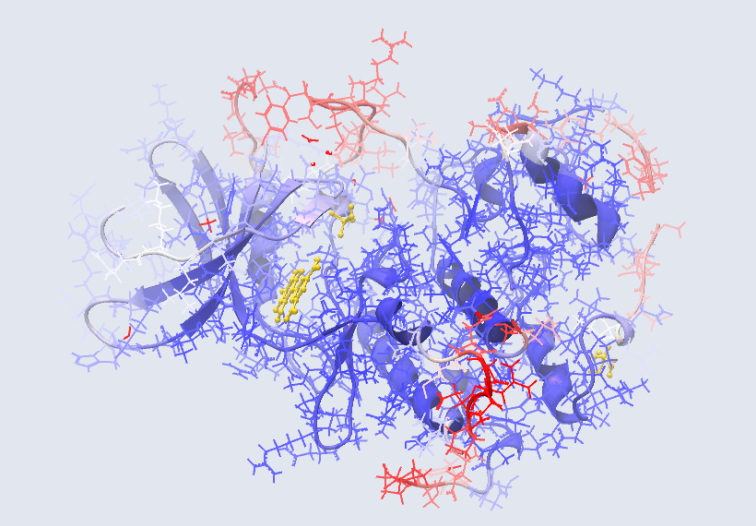


Figure 13.2: *Structure Model of CDK5_HUMAN. The atoms and backbone are colored by temperature, showing uncertain structure in red and well defined structure in blue.*

The template structure is also available from the Proteins category in the Project Tree, but hidden in the initial view. The initial view settings are saved on the Molecule Project as "Initial visualization", and can always be reapplied from the View Settings menu ( ) found in the bottom right corner of the Molecule Project (see section 4.5).

Protein coloring to visualize local structural uncertainties

The default coloring scheme for modeled structures in *CLC Drug Discovery Workbench* is "Color by Temperature". This coloring indicates the uncertainty or disorder of each atom position in the structure.

For crystal structures, the temperature factor (also called the B-factor) is given in the PDB file as a measure of the uncertainty or disorder of each atom position. The temperature factor has the unit $Å^2$, and is typically in the range [0, 100].

The temperature color scale ranges from blue (0) over white (50) to red (100) (see section 9.2.1).

For structure models created in *CLC Drug Discovery Workbench*, the temperature factor assigned to each atom combines three sources of positional uncertainty:

- **PDB Temp.** The atom position uncertainty for the template structure, represented by the temperature factor of the backbone atoms in the template structure.
- **P(alignment)** The probability that the alignment of a residue in the query sequence to a particular position on the structure is correct.
- **Clash?** It is evaluated if atoms in the structure model seem to clash, thereby indicating a problem with the model.

The three aspects are combined to give a temperature value between zero and 100, as illustrated in figure 13.3 and 13.4.

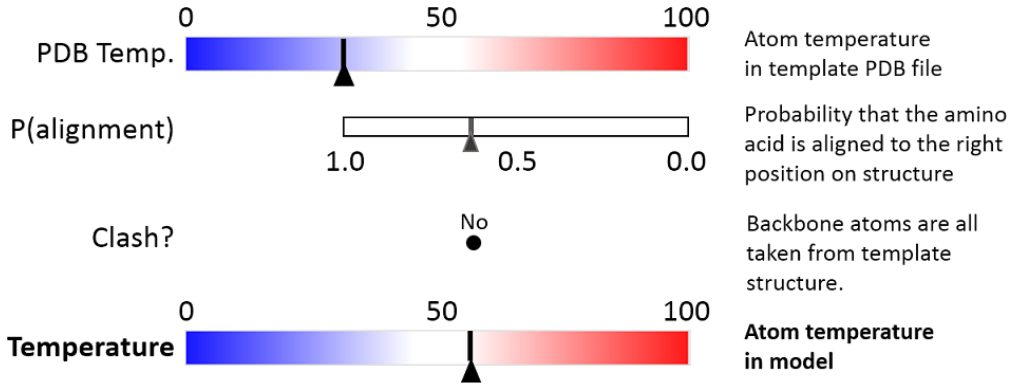


Figure 13.3: *Evaluation of temperature color for backbone atoms in structure models.*

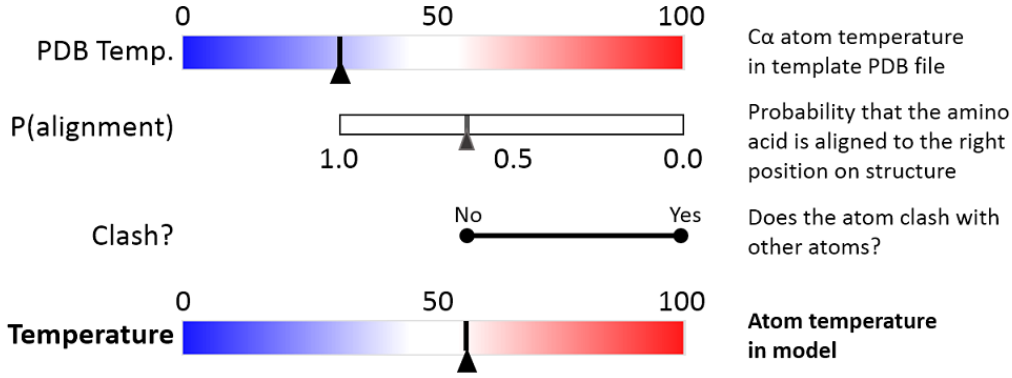


Figure 13.4: *Evaluation of temperature color for side chain atoms in structure models.*

When holding the mouse over an atom, the Property Viewer in the Side Panel will show various information about the atom. For atoms in structure models, the contributions to the assigned temperature are listed as seen in figure 13.5.

Note: For NMR structures, the temperature factor is set to zero in the PDB file, and the "Color by Temperature" will therefore suggest that the structure is more well determined than is actually the case.

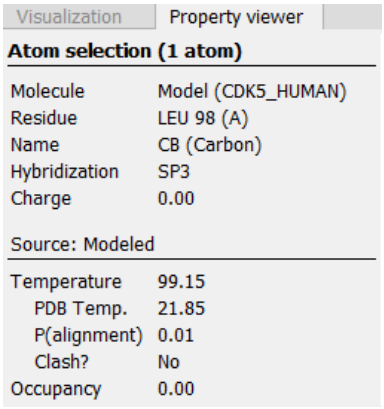


Figure 13.5: *Information displayed in the Side Panel Property viewer for a modeled atom.*

P(alignment)

Alignment error is one of the largest causes of model inaccuracy, particularly when the model is built from a template sharing low sequence identity (e.g. lower than 60%). Misaligning a single amino acid by one position will cause a ca. 3.5 Å shift of its atoms from their true positions.

The estimate of the probability that two amino acids are correctly aligned, P(alignment), is obtained by averaging over all the possible alignments between two sequences, similar to [Knudsen and Miyamoto, 2003].

This allows local alignment uncertainty to be detected even in similar sequences. For example the position of the D in this alignment:

Template	GGACDAEDRSTRSTACE---GG
Target	GGACD---RSTRSTACEKLMGG

is uncertain, because an alternative alignment is as likely:

Template	GGACDAEDRSTRSTACE---GG
Target	GGAC---DRSTRSTACEKLMGG

For manual mutations on a crystal structure (section 9.10), P(alignment) is set to 1.0. For manual mutations on a model created with the Find and Model Structure tool, the P(alignment) value is carried over from the model.

Clash?

Clashes are evaluated separately for each atom in a side chain. The scoring function used to evaluate protein-ligand interactions for docking and ligand optimization in *CLC Drug Discovery Workbench* is also used to evaluate the interaction between a given side chain atom and its surroundings.

Read more about the scoring function in section 9.12.4.

If the total score for one atom is higher than 3.0 (see figure 9.58), then this atom is considered to clash, and will be assigned a temperature of 100.

Note: For structure models created with the Find and Model Structure tool, clashes within the modeled protein chain as well as with all other molecules in the PDB file (except water) are considered.

For manual mutations introduced to a protein structure from an imported PDB file (see section 9.10), clashes are only considered within the mutated protein chain.

When a manual mutation is introduced to a structure model, "Clash?" will be re-evaluated for all residues, to only consider interactions within the mutated protein chain, thus ignoring clashes with other molecules in the Molecule Project.

13.1.2 Method details

Evaluating the rank of available structures

A *template quality score* is calculated for the available structures found for the query sequence. The purpose of the score is to rank structures considering both their quality and their homology to the query sequence.

The five descriptors contributing to the score can be found in the columns of the output table (see figure 13.1):

- E-value
- % Match identity
- % Coverage
- Resolution (of crystal structure)
- Free R-value (R_{free} of crystal structure)

Each of the five descriptors are scaled to [0,1], based on the linear functions seen in figure 13.7. The five scaled descriptors are combined into the *template quality score*, weighting them to emphasize homology over structure qualities.

$$\text{Template quality score} = 3 \cdot S_{\text{E-value}} + 3 \cdot S_{\text{Identity}} + 1.5 \cdot S_{\text{Coverage}} + S_{\text{Resolution}} + 0.5 \cdot S_{\text{Rfree}}$$

E-value is a measure of the quality of the match returned from the BLAST search. You can read more about BLAST and E-values in section 12.6.

% Match identity is the identity between the query sequence and the BLAST hit in the matched region. It is evaluated as

$$\text{\% Match identity} = 100\% \cdot (\text{Identity in BLAST alignment})/L_{\text{B}}$$

where L_{B} is the length of the BLAST alignment of the matched region, as indicated in figure 13.6, and "Identity in BLAST alignment" is the number of identical positions in the matched region.

% Coverage indicates how much of the query sequence has been covered by a given BLAST hit (see figure 13.6). It is evaluated as

$$\% \text{ Coverage} = 100\% \cdot (L_B - L_G)/L_Q$$

where L_G is the total length of gaps in the BLAST alignment and L_Q is the length of the query sequence.

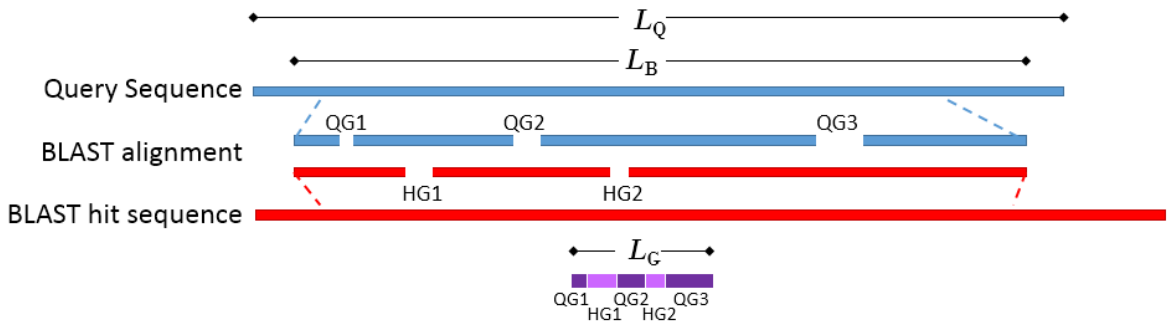


Figure 13.6: *Schematic of a query sequence matched to a BLAST hit. L_Q is the length of the query sequence, L_B is the length of the BLAST alignment of the matched region, QG1-3 are gaps in the matched region of the query sequence, HG1-2 are gaps in the matched region of the BLAST hit sequence, L_G is the total length of gaps in the BLAST alignment.*

The **resolution** of a crystal structure is related to the size of structural features that can be resolved from the raw experimental data.

R_{free} is used to assess possible overmodeling of the experimental data.

Resolution and R_{free} are only given for crystal structures. NMR structures will therefore usually be ranked lower than crystal structures. Likewise, structures where R_{free} has not been given will tend to receive a lower rank. This often coincides with structures of older date.

How a model structure is created

A structure model is created by mapping the query sequence onto the template structure based on a sequence alignment (see figure 13.8):

- For identical amino acids (example 1 in figure 13.8) => Copy atom positions from the PDB file. If the side chain is missing atoms in the PDB file, the side chain is rebuilt (section 9.10.6).
- For amino acid changes (example 2 in figure 13.8) => Copy backbone atom positions from the PDB file. Model side chain atom positions to match the query sequence (section 9.10.6).
- For amino acids in the query sequence not aligned to a position on the template structure (example 3 in figure 13.8) => No atoms are modeled. The model backbone will have a gap at this position and a "Structure modeling" issue is raised (see section 6.2.9).
- For amino acids on the template structure, not aligned to the query sequence (example 4 in figure 13.8) => The residues are deleted from the structure and a "Structure modeling" issue is raised (see section 6.2.9).

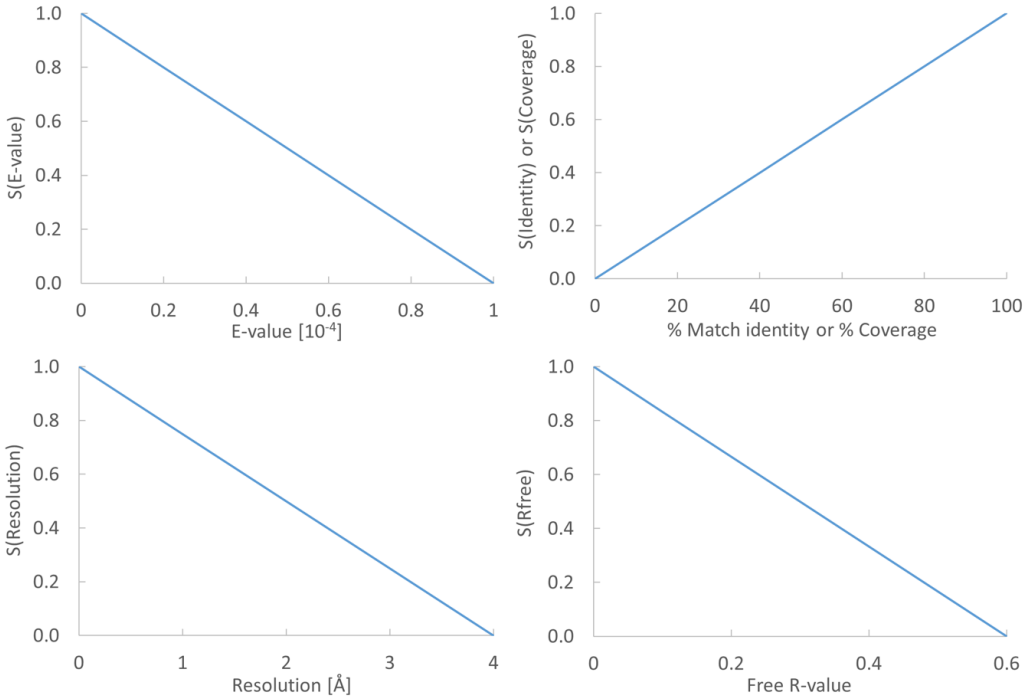


Figure 13.7: *From the E-value, % Match identity, % Coverage, Resolution, and Free R-value, the contributions to the "Template quality score" are determined from the linear functions shown in the graphs.*

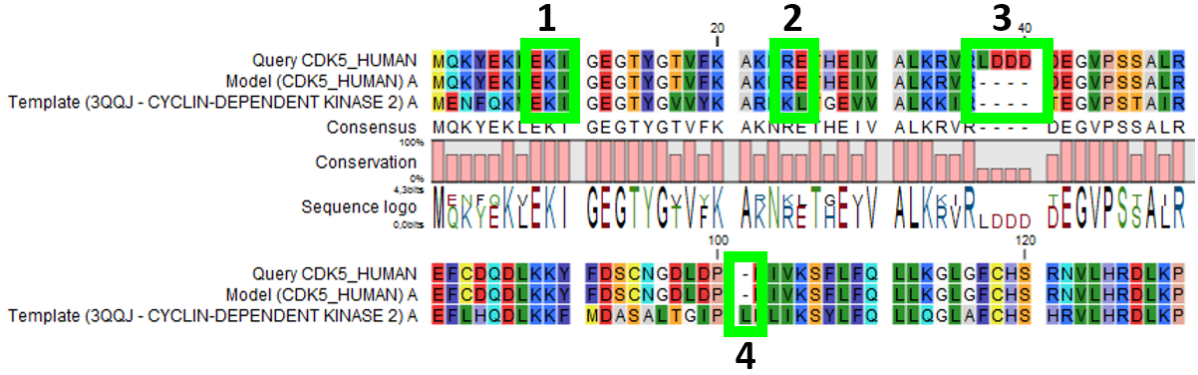


Figure 13.8: *Sequence alignment mapping query sequence (Query CDK5_HUMAN) to the structure with sequence "Template(3QQJ - CYCLIN-DEPENDENT KINASE 2)", producing a structure with sequence "Model(CDK5_HUMAN)". Examples are highlighted: 1. Identical amino acids, 2. Amino acid changes, 3. Amino acids in query sequence not aligned to a position on the template structure, and 4. Amino acids on the template structure, not aligned to query sequence.*

13.2 Download Find Structure Database

This tool downloads the Find Structure Database from a public accessible HTTP location hosted by QIAGEN Aarhus.

The database contains a curated set of sequences with known 3D structures, which are obtained from the Protein Data Bank (`http://www.wwpdb.org`) [Berman et al., 2003]. The information stored in the database (e.g. protein sequence, X-ray resolution) is used to identify suitable structural templates when using the **Find and Model Structure** tool.

To download the database, select:

Toolbox | **Sequence Analysis** ( ) | **Download Find Structure Database** ( )

If you are connected to a server, you will first be asked about whether you want to download the data locally or on a server. In the next wizard step you are asked to select the download location (see figure 13.9).

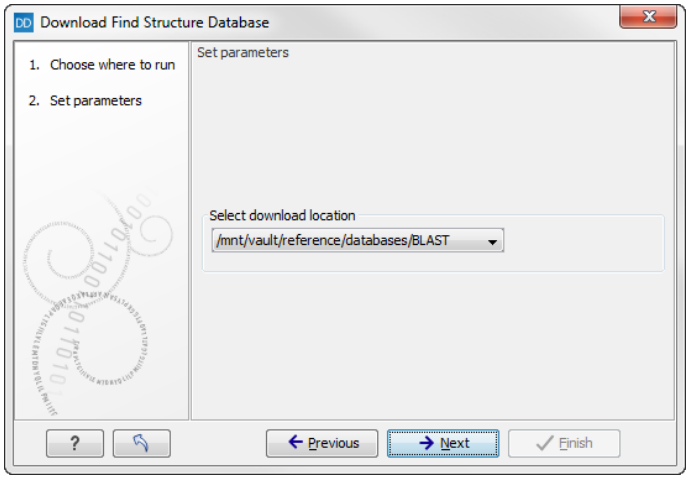


Figure 13.9: *Select the download location.*

The downloaded database will be installed in the same location as local BLAST databases (e.g. <username>/CLCdatabases) or at a server location if the tool was executed on a CLC Server. From the wizard it is possible to select alternative locations if more than one location is available.

When new databases are released, a new version of the database can be downloaded by invoking the tool again (the existing database will be replaced).

If needed, the **Manage BLAST Databases** tool can be used to inspect or delete the database (the database is listed with the name 'ProteinStructureSequences'). You can find the tool here:

Sequence Analysis ( ) | **BLAST ( )**| **Manage BLAST Databases** ( )

13.3 Extract sequences

This tool allows the extraction of sequences from other types of data in the Workbench, such as sequence lists or alignments. The data types you can extract sequences from are:

- Alignments ( )
- BLAST result ( )
- BLAST overview tables ( )
- sequence lists ( )

Note! When the Extract Sequences tool is run via the Workbench toolbox on an entire file of one of the above types, **all** sequences are extracted from the data used as input. If only a **subset** of the sequences is desired, for example, the reads from just a small area of a mapping, or the

sequences for only a few blast results, then a data set containing just this subsection or subset should be created and the Extract Sequences tool should be run on that.

For extracting a subset of a sequence list, you can highlight the sequences of interest in the table view of the sequence list, right click on the selection and launch the Extract Sequences tool.

The Extract Sequences tool can be launched via the Toolbox menu, by going to:

Toolbox | **Sequence Analysis** ( )| **Extract Sequences** ( )

Alternatively, on all the data types listed above except sequence lists, the option to run this tool appears by right clicking in the relevant area; a row in a table or in the read area of mapping data. An example is shown in figure 13.10.

Please note that when extracting sequences from BLAST results, the sequence hits are extracted, not the original query sequence or a consensus sequence.

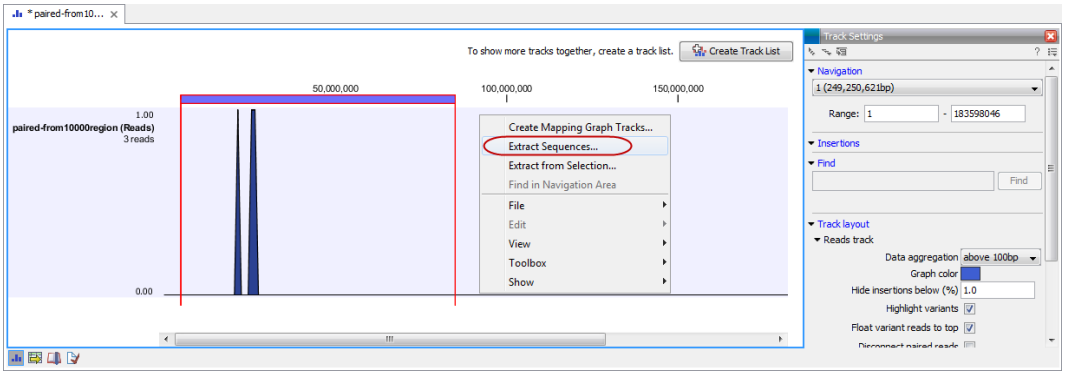


Figure 13.10: *Right click somewhere in the reads track area and select "Extract Sequences".*

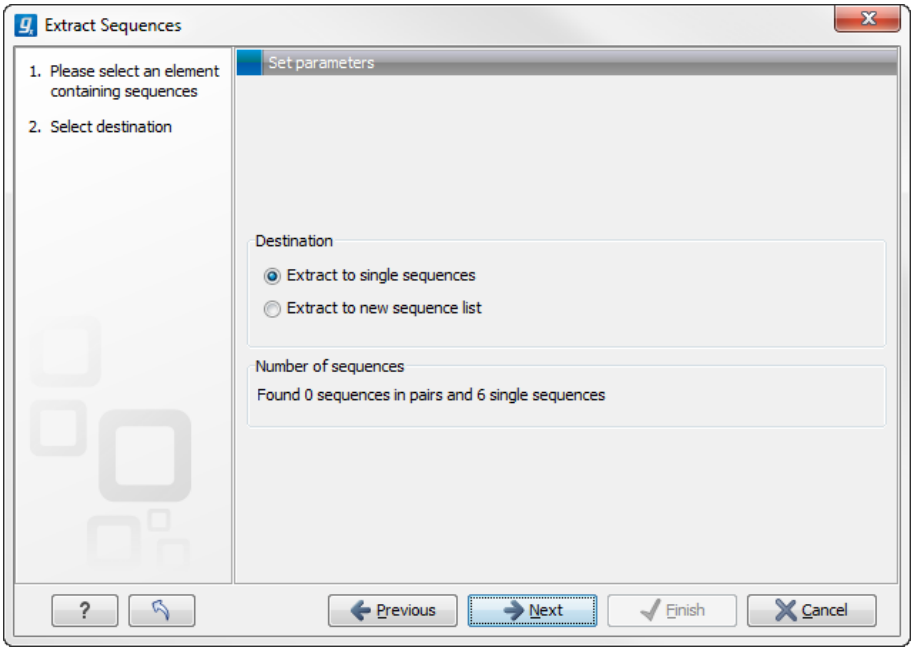


Figure 13.11: *Choosing whether the extracted sequences should be placed in a new list or as single sequences.*

The dialog allows you to select the **Destination**. Here you can choose whether the extracted sequences should be extracted as single sequences or placed in a new sequence list. For most data types, it will make most sense to choose to extract the sequences into a sequence list. The exception to this is when working with a sequence list, where choosing to extract to a sequence list would create a copy of the same sequence list. In this case, the other option would generally be chosen. This would then result in the generation of individual sequence objects for each sequence in the sequence list.

Below these options, in the dialog, you can see the number of sequences that will be extracted.

13.4 Dot plots

Dot plots provide a powerful visual comparison of two sequences. Dot plots can also be used to compare regions of similarity within a sequence. This chapter first describes how to create and second how to

This section describes how to

adjust the view of the plot.

13.4.1 Create dot plots

A dot plot is a simple, yet intuitive way of comparing two sequences, either DNA or protein, and is probably the oldest way of comparing two sequences [Maizel and Lenk, 1981]. A dot plot is a 2 dimensional matrix where each axis of the plot represents one sequence. By sliding a fixed size window over the sequences and making a sequence match by a dot in the matrix, a diagonal line will emerge if two identical (or very homologous) sequences are plotted against each other. Dot plots can also be used to visually inspect sequences for direct or inverted repeats or regions with low sequence complexity. Various smoothing algorithms can be applied to the dot plot calculation to avoid noisy background of the plot. Moreover, various substitution matrices can be applied in order to take the evolutionary distance of the two sequences into account.

To create a dot plot, go to:

Toolbox | **Sequence Analysis (**  **)**| **Create Dot Plot (**  **)**

This opens the dialog shown in figure 13.12.

If a sequence was selected before choosing the **Toolbox** action, this sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove elements from the selected elements. Click **Next** to adjust dot plot parameters. Clicking **Next** opens the dialog shown in figure 13.13.

Note! Calculating dot plots takes up a considerable amount of memory in the computer. Therefore, you will see a warning message if the sum of the number of nucleotides/amino acids in the sequences is higher than 8000. If you insist on calculating a dot plot with more residues the Workbench may shut down, but still allowing you to save your work first. However, this depends on your computer's memory configuration.

Adjust dot plot parameters

There are two parameters for calculating the dot plot:

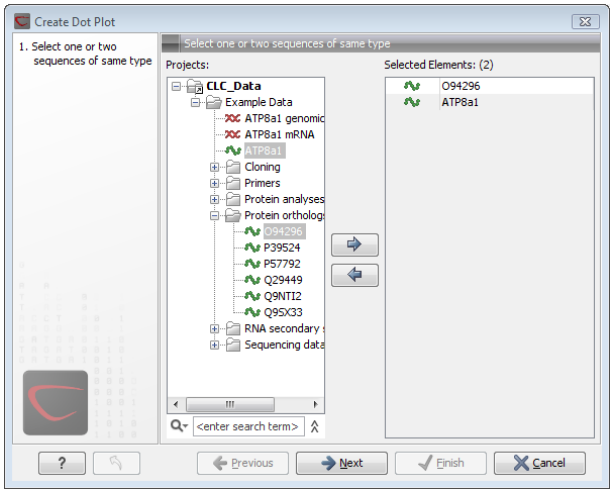


Figure 13.12: *Selecting sequences for the dot plot.*

- **Distance correction (only valid for protein sequences)** In order to treat evolutionary transitions of amino acids, a distance correction measure can be used when calculating the dot plot. These distance correction matrices (substitution matrices) take into account the likeliness of one amino acid changing to another.
- **Window size** A residue by residue comparison (window size = 1) would undoubtedly result in a very noisy background due to a lot of similarities between the two sequences of interest. For DNA sequences the background noise will be even more dominant as a match between only four nucleotide is very likely to happen. Moreover, a residue by residue comparison (window size = 1) can be very time consuming and computationally demanding. Increasing the window size will make the dot plot more 'smooth'.

Click **Next** if you wish to adjust how to handle the results (see section 7.2). If not, click **Finish**.

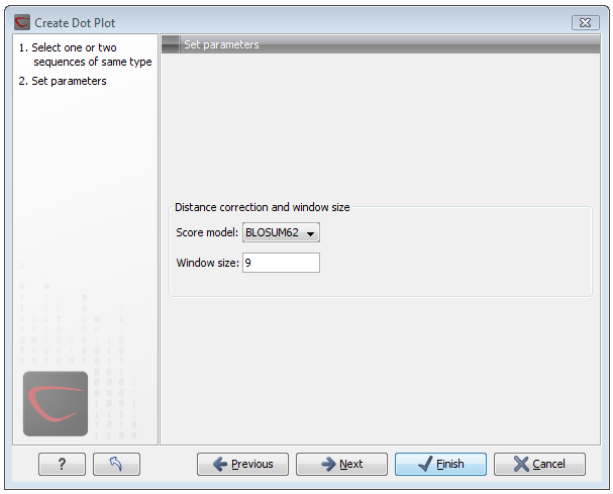


Figure 13.13: *Setting the dot plot parameters.*

13.4.2 View dot plots

A view of a dot plot can be seen in figure 13.14. You can select **Zoom in** ( ) in the Toolbar and click the dot plot to zoom in to see the details of particular areas.

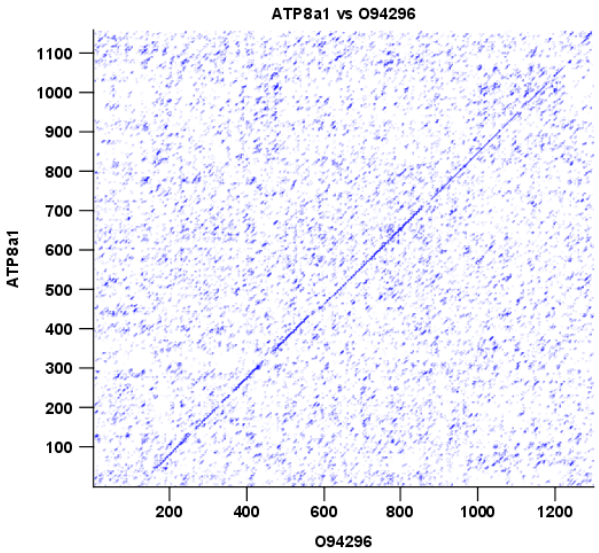


Figure 13.14: *A view is opened showing the dot plot.*

The **Side Panel** to the right let you specify the dot plot preferences. The gradient color box can be adjusted to get the appropriate result by dragging the small pointers at the top of the box. Moving the slider from the right to the left lowers the thresholds which can be directly seen in the dot plot, where more diagonal lines will emerge. You can also choose another color gradient by clicking on the gradient box and choose from the list.

Adjusting the sliders above the gradient box is also practical, when producing an output for printing. (Too much background color might not be desirable). By crossing one slider over the other (the two sliders change side) the colors are inverted, allowing for a white background. (If you choose a color gradient, which includes white). Se figure 13.14.

13.4.3 Bioinformatics explained: Dot plots

Realization of dot plots

Dot plots are two-dimensional plots where the x-axis and y-axis each represents a sequence and the plot itself shows a comparison of these two sequences by a calculated score for each position of the sequence. If a window of fixed size on one sequence (one axis) match to the other sequence a dot is drawn at the plot. Dot plots are one of the oldest methods for comparing two sequences [Maizel and Lenk, 1981].

The scores that are drawn on the plot are affected by several issues.

- Scoring matrix for distance correction.
  Scoring matrices (BLOSUM and PAM) contain substitution scores for every combination of two amino acids. Thus, these matrices can only be used for dot plots of protein sequences.
- Window size

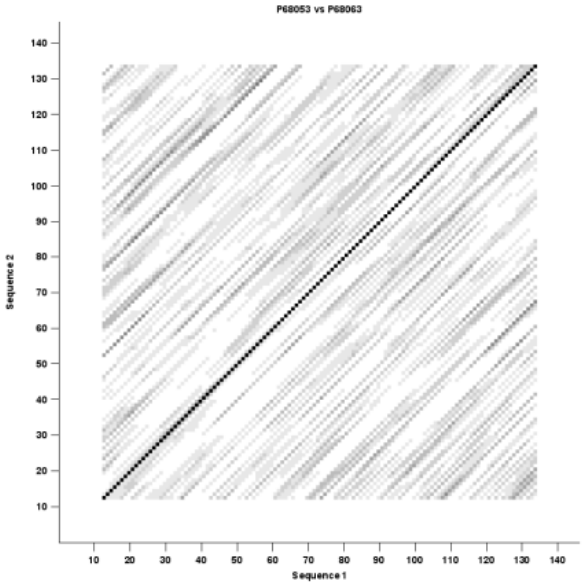


Figure 13.15: *Dot plot with inverted colors, practical for printing.*

The single residue comparison (bit by bit comparison(window size = 1)) in dot plots will undoubtedly result in a noisy background of the plot. You can imagine that there are many successes in the comparison if you only have four possible residues like in nucleotide sequences. Therefore you can set a window size which is smoothing the dot plot. Instead of comparing single residues it compares subsequences of length set as window size. The score is now calculated with respect to aligning the subsequences.

- Threshold
  The dot plot shows the calculated scores with colored threshold. Hence you can better recognize the most important similarities.

Examples and interpretations of dot plots

Contrary to simple sequence alignments dot plots can be a very useful tool for spotting various evolutionary events which may have happened to the sequences of interest.

Below is shown some examples of dot plots where sequence insertions, low complexity regions, inverted repeats etc. can be identified visually.

Similar sequences

The most simple example of a dot plot is obtained by plotting two homologous sequences of interest. If very similar or identical sequences are plotted against each other a diagonal line will occur.

The dot plot in figure 13.16 shows two related sequences of the Influenza A virus nucleoproteins infecting ducks and chickens. Accession numbers from the two sequences are: DQ232610 and DQ023146. Both sequences can be retrieved directly from `http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi`.

Repeated regions

Sequence repeats can also be identified using dot plots. A repeat region will typically show up as

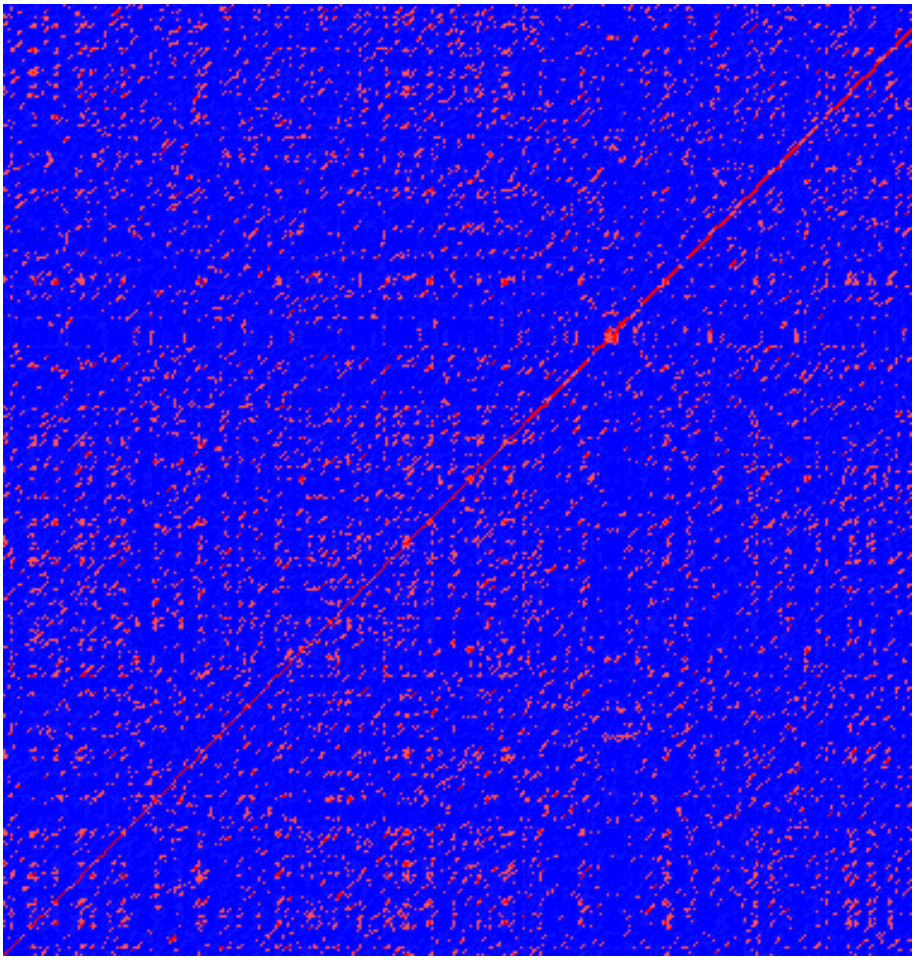


Figure 13.16: *Dot plot of DQ232610 vs. DQ023146 (Influenza A virus nucleoproteins) showing and overall similarity*

lines parallel to the diagonal line.

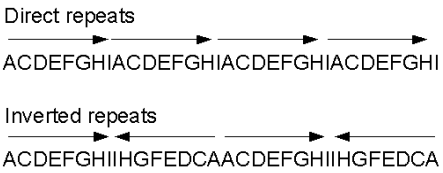


Figure 13.17: *Direct and inverted repeats shown on an amino acid sequence generated for demonstration purposes.*

If the dot plot shows more than one diagonal in the same region of a sequence, the regions depending to the other sequence are repeated. In figure 13.18 you can see a sequence with repeats.

Frame shifts

Frame shifts in a nucleotide sequence can occur due to insertions, deletions or mutations. Such frame shifts can be visualized in a dot plot as seen in figure 13.19. In this figure, three frame shifts for the sequence on the y-axis are found.

1. Deletion of nucleotides

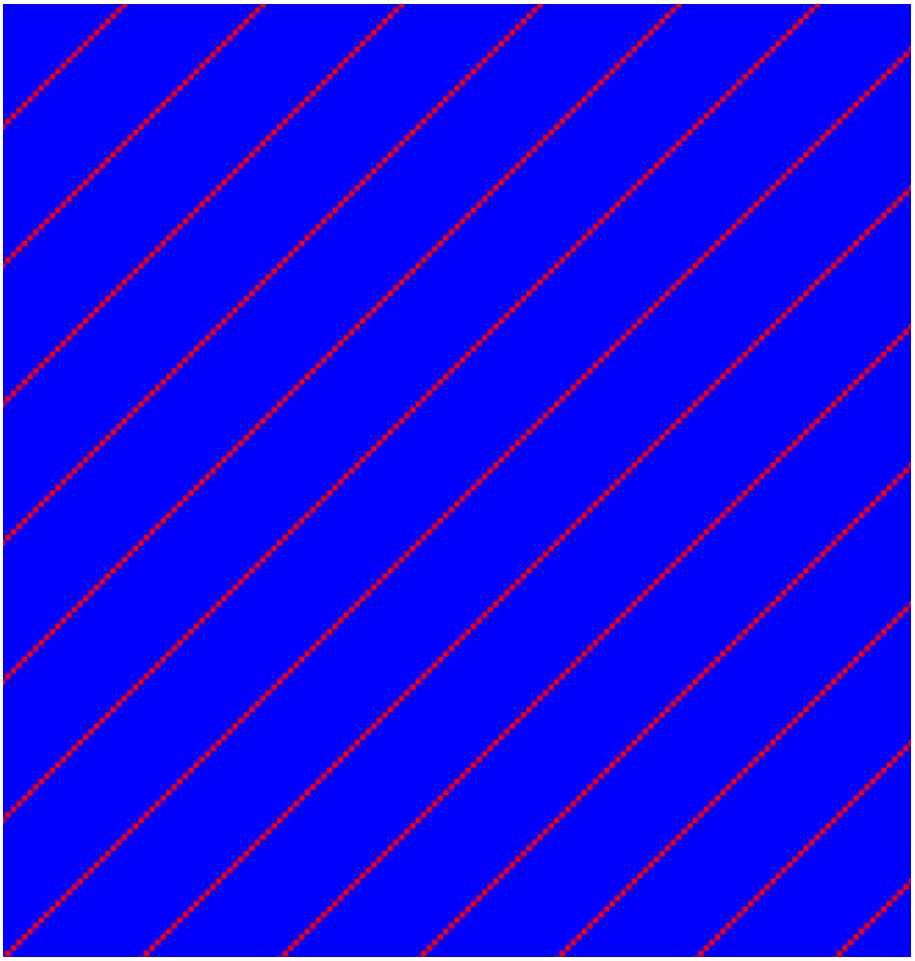


Figure 13.18: *The dot plot of a sequence showing repeated elements. See also figure 13.17.*

2. Insertion of nucleotides
3. Mutation (out of frame)

Sequence inversions

In dot plots you can see an inversion of sequence as contrary diagonal to the diagonal showing similarity. In figure 13.20 you can see a dot plot (window length is 3) with an inversion.

Low-complexity regions

Low-complexity regions in sequences can be found as regions around the diagonal all obtaining a high score. Low complexity regions are calculated from the redundancy of amino acids within a limited region [Wootton and Federhen, 1993]. These are most often seen as short regions of only a few different amino acids. In the middle of figure 13.21 is a square shows the low-complexity region of this sequence.

13.4.4 Bioinformatics explained: Scoring matrices

Biological sequences have evolved throughout time and evolution has shown that not all changes to a biological sequence is equally likely to happen. Certain amino acid substitutions (change of one amino acid to another) happen often, whereas other substitutions are very rare. For instance,

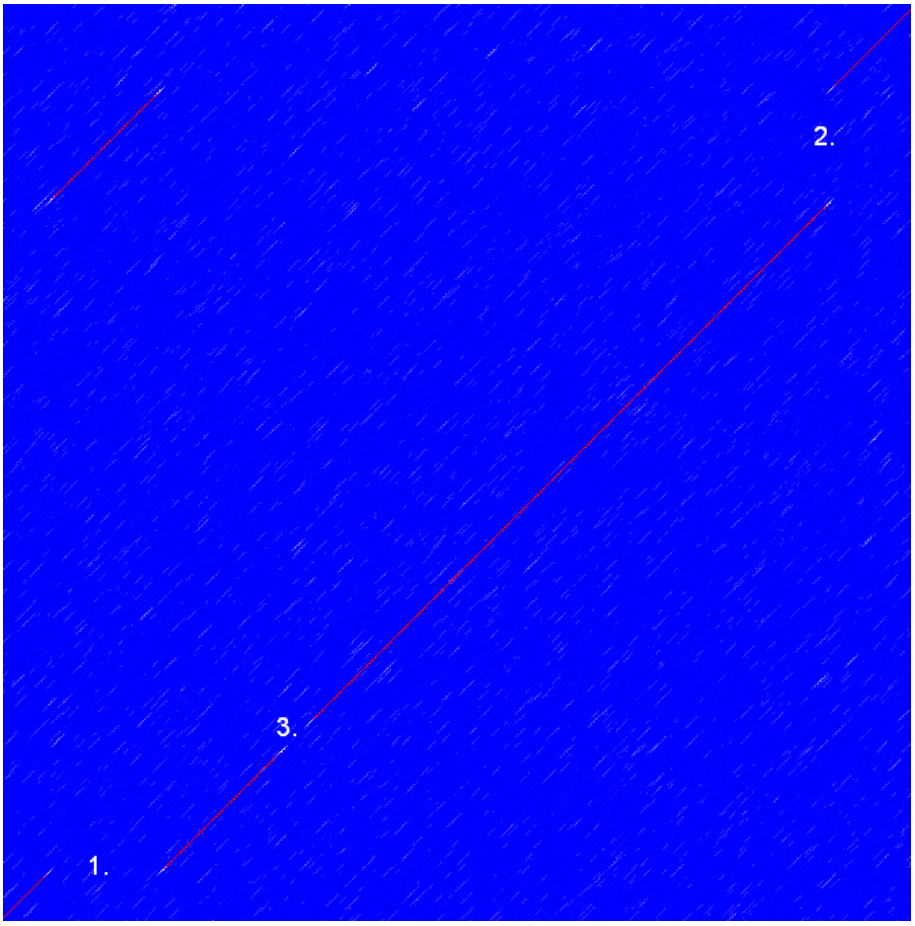


Figure 13.19: *This dot plot show various frame shifts in the sequence. See text for details.*

tryptophan (W) which is a relatively rare amino acid, will only — on very rare occasions — mutate into a leucine (L).

Based on evolution of proteins it became apparent that these changes or substitutions of amino acids can be modeled by a scoring matrix also refereed to as a substitution matrix. See an example of a scoring matrix in table 13.1. This matrix lists the substitution scores of every single amino acid. A score for an aligned amino acid pair is found at the intersection of the corresponding column and row. For example, the substitution score from an arginine (R) to a lysine (K) is 2. The diagonal show scores for amino acids which have not changed. Most substitutions changes have a negative score. Only rounded numbers are found in this matrix.

The two most used matrices are the BLOSUM [Henikoff and Henikoff, 1992] and PAM [Dayhoff and Schwartz, 1978].

Different scoring matrices

PAM

The first PAM matrix (Point Accepted Mutation) was published in 1978 by Dayhoff et al. The PAM matrix was build through a global alignment of related sequences all having sequence similarity above 85% [Dayhoff and Schwartz, 1978]. A PAM matrix shows the probability that any given amino acid will mutate into another in a given time interval. As an example, PAM1 gives that one amino acid out of a 100 will mutate in a given time interval. In the other end of the scale, a

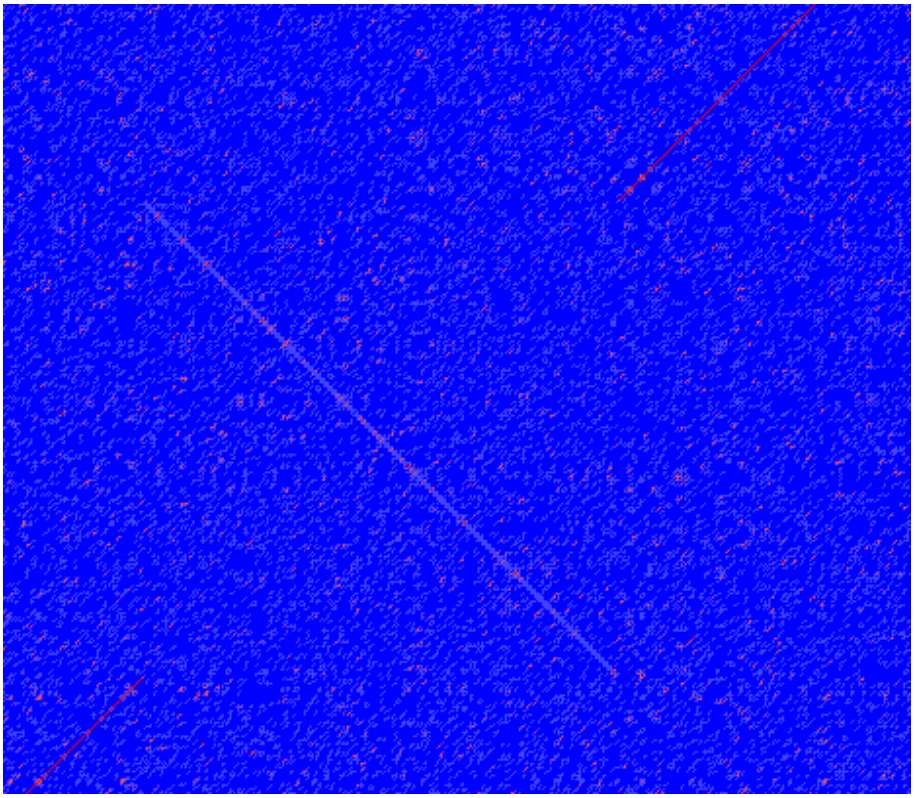


Figure 13.20: *The dot plot showing an inversion in a sequence. See also figure 13.17.*

PAM256 matrix, gives the probability of 256 mutations in a 100 amino acids (see figure 13.22).

There are some limitation to the PAM matrices which makes the BLOSUM matrices somewhat more attractive. The dataset on which the initial PAM matrices were build is very old by now, and the PAM matrices assume that all amino acids mutate at the same rate - this is not a correct assumption.

BLOSUM

In 1992, 14 years after the PAM matrices were published, the BLOSUM matrices (BLOcks SUbstitution Matrix) were developed and published [Henikoff and Henikoff, 1992].

Henikoff et al. wanted to model more divergent proteins, thus they used locally aligned sequences where none of the aligned sequences share less than 62% identity. This resulted in a scoring matrixï¿½called BLOSUM62. In contrast to the PAM matrices the BLOSUM matrices are calculated from alignments without gaps emerging from the BLOCKS database `http://blocks.fhcrc.org/`.

Sean Eddy recently wrote a paper reviewing the BLOSUM62 substitution matrix and how to calculate the scores [Eddy, 2004].

Use of scoring matrices

Deciding which scoring matrix you should use in order of obtain the best alignment results is a difficult task. If you have no prior knowledge on the sequence the BLOSUM62 is probably the best choice. This matrix has become the *de facto* standard for scoring matrices and is also used as the default matrix in BLAST searches. The selection of a "wrong" scoring matrix will most

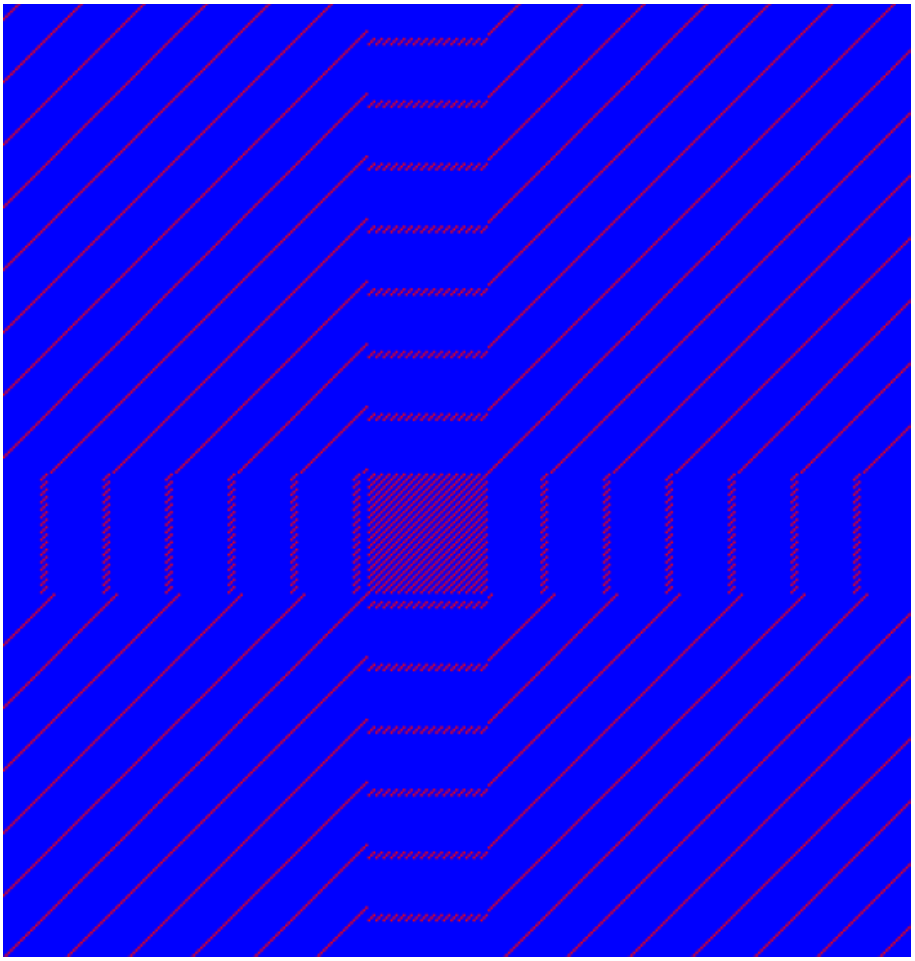


Figure 13.21: *The dot plot showing a low-complexity region in the sequence. The sequence is artificial and low complexity regions do not always show as a square.*

probable strongly influence on the outcome of the analysis. In general a few rules apply to the selection of scoring matrices.

- For closely related sequences choose BLOSUM matrices created for highly similar alignments, like BLOSUM80. You can also select low PAM matrices such as PAM1.
- For distant related sequences, select low BLOSUM matrices (for example BLOSUM45) or high PAM matrices such as PAM250.

The BLOSUM matrices with low numbers correspond to PAM matrices with high numbers. (See figure 13.22) for correlations between the PAM and BLOSUM matrices. To summarize, if you want to find distant related proteins to a sequence of interest using BLAST, you could benefit of using BLOSUM45 or similar matrices.

Other useful resources

BLOKS database
http://blocks.fhcrc.org/

NCBI help site
http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Table 13.1: **The BLOSUM62 matrix**. A tabular view of the BLOSUM62 matrix containing all possible substitution scores [Henikoff and Henikoff, 1992].

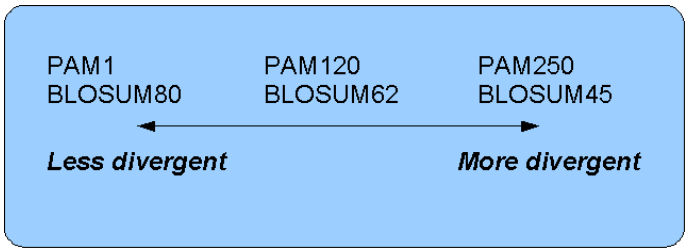


Figure 13.22: *Relationship between scoring matrices. The BLOSUM62 has become a* de facto *standard scoring matrix for a wide range of alignment programs. It is the default matrix in BLAST.*

13.5 Sequence statistics

CLC Drug Discovery Workbench can produce an output with many relevant statistics for protein sequences. Some of the statistics are also relevant to produce for DNA sequences. Therefore, this section deals with both types of statistics. The required steps for producing the statistics are the same.

To create a statistic for the sequence, do the following:

Toolbox | **Sequence Analysis** ( )| **Create Sequence Statistics** ( )

This opens a dialog where you can alter your choice of sequences. If you had already selected sequences in the Navigation Area, these will be shown in the **Selected Elements** window. However you can remove these, or add others, by using the arrows to move sequences in or out of the **Selected Elements** window. You can also add sequence lists.

Note! You cannot create statistics for DNA and protein sequences at the same time; they must be run separately.

When the sequences are selected, click **Next**.

This opens the dialog displayed in figure 13.23.

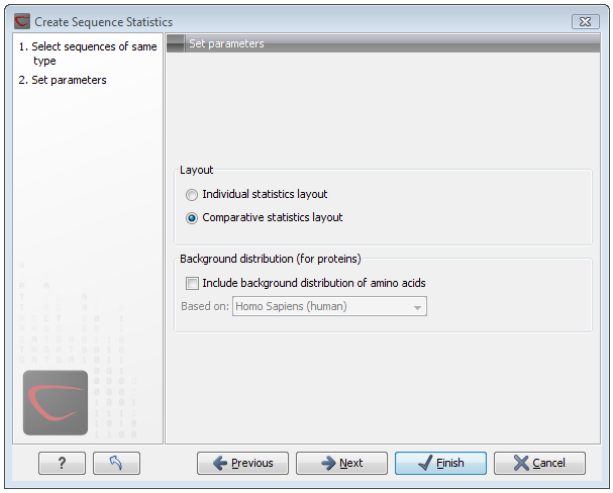


Figure 13.23: *Setting parameters for the sequence statistics.*

The dialog offers to adjust the following parameters:

- **Individual statistics layout.** If more sequences were selected in **Step 1**, this function generates separate statistics for each sequence.
- **Comparative statistics layout.** If more sequences were selected in **Step 1**, this function generates statistics with comparisons between the sequences.

You can also choose to include Background distribution of amino acids. If this box is ticked, an extra column with amino acid distribution of the chosen species, is included in the table output. (The distributions are calculated from UniProt `www.uniprot.org` version 6.0, dated September 13 2005.)

Click **Next** if you wish to adjust how to handle the results (see section 7.2). If not, click **Finish**. An example of protein sequence statistics is shown in figure 13.24.

1 Protein statistics

1.1 Sequence information

Sequence type	Protein
Length	147
Organism	Mus musculus
Name	CAA32220
Description	haemoglobin beta-h0 chain [Mus musculus].
Modification Date	18-APR-2005
Weight	16,412 kDa

1.2 Half-life

N-terminal aa	Half-life mammals	Half-life yeast	Half-life E.Coli

Figure 13.24: *Example of protein sequence statistics.*

Nucleotide sequence statistics are generated using the same dialog as used for protein sequence statistics. However, the output of Nucleotide sequence statistics is less extensive than that of the protein sequence statistics.

Note! The headings of the tables change depending on whether you calculate 'individual' or 'comparative' sequence statistics.

The output of comparative protein sequence statistics include:

- Sequence information:
  - Sequence type
  - Length
  - Organism
  - Name
  - Description
  - Modification Date
  - Weight. This is calculated like this: $sum_{unitsinsequence}(weight(unit)) - links * weight(H2O)$ where `links` is the sequence length minus one and `units` are amino acids. The atomic composition is defined the same way.
  - Isoelectric point
  - Aliphatic index
- Half-life
- Extinction coefficient
- Counts of Atoms
- Frequency of Atoms
- Count of hydrophobic and hydrophilic residues
- Frequencies of hydrophobic and hydrophilic residues
- Count of charged residues
- Frequencies of charged residues
- Amino acid distribution
- Histogram of amino acid distribution
- Annotation table
- Counts of di-peptides
- Frequency of di-peptides

- Sequence Information:
  - Sequence type
  - Length
  - Organism
  - Name

  - Description
  - Modification Date
  - Weight
  - Isoelectric point
  - Aliphatic index
- Amino acid distribution
- Annotation table

The output of nucleotide sequence statistics include:

- General statistics:
  - Sequence type
  - Length
  - Organism
  - Name
  - Description
  - Modification Date
  - Weight. This is calculated like this: $sum_{unitsinsequence}(weight(unit)) - links * weight(H2O)$ where `links` is the sequence length minus one for linear sequences and sequence length for circular molecules. The `units` are monophosphates. Both the weight for single- and double stranded molecules are includes. The atomic composition is defined the same way.
- Atomic composition
- Nucleotide distribution table
- Nucleotide distribution histogram
- Annotation table
- Counts of di-nucleotides
- Frequency of di-nucleotides

- General statistics:
  - Sequence type
  - Length
  - Organism
  - Name
  - Description
  - Modification Date
  - Weight (calculated as single-stranded DNA)

- Nucleotide distribution table
- Annotation table

If nucleotide sequences are used as input, and these are annotated with CDS, a section on Codon statistics for Coding Regions is included.

A short description of the different areas of the statistical output is given in section 13.5.1.

13.5.1 Bioinformatics explained: Protein statistics

Every protein holds specific and individual features which are unique to that particular protein. Features such as isoelectric point or amino acid composition can reveal important information of a novel protein. Many of the features described below are calculated in a simple way.

Molecular weight

The molecular weight is the mass of a protein or molecule. The molecular weight is simply calculated as the sum of the atomic mass of all the atoms in the molecule.

The weight of a protein is usually represented in Daltons (Da).

A calculation of the molecular weight of a protein does not usually include additional posttranslational modifications. For native and unknown proteins it tends to be difficult to assess whether posttranslational modifications such as glycosylations are present on the protein, making a calculation based solely on the amino acid sequence inaccurate. The molecular weight can be determined very accurately by mass-spectrometry in a laboratory.

Isoelectric point

The isoelectric point (pI) of a protein is the pH where the proteins has no net charge. The pI is calculated from the pKa values for 20 different amino acids. At a pH below the pI, the protein carries a positive charge, whereas if the pH is above pI the proteins carry a negative charge. In other words, pI is high for basic proteins and low for acidic proteins. This information can be used in the laboratory when running electrophoretic gels. Here the proteins can be separated, based on their isoelectric point.

Aliphatic index

The aliphatic index of a protein is a measure of the relative volume occupied by aliphatic side chain of the following amino acids: alanine, valine, leucine and isoleucine. An increase in the aliphatic index increases the thermostability of globular proteins. The index is calculated by the following formula.

$$Aliphaticindex = X(Ala) + a * X(Val) + b * X(Leu) + b * (X)Ile$$

X(Ala), *X(Val)*, *X(Ile)* and *X(Leu)* are the amino acid compositional fractions. The constants a and b are the relative volume of valine (a=2.9) and leucine/isoleucine (b=3.9) side chains compared to the side chain of alanine [Ikai, 1980].

Amino acid	Mammalian	Yeast	E. coli
Ala (A)	4.4 hour	>20 hours	>10 hours
Cys (C)	1.2 hours	>20 hours	>10 hours
Asp (D)	1.1 hours	3 min	>10 hours
Glu (E)	1 hour	30 min	>10 hours
Phe (F)	1.1 hours	3 min	2 min
Gly (G)	30 hours	>20 hours	>10 hours
His (H)	3.5 hours	10 min	>10 hours
Ile (I)	20 hours	30 min	>10 hours
Lys (K)	1.3 hours	3 min	2 min
Leu (L)	5.5 hours	3 min	2 min
Met (M)	30 hours	>20 hours	>10 hours
Asn (N)	1.4 hours	3 min	>10 hours
Pro (P)	>20 hours	>20 hours	?
Gln (Q)	0.8 hour	10 min	>10 hours
Arg (R)	1 hour	2 min	2 min
Ser (S)	1.9 hours	>20 hours	>10 hours
Thr (T)	7.2 hours	>20 hours	>10 hours
Val (V)	100 hours	>20 hours	>10 hours
Trp (W)	2.8 hours	3 min	2 min
Tyr (Y)	2.8 hours	10 min	2 min

Table 13.2: **Estimated half life**. Half life of proteins where the N-terminal residue is listed in the first column and the half-life in the subsequent columns for mammals, yeast and *E. coli*.

Estimated half-life

The half life of a protein is the time it takes for the protein pool of that particular protein to be reduced to the half. The half life of proteins is highly dependent on the presence of the N-terminal amino acid, thus overall protein stability [Bachmair et al., 1986, Gonda et al., 1989, Tobias et al., 1991]. The importance of the N-terminal residues is generally known as the 'N-end rule'. The N-end rule and consequently the N-terminal amino acid, simply determines the half-life of proteins. The estimated half-life of proteins have been investigated in mammals, yeast and *E. coli* (see Table 13.2). If leucine is found N-terminally in mammalian proteins the estimated half-life is 5.5 hours.

Extinction coefficient

This measure indicates how much light is absorbed by a protein at a particular wavelength. The extinction coefficient is measured by UV spectrophotometry, but can also be calculated. The amino acid composition is important when calculating the extinction coefficient. The extinction coefficient is calculated from the absorbance of cysteine, tyrosine and tryptophan using the following equation:

$$Ext(Protein) = count(Cystine) * Ext(Cystine) + count(Tyr) * Ext(Tyr) + count(Trp) * Ext(Trp)$$

where Ext is the extinction coefficient of amino acid in question. At 280nm the extinction coefficients are: Cys=120, Tyr=1280 and Trp=5690.

This equation is only valid under the following conditions:

- pH 6.5
- 6.0 M guanidium hydrochloride
- 0.02 M phosphate buffer

The extinction coefficient values of the three important amino acids at different wavelengths are found in [Gill and von Hippel, 1989].

Knowing the extinction coefficient, the absorbance (optical density) can be calculated using the following formula:

$$Absorbance(Protein) = \frac{Ext(Protein)}{Molecular\ weight}$$

Two values are reported. The first value is computed assuming that all cysteine residues appear as half cystines, meaning they form di-sulfide bridges to other cysteines. The second number assumes that no di-sulfide bonds are formed.

Atomic composition

Amino acids are indeed very simple compounds. All 20 amino acids consist of combinations of only five different atoms. The atoms which can be found in these simple structures are: Carbon, Nitrogen, Hydrogen, Sulfur, Oxygen. The atomic composition of a protein can for example be used to calculate the precise molecular weight of the entire protein.

Total number of negatively charged residues (Asp+Glu)

At neutral pH, the fraction of negatively charged residues provides information about the location of the protein. Intracellular proteins tend to have a higher fraction of negatively charged residues than extracellular proteins.

Total number of positively charged residues (Arg+Lys)

At neutral pH, nuclear proteins have a high relative percentage of positively charged amino acids. Nuclear proteins often bind to the negatively charged DNA, which may regulate gene expression or help to fold the DNA. Nuclear proteins often have a low percentage of aromatic residues [Andrade et al., 1998].

Amino acid distribution

Amino acids are the basic components of proteins. The amino acid distribution in a protein is simply the percentage of the different amino acids represented in a particular protein of interest. Amino acid composition is generally conserved through family-classes in different organisms which can be useful when studying a particular protein or enzymes across species borders. Another interesting observation is that amino acid composition variate slightly between proteins from different subcellular localizations. This fact has been used in several computational methods, used for prediction of subcellular localization.

Annotation table

This table provides an overview of all the different annotations associated with the sequence and their incidence.

Dipeptide distribution

This measure is simply a count, or frequency, of all the observed adjacent pairs of amino acids (dipeptides) found in the protein. It is only possible to report neighboring amino acids. Knowledge on dipeptide composition have previously been used for prediction of subcellular localization.

13.6 Pattern discovery

With *CLC Drug Discovery Workbench* you can perform pattern discovery on both DNA and protein sequences. Advanced hidden Markov models can help to identify unknown sequence patterns across single or even multiple sequences.

In order to search for unknown patterns:

Toolbox | **Sequence Analysis** ( )| **Pattern Discovery** ( )

If a sequence was selected before choosing the Toolbox action, the sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the selected elements.

You can perform the analysis on several DNA or several protein sequences at a time. If the analysis is performed on several sequences at a time the method will search for patterns which is common between all the sequences. Annotations will be added to all the sequences and a view is opened for each sequence.

Click **Next** to adjust parameters (see figure 13.25).

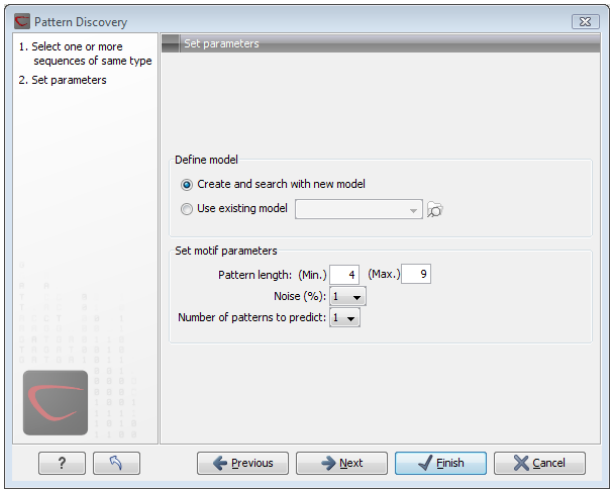


Figure 13.25: *Setting parameters for the pattern discovery. See text for details.*

In order to search unknown sequences with an already existing model:

Select to use an already existing model which is seen in figure 13.25. Models are represented with the following icon in the **Navigation Area** ( ).

13.6.1 Pattern discovery search parameters

Various parameters can be set prior to the pattern discovery. The parameters are listed below and a screenshot of the parameter settings can be seen in figure 13.25.

- **Create and search with new model.** This will create a new HMM model based on the selected sequences. The found model will be opened after the run and presented in a table view. It can be saved and used later if desired.
- **Use existing model.** It is possible to use already created models to search for the same pattern in new sequences.
- **Minimum pattern length.** Here, the minimum length of patterns to search for, can be specified.
- **Maximum pattern length.** Here, the maximum length of patterns to search for, can be specified.
- **Noise (%).** Specify noise-level of the model. This parameter has influence on the level of degeneracy of patterns in the sequence(s). The noise parameter can be 1,2,5 or 10 percent.
- **Number of different kinds of patterns to predict.** Number of iterations the algorithm goes through. After the first iteration, we force predicted pattern-positions in the first run to be member of the background: In that way, the algorithm finds new patterns in the second iteration. Patterns marked 'Pattern1' have the highest confidence. The maximal iterations to go through is 3.
- **Include background distribution.** For protein sequences it is possible to include information on the background distribution of amino acids from a range of organisms.

Click **Next** if you wish to adjust how to handle the results (see section 7.2). If not, click **Finish**. This will open a view showing the patterns found as annotations on the original sequence (see figure 13.26). If you have selected several sequences, a corresponding number of views will be opened.

Figure 13.26: *Sequence view displaying two discovered patterns.*

13.6.2 Pattern search output

If the analysis is performed on several sequences at a time the method will search for patterns in the sequences and open a new view for each of the sequences, in which a pattern was discovered. Each novel pattern will be represented as an annotation of the type **Region**. More information on each found pattern is available through the tool-tip, including detailed information on the position of the pattern and quality scores.

It is also possible to get a tabular view of all found patterns in one combined table. Then each found pattern will be represented with various information on obtained scores, quality of the pattern and position in the sequence.

A table view of emission values of the actual used HMM model is presented in a table view. This model can be saved and used to search for a similar pattern in new or unknown sequences.

13.7 Motif Search

CLC Drug Discovery Workbench offers advanced and versatile options to search for known motifs represented either by a simple sequence or a more advanced regular expression. These advanced search capabilities are available for use in both DNA and protein sequences.

There are two ways to access this functionality:

- When viewing sequences, it is possible to have motifs calculated and shown on the sequence . This approach is called *Dynamic motifs* and is an easy way to spot known sequence motifs when working with sequences for cloning etc.
- A more refined and systematic search for motifs can be performed through the **Toolbox**. This will generate a table and optionally add annotations to the sequences.

The two approaches are described below.

13.7.1 Dynamic motifs

In the **Side Panel** of sequence views, there is a group called **Motifs** (see figure 13.27).

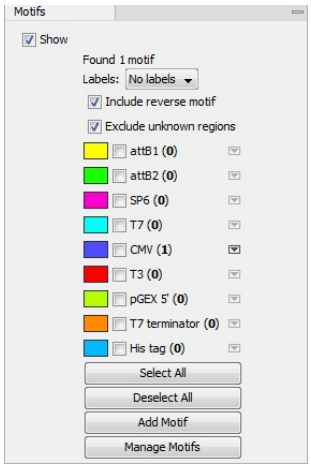


Figure 13.27: *Dynamic motifs in the Side Panel.*

The Workbench will look for the listed motifs in the sequence that is open and by clicking the check box next to the motif it will be shown in the view as illustrated in figure 13.28.

This case shows the CMV promoter primer sequence which is one of the pre-defined motifs in *CLC Drug Discovery Workbench*. The motif is per default shown as a faded arrow with no text. The direction of the arrow indicates the strand of the motif.

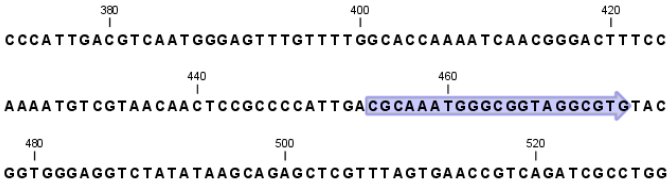


Figure 13.28: *Showing dynamic motifs on the sequence.*

Placing the mouse cursor on the arrow will display additional information about the motif as illustrated in figure 13.29.

Figure 13.29: *Showing dynamic motifs on the sequence.*

To add **Labels** to the motif, select the **Flag** or **Stacked** option. They will put the name of the motif as a flag above the sequence. The stacked option will stack the labels when there is more than one motif so that all labels are shown.

Below the labels option there are two options for controlling the way the sequence should be searched for motifs:

- **Include reverse motifs**. This will also find motifs on the negative strand (only available for nucleotide sequences)
- **Exclude matches in N-regions for simple motifs**. The motif search handles ambiguous characters in the way that two residues are different if they do not have any residues in common. For example: For nucleotides, *N* matches any character and *R* matches *A,G*. For proteins, *X* matches any character and *Z* matches *E,Q*. Genome sequence often have large regions with unknown sequence. These regions are very often padded with N's. Ticking this checkbox will not display hits found in N-regions and if a one residue in a motif matches to an N, it will be treated as a mismatch.

The list of motifs shown in figure 13.27 is a pre-defined list that is included with the *CLC Drug Discovery Workbench*. You can define your own set of motifs to use instead. In order to do this, you can either click on the **Add Motif** button in the side panel (see figure 13.27) and directly define and add motifs of choice as illustrated in figure 13.33. Alternatively, you can create and save a **Motif list** ( ) (see section 13.8). Subsequently, in the sequence view click the **Manage Motifs** button in the side panel which will bring up the dialog shown in figure 13.30.

At the top, select a motif list by clicking the **Browse** ( ) button. When the motif list is selected, its motifs are listed in the panel in the left-hand side of the dialog. The right-hand side panel contains the motifs that will be listed in the **Side Panel** when you click **Finish**.

13.7.2 Motif search from the Toolbox

The dynamic motifs described in section 13.7.1 provide a quick way of routinely scanning a sequence for commonly used motifs, but in some cases a more systematic approach is needed.

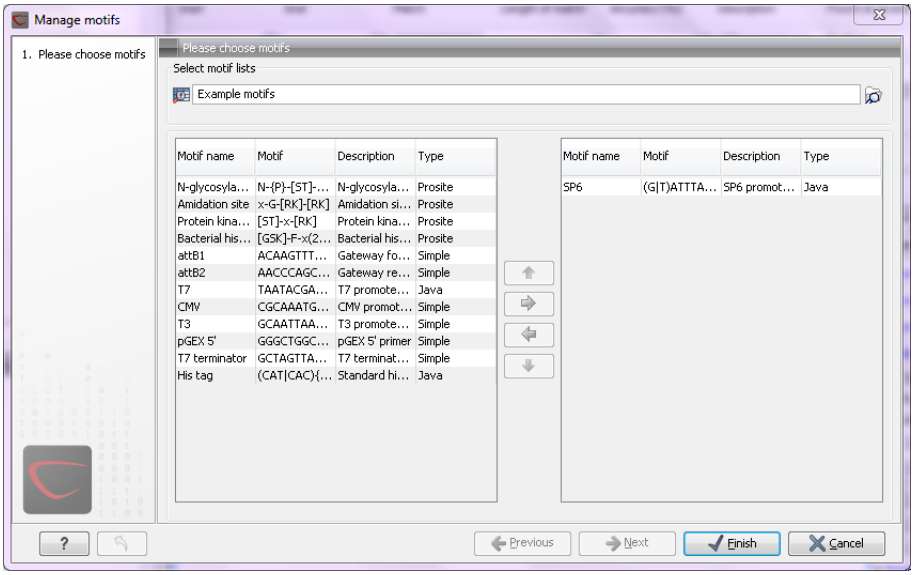


Figure 13.30: *Managing the motifs to be shown.*

The motif search in the **Toolbox** provides an option to search for motifs with a user-specified similarity to the target sequence, and furthermore the motifs found can be displayed in an overview table. This is particularly useful when searching for motifs on many sequences.

To start the Toolbox motif search, go to:

Toolbox | **Sequence Analysis** ( )| **Motif Search** ( )

A dialog window will be launched. Use the arrows to add or remove sequences or sequence lists between the Navigation Area and the selected elements.

You can perform the analysis on several DNA or several protein sequences at a time. In this case, the method will search for patterns in the sequences and create an overview table of the motifs found in all sequences.

Click **Next** to adjust parameters (see figure 13.31).

The options for the motif search are:

- **Motif types.** Choose what kind of motif to be used:
  - Simple motif. Choosing this option means that you enter a simple motif, e.g. ATGATGNNATG.
  - Java regular expression. See section 13.7.3.
  - Prosite regular expression. For proteins, you can enter different protein patterns from the PROSITE database (protein patterns using regular expressions and describing specific amino acid sequences). The PROSITE database contains a great number of patterns and have been used to identify related proteins (see `http://www.expasy.org/cgi-bin/prosite-list.pl`).
  - Use motif list. Clicking the small button ( ) will allow you to select a saved motif list (see section 13.8).
- **Motif.** If you choose to search with a simple motif, you should enter a literal string as your motif. Ambiguous amino acids and nucleotides are allowed. Example; ATGATGNNATG. If

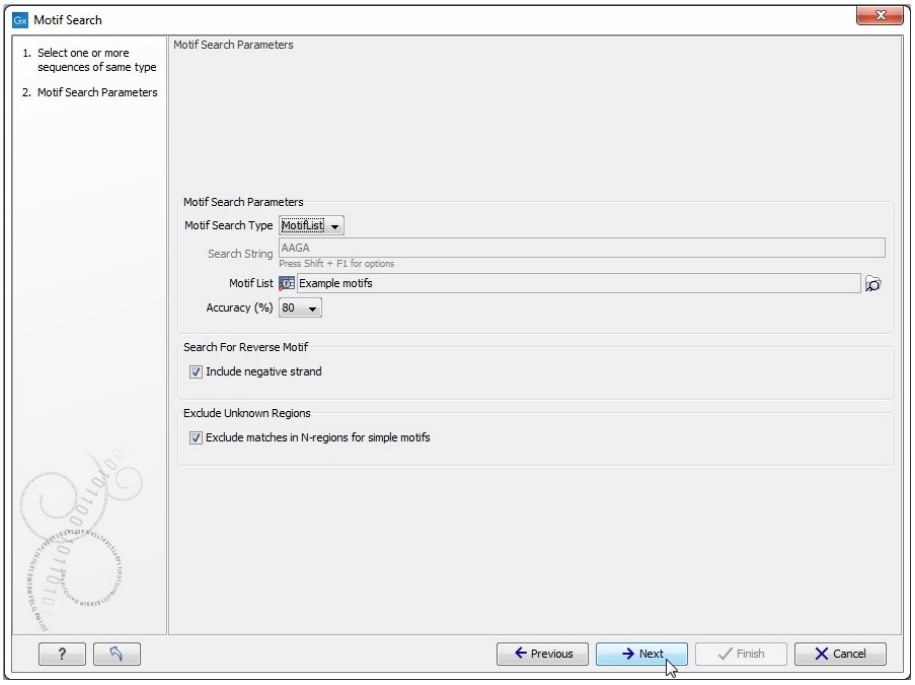


Figure 13.31: *Setting parameters for the motif search.*

your motif type is Java regular expression, you should enter a regular expression according to the syntax rules described in section 13.7.3. Press **Shift + F1** key for options. For proteins, you can search with a Prosite regular expression and you should enter a protein pattern from the PROSITE database.

- **Accuracy.** If you search with a simple motif, you can adjust the accuracy of the motif to the match on the sequence. If you type in a simple motif and let the accuracy be 80%, the motif search algorithm runs through the input sequence and finds all subsequences of the same length as the simple motif such that the fraction of identity between the subsequence and the simple motif is at least 80%. A motif match is added to the sequence as an annotation with the exact fraction of identity between the subsequence and the simple motif. If you use a list of motifs, the accuracy applies only to the simple motifs in the list.

- **Search for reverse motif.** This enables searching on the negative strand on nucleotide sequences.

- **Exclude unknown regions.** Genome sequence often have large regions with unknown sequence. These regions are very often padded with N's. Ticking this checkbox will not display hits found in N-regions.Motif search handles ambiguous characters in the way that two residues are different if they do not have any residues in common. For example: For nucleotides, *N* matches any character and *R* matches *A,G*. For proteins, *X* matches any character and *Z* matches *E,Q*.

Click **Next** to adjust how to handle the results and then click **Finish**. There are two types of results that can be produced:

- **Add annotations**. This will add an annotation to the sequence when a motif is found (an example is shown in figure 13.32.

- **Create table**. This will create an overview table of all the motifs found for all the input sequences.

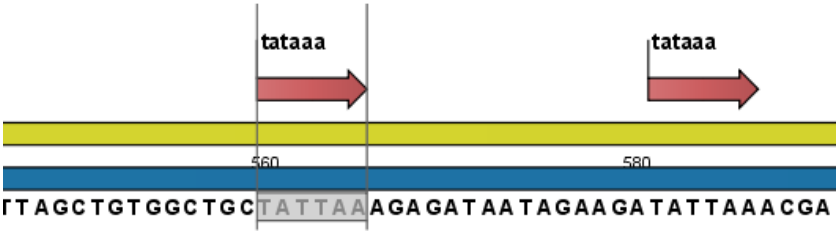


Figure 13.32: *Sequence view displaying the pattern found. The search string was 'tataaa'.*

13.7.3 Java regular expressions

A regular expressions is a string that describes or matches a set of strings, according to certain syntax rules. They are usually used to give a concise description of a set, without having to list all elements. The simplest form of a regular expression is a literal string. The syntax used for the regular expressions is the Java regular expression syntax (see `http://java.sun.com/docs/books/tutorial/essential/regex/index.html`). Below is listed some of the most important syntax rules which are also shown in the help pop-up when you press Shift + F1:

[A-Z] will match the characters *A* through *Z* (Range). You can also put single characters between the brackets: The expression *[AGT]* matches the characters *A*, *G* or *T*.

[A-D[M-P]] will match the characters *A* through *D* and *M* through *P* (Union). You can also put single characters between the brackets: The expression *[AG[M-P]]* matches the characters *A*, *G* and *M* through *P*.

[A-M&&[H-P]] will match the characters between *A* and *M* lying between *H* and *P* (Intersection). You can also put single characters between the brackets. The expression *[A-M&&[HGTDA]]* matches the characters *A* through *M* which is *H*, *G*, *T*, *D* or *A*.

[^A-M] will match any character except those between *A* and *M* (Excluding). You can also put single characters between the brackets: The expression *[^AG]* matches any character except *A* and *G*.

[A-Z&&[^M-P]] will match any character *A* through *Z* except those between *M* and *P* (Subtraction). You can also put single characters between the brackets: The expression *[A-P&&[^CG]]* matches any character between *A* and *P* except *C* and *G*.

The symbol . matches any character.

X{n} will match a repetition of an element indicated by following that element with a numerical value or a numerical range between the curly brackets. For example, *ACG{2}* matches the string *ACGG* and *(ACG){2}* matches *ACGACG*.

X{n,m} will match a certain number of repetitions of an element indicated by following that element with two numerical values between the curly brackets. The first number is a lower limit on the number of repetitions and the second number is an upper limit on the number of repetitions. For example, *ACT{1,3}* matches *ACT, ACTT* and *ACTTT*.

X{n,} represents a repetition of an element at least *n* times. For example, *(AC){2,}* matches all strings *ACAC*, *ACACAC*, *ACACACAC*,...

The symbol ^ restricts the search to the beginning of your sequence. For example, if you search through a sequence with the regular expression *^AC*, the algorithm will find a match if *AC* occurs in the beginning of the sequence.

The symbol *$* restricts the search to the end of your sequence. For example, if you search through a sequence with the regular expression *GT$*, the algorithm will find a match if *GT* occurs in the end of the sequence.

Examples
The expression *[ACG][^AC]G{2}* matches all strings of length *4*, where the first character is *A,C* or *G* and the second is any character except *A,C* and the third and fourth character is *G*. The expression *G.[^A]$* matches all strings of length *3* in the end of your sequence, where the first character is *C*, the second any character and the third any character except *A*.

13.8 Create motif list

CLC Drug Discovery Workbench offers advanced and versatile options to create lists of sequence patterns or known motifs, represented either by a literal string or a regular expression.

A motif list can be created using:

Toolbox | **Sequence Analysis (**  **)**| **Create Motif List (**  **)**

File | **New** | **Motif List (**  **)**

Add ( ) button at the bottom of the view. This will open a dialog shown in figure 13.33.

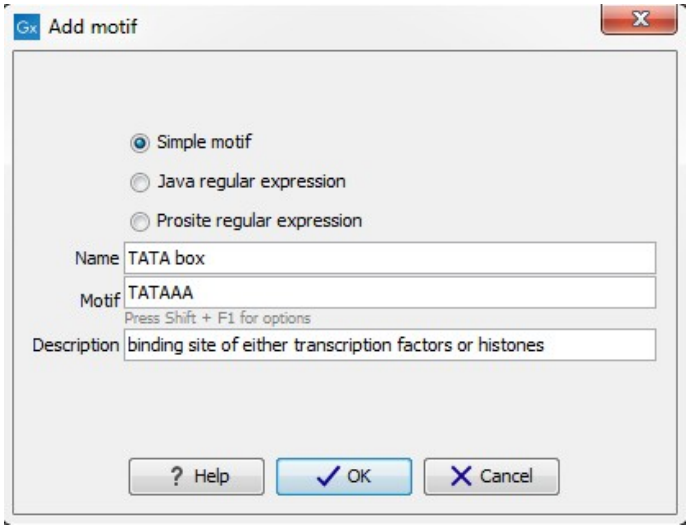


Figure 13.33: *Entering a new motif in the list.*

In this dialog, you can enter the following information:

- **Name**. The name of the motif. In the result of a motif search, this name will appear as the name of the annotation and in the result table.

- **Motif**. The actual motif. See section 13.7.2 for more information about the syntax of motifs.
- **Description**. You can enter a description of the motif. In the result of a motif search, the description will appear in the result table and will be added as a note to the annotation on the sequence (visible in the **Annotation table** ( ) or by placing the mouse cursor on the annotation).
- **Type**. You can enter three different types of motifs: Simple motifs, java regular expressions or PROSITE regular expression. Read more in section 13.7.2.

The motif list can contain a mix of different types of motifs. This is practical because some motifs can be described with the simple syntax, whereas others need the more advanced regular expression syntax.

Instead of manually adding motifs, you can **Import From Fasta File** ( ). This will show a dialog where you can select a fasta file on your computer and use this to create motifs. This will automatically take the name, description and sequence information from the fasta file, and put it into the motif list. The motif type will be "simple".

Besides adding new motifs, you can also edit and delete existing motifs in the list. To edit a motif, either double-click the motif in the list, or select and click the **Edit** ( ) button at the bottom of the view.

To delete a motif, select it and press the Delete key on the keyboard. Alternatively, click **Delete** ( ) in the **Tool bar**.

Save the motif list in the **Navigation Area**, and you will be able to use for Motif Search ( ) (see section 13.7).

13.9 Signal peptide prediction

Signal peptides target proteins to the extracellular environment either through direct plasmamembrane translocation in prokaryotes or is routed through the Endoplasmatic Reticulum in eukaryotic cells. The signal peptide is removed from the resulting mature protein during translocation across the membrane. For prediction of signal peptides, we query SignalP [Nielsen et al., 1997, Bendtsen et al., 2004b] located at `http://www.cbs.dtu.dk/services/SignalP/`. Thus an active internet connection is required to run the signal peptide prediction. Additional information on SignalP and Center for Biological Sequence analysis (CBS) can be found at `http://www.cbs.dtu.dk` and in the original research papers [Nielsen et al., 1997, Bendtsen et al., 2004b].

In order to predict potential signal peptides of proteins, the D-score from the SignalP output is used for discrimination of signal peptide versus non-signal peptide (see section 13.9.3). This score has been shown to be the most accurate [Klee and Ellis, 2005] in an evaluation study of signal peptide predictors.

In order to use SignalP, *you need to download the SignalP plugin* using the plugin manager, see section 1.7.1.

When the plugin is downloaded and installed, you can use it to predict signal peptides:

Toolbox | **Sequence Analysis ( )**| **Signal Peptide Prediction ( )**

If a sequence was selected before choosing the Toolbox action, this sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the selected elements. The SignalP service is limited to 2,000 sequences and 200,000 amino acids for one submission. Each sequence may be no longer than 6,000 amino acids.

Click **Next** to set parameters for the SignalP analysis.

13.9.1 Signal peptide prediction parameter settings

You should select which organism group the input sequences belong to. the default is eukaryote (see figure 13.34).

- Eukaryote (default)
- Gram-negative bacteria
- Gram-positive bacteria

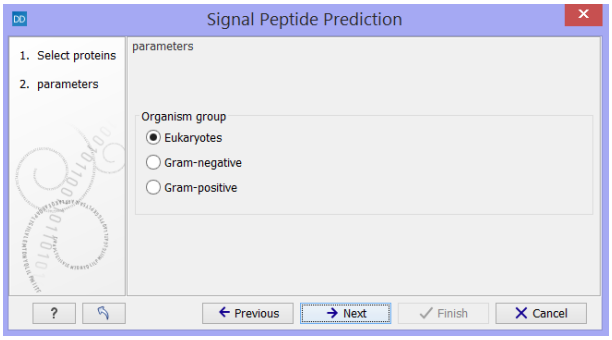


Figure 13.34: *Setting the parameters for signal peptide prediction.*

You can perform the analysis on several protein sequences at a time. This will add annotations to all the sequences and open a view for each sequence if a signal peptide is found. If no signal peptide is found in the sequence, a dialog box will be shown.

The predictions obtained can either be shown as annotations on the sequence, listed in a table or be shown as the detailed and full text output from the SignalP method. This can be used to interpret borderline predictions:

- Add annotations to sequence
- Create table
- Text

Click **Next** to adjust how to handle the results, then click **Finish**.

13.9.2 Signal peptide prediction output

After running the prediction as described above, the protein sequence will show predicted signal peptide as annotations on the original sequence (see figure 13.35). Make sure the Side Panel

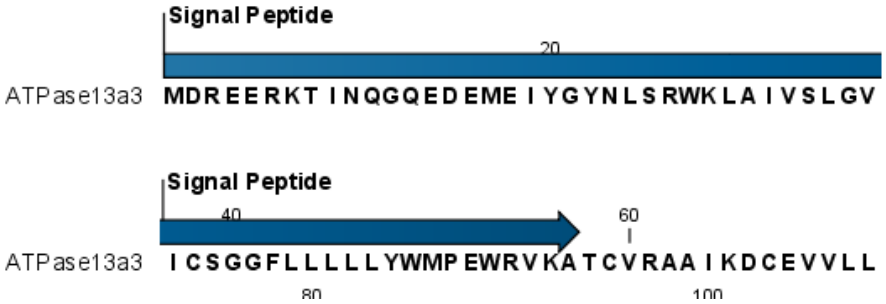


Figure 13.35: *N-terminal signal peptide shown as annotation on the sequence.*

settings of the sequence is so that 'Show annotations' is checked in the 'Annotation layout' palette, and that the annotation type 'Signal peptide' is checked in the 'Annotation types' palette.

Additional notes can be added through the **Edit annotation** ( ) right-click mouse menu. See section 10.3.2.

Undesired annotations can be removed through the **Delete Annotation ( )** right-click mouse menu. See section 10.3.4.

13.9.3 Bioinformatics explained: Prediction of signal peptides

Why the interest in signal peptides?

The importance of signal peptides was shown in 1999 when Günter Blobel received the Nobel Prize in physiology or medicine for his discovery that "proteins have intrinsic signals that govern their transport and localization in the cell" [Blobel, 2000]. He pointed out the importance of defined peptide motifs for targeting proteins to their site of function.

Performing a query to PubMed[1] reveals that thousands of papers have been published, regarding signal peptides, secretion and subcellular localization, including knowledge of using signal peptides as vehicles for chimeric proteins for biomedical and pharmaceutical industry. Many papers describe statistical or machine learning methods for prediction of signal peptides and prediction of subcellular localization in general. After the first published method for signal peptide prediction [von Heijne, 1986], more and more methods have surfaced, although not all methods have been made available publicly.

Different types of signal peptides

Soon after Günter Blobel's initial discovery of signal peptides, more targeting signals were found. Most cell types and organisms employ several ways of targeting proteins to the extracellular environment or subcellular locations. Most of the proteins targeted for the extracellular space or subcellular locations carry specific sequence motifs (signal peptides) characterizing the type of secretion/targeting it undergoes.

Several new different signal peptides or targeting signals have been found during the later years, and papers often describe a small amino acid motif required for secretion of that particular protein. In most of the latter cases, the identified sequence motif is only found in this particular protein and as such cannot be described as a new group of signal peptides.

[1] http://www.ncbi.nlm.nih.gov/entrez/

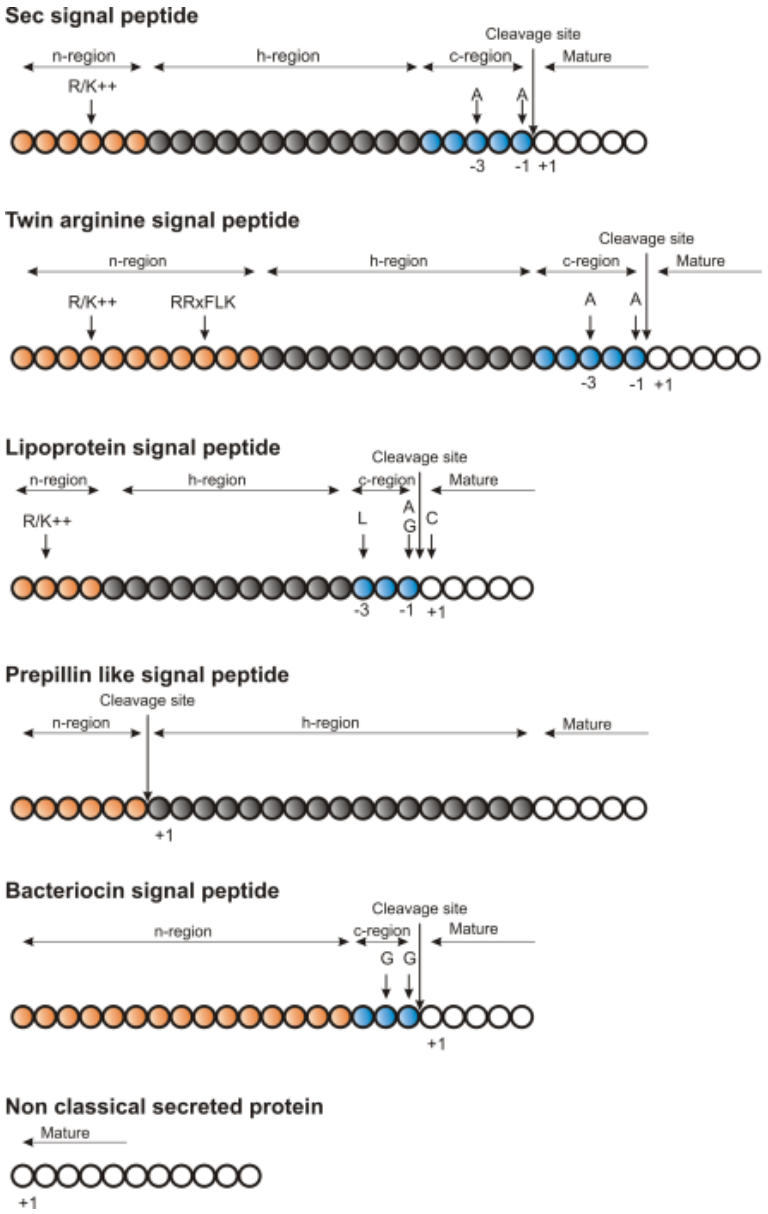


Figure 13.36: *Schematic representation of various signal peptides. Red color indicates n-region, gray color indicates h-region, cyan indicates c-region. All white circles are part of the mature protein.* $+1$ *indicates the first position of the mature protein. The length of the signal peptides is not drawn to scale.*

Describing the various types of signal peptides is beyond the scope of this text but several review papers on this topic can be found on PubMed. Targeting motifs can either be removed from, or retained in the mature protein after the protein has reached the correct and final destination. Some of the best characterized signal peptides are depicted in figure 13.36.

Numerous methods for prediction of protein targeting and signal peptides have been developed; some of them are mentioned and cited in the introduction of the SignalP research paper [Bendtsen et al., 2004b]. However, no prediction method will be able to cover all the different types of signal peptides. Most methods predicts classical signal peptides targeting to the general secretory pathway in bacteria or classical secretory pathway in eukaryotes. Furthermore, a few methods for prediction of non-classically secreted proteins have emerged [Bendtsen et al., 2004a, Bendtsen

et al., 2005].

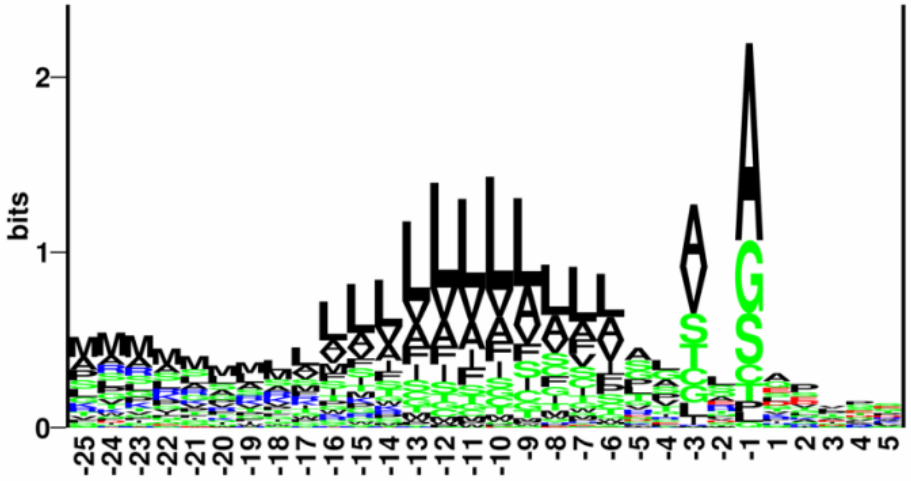


Figure 13.37: *Sequence logo of eukaryotic signal peptides, showing conservation of amino acids in bits [Schneider and Stephens, 1990]. Polar and hydrophobic residues are shown in green and black, respectively, while blue indicates positively charged residues and red negatively charged residues. The logo is based on an ungapped sequence alignment fixed at the -1 position of the signal peptides.*

Prediction of signal peptides and subcellular localization

In the search for accurate prediction of signal peptides, many approaches have been investigated. Almost 20 years ago, the first method for prediction of classical signal peptides was published [von Heijne, 1986]. Nowadays, more sophisticated machine learning methods, such as neural networks, support vector machines, and hidden Markov models have arrived along with the increasing computational power and they all perform superior to the old weight matrix based methods [Menne et al., 2000]. Also, many other "classical" statistical approaches have been carried out, often in conjunction with machine learning methods. In the following sections, a wide range of different signal peptide and subcellular prediction methods will be described.

Most signal peptide prediction methods require the presence of the correct N-terminal end of the preprotein for correct classification. As large scale genome sequencing projects sometimes assign the 5'-end of genes incorrectly, many proteins are annotated without the correct N-terminal [Reinhardt and Hubbard, 1998] leading to incorrect prediction of subcellular localization. These erroneous predictions can be ascribed directly to poor gene finding. Other methods for prediction of subcellular localization use information within the mature protein and therefore they are more robust to N-terminal truncation and gene finding errors.

The SignalP method

One of the most cited and best methods for prediction of classical signal peptides is the SignalP method [Nielsen et al., 1997, Bendtsen et al., 2004b]. In contrast to other methods, SignalP also predicts the actual cleavage site; thus the peptide which is cleaved off during translocation over the membrane. Recently, an independent research paper has rated SignalP version 3.0 to be the best standalone tool for signal peptide prediction. It was shown that the D-score which is reported by the SignalP method is the best measure for discriminating secretory from

non-secretory proteins [Klee and Ellis, 2005].

SignalP is located at http://www.cbs.dtu.dk/services/SignalP/

What do the SignalP scores mean?

Many bioinformatics approaches or prediction tools do not give a yes/no answer. Often the user is facing an interpretation of the output, which can be either numerical or graphical. Why is that? In clear-cut examples there are no doubt; yes: this is a signal peptide! But, in borderline cases it is often convenient to have more information than just a yes/no answer. Here a graphical output can aid to interpret the correct answer. An example is shown in figure 13.38.

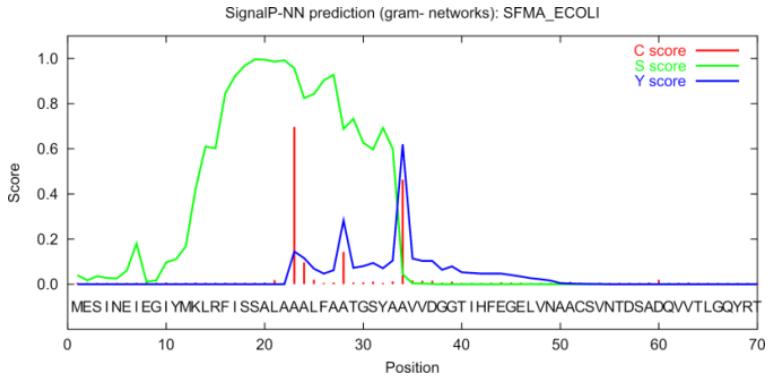


Figure 13.38: *Graphical output from the SignalP method of Swiss-Prot entry SFMA_ECOLI. Initially this seemed like a borderline prediction, but closer inspection of the sequence revealed an internal methionine at position 12, which could indicate a erroneously annotated start of the protein. Later this protein was re-annotated by Swiss-Prot to start at the M in position 12. See the text for description of the scores.*

The graphical output from SignalP (neural network) comprises three different scores, *C*, *S* and *Y*. Two additional scores are reported in the SignalP3-NN output, namely the *S-mean* and the *D-score*, but these are only reported as numerical values.

For each organism class in SignalP; Eukaryote, Gram-negative and Gram-positive, two different neural networks are used, one for predicting the actual signal peptide and one for predicting the position of the signal peptidase I (SPase I) cleavage site. The *S-score* for the signal peptide prediction is reported for every single amino acid position in the submitted sequence, with high scores indicating that the corresponding amino acid is part of a signal peptide, and low scores indicating that the amino acid is part of a mature protein.

The *C-score* is the "cleavage site" score. For each position in the submitted sequence, a C-score is reported, which should only be significantly high at the cleavage site. Confusion is often seen with the position numbering of the cleavage site. When a cleavage site position is referred to by a single number, the number indicates the first residue in the mature protein. This means that a reported cleavage site between amino acid 26-27 corresponds to the mature protein starting at (and include) position 27.

Y-max is a derivative of the C-score combined with the S-score resulting in a better cleavage site prediction than the raw C-score alone. This is due to the fact that multiple high-peaking C-scores can be found in one sequence, where only one is the true cleavage site. The cleavage site is assigned from the Y-score where the slope of the S-score is steep and a significant C-score is found.

The *S-mean* is the average of the S-score, ranging from the N-terminal amino acid to the amino acid assigned with the highest Y-max score, thus the S-mean score is calculated for the length of the predicted signal peptide. The S-mean score was in SignalP version 2.0 used as the criteria for discrimination of secretory and non-secretory proteins.

The *D-score* is introduced in SignalP version 3.0 and is a simple average of the S-mean and Y-max score. The score shows superior discrimination performance of secretory and non-secretory proteins to that of the S-mean score which was used in SignalP version 1 and 2.

For non-secretory proteins all the scores represented in the SignalP3-NN output should ideally be very low.

The hidden Markov model calculates the probability of whether the submitted sequence contains a signal peptide or not. The eukaryotic HMM model also reports the probability of a signal anchor, previously named uncleaved signal peptides. Furthermore, the cleavage site is assigned by a probability score together with scores for the n-region, h-region, and c-region of the signal peptide, if it is found.

Other useful resources

http://www.cbs.dtu.dk/services/SignalP

Pubmed entries for some of the original papers.
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=9051728&query_hl=1&itool=pubmed_docsum

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=15223320&dopt=Citation

13.10 Transmembrane helix prediction

Many proteins are integral membrane proteins. Most membrane proteins have hydrophobic regions which span the hydrophobic core of the membrane bi-layer and hydrophilic regions located on the outside or the inside of the membrane. Many receptor proteins have several transmembrane helices spanning the cellular membrane.

For prediction of transmembrane helices, *CLC Drug Discovery Workbench* uses TMHMM version 2.0 [Krogh et al., 2001] located at http://www.cbs.dtu.dk/services/TMHMM/, thus an active internet connection is required to run the transmembrane helix prediction. Additional information on THMHH and Center for Biological Sequence analysis (CBS) can be found at http://www.cbs.dtu.dk and in the original research paper [Krogh et al., 2001].

In order to use the transmembrane helix prediction, *you need to download the plugin* using the plugin manager (see section 1.7.1).

When the plugin is downloaded and installed, you can use it to predict transmembrane helices:

Toolbox | **Sequence Analysis** ()| **Transmembrane Helix Prediction** ()

If a sequence was selected before choosing the Toolbox action, this sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the selected elements.

The predictions obtained can either be shown as annotations on the sequence, in a table or as

the detailed and text output from the TMHMM method.

- Add annotations to sequence
- Create table
- Text

Click **Next** if you wish to adjust how to handle the results (see section 7.2). If not, click **Finish**.

You can perform the analysis on several protein sequences at a time. This will add annotations to all the sequences and open a view for each sequence if a transmembrane helix is found. If a transmembrane helix is not found a dialog box will be presented.

After running the prediction as described above, the protein sequence will show predicted transmembrane helices as annotations on the original sequence (see figure 13.39). Moreover, annotations showing the topology will be shown. That is, which part the proteins is located on the inside or on the outside.

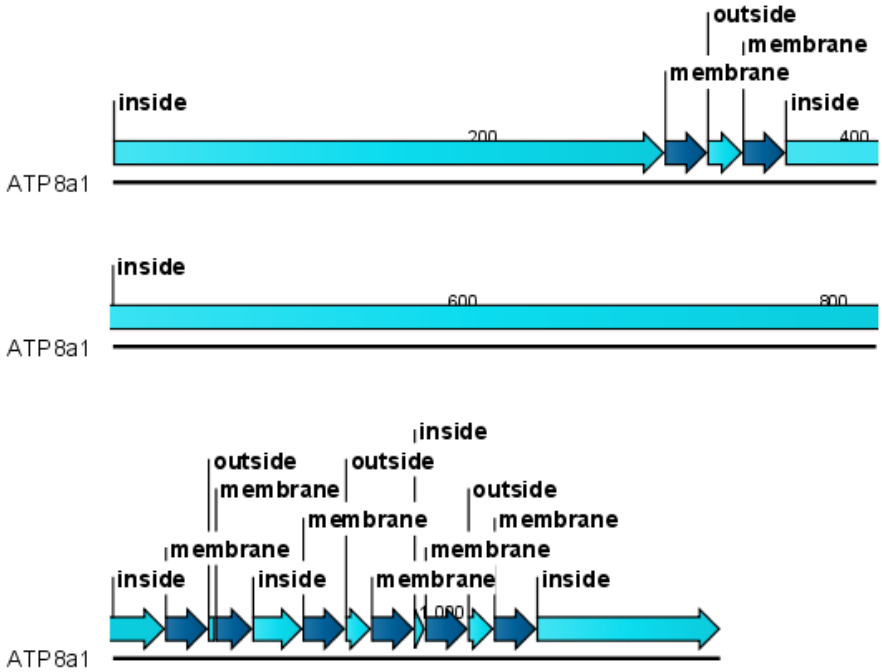


Figure 13.39: *Transmembrane segments shown as annotation on the sequence and the topology.*

Each annotation will carry a tooltip note saying that the corresponding annotation is predicted with TMHMM version 2.0. Additional notes can be added through the **Edit annotation** ( ) right-click mouse menu. See section 10.3.2.

Undesired annotations can be removed through the **Delete Annotation ( )** right-click mouse menu. See section 10.3.4.

13.11 Hydrophobicity

CLC Drug Discovery Workbench can calculate the hydrophobicity of protein sequences in different ways, using different algorithms. (See section 13.11.3). Furthermore, hydrophobicity of se-

quences can be displayed as hydrophobicity plots and as graphs along sequences. In addition, *CLC Drug Discovery Workbench* can calculate hydrophobicity for several sequences at the same time, and for alignments.

13.11.1 Hydrophobicity plot

Displaying the hydrophobicity for a protein sequence in a plot is done in the following way:

Toolbox | **Sequence Analysis** ( )| **Create Hydrophobicity Plot** ( )

This opens a dialog. The first step allows you to add or remove sequences. If you had already selected a sequence in the Navigation Area, this will be shown in the **Selected Elements**.Clicking **Next** takes you through to **Step 2**, which is displayed in figure 13.40.

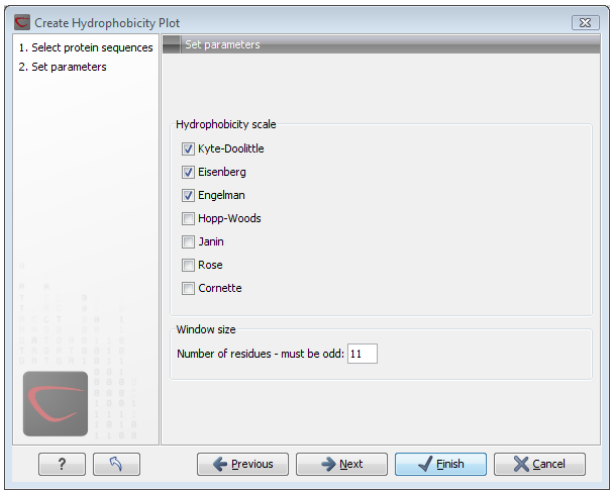


Figure 13.40: *Step two in the Hydrophobicity Plot allows you to choose hydrophobicity scale and the window size.*

The **Window size** is the width of the window where the hydrophobicity is calculated. The wider the window, the less volatile the graph. You can chose from a number of hydrophobicity scales which are further explained in section 13.11.3 Click **Next** if you wish to adjust how to handle the results (see section 7.2). If not, click **Finish**. The result can be seen in figure 13.41.

See section B in the appendix for information about the graph view.

13.11.2 Hydrophobicity graphs along sequence

Hydrophobicity graphs along sequence can be displayed easily by activating the calculations from the **Side Panel** for a sequence.

right-click protein sequence in Navigation Area | **Show** | **Sequence** | **open Protein info in Side Panel**

or **double-click protein sequence in Navigation Area** | **Show** | **Sequence** | **open Protein info in Side Panel**

These actions result in the view displayed in figure 13.42.

The level of hydrophobicity is calculated on the basis of the different scales. The different scales

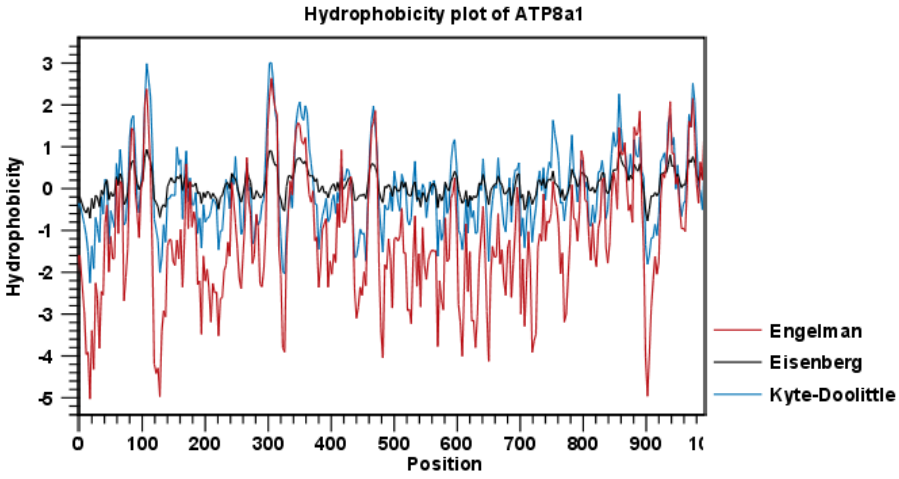


Figure 13.41: *The result of the hydrophobicity plot calculation and the associated Side Panel.*

Figure 13.42: *The different available scales in Protein info in* **CLC Drug Discovery Workbench**.

add different values to each type of amino acid. The hydrophobicity score is then calculated as the sum of the values in a 'window', which is a particular range of the sequence. The window length can be set from 5 to 25 residues. The wider the window, the less fluctuations in the hydrophobicity scores. (For more about the theory behind hydrophobicity, see 13.11.3).

In the following we will focus on the different ways that *CLC Drug Discovery Workbench* offers to display the hydrophobicity scores. We use Kyte-Doolittle to explain the display of the scores, but the different options are the same for all the scales. Initially there are three options for displaying the hydrophobicity scores. You can choose one, two or all three options by selecting the boxes. (See figure 13.43).

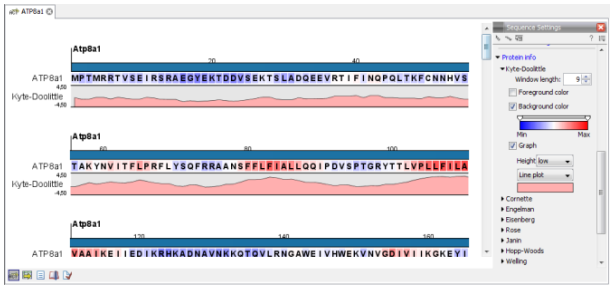


Figure 13.43: *The different ways of displaying the hydrophobicity scores, using the Kyte-Doolittle scale.*

Coloring the letters and their background. When choosing coloring of letters or coloring of their background, the color red is used to indicate high scores of hydrophobicity. A 'color-slider' allows you to amplify the scores, thereby emphasizing areas with high (or low, blue) levels of hydrophobicity. The color settings mentioned are default settings. By clicking the color bar just below the color slider you get the option of changing color settings.

Graphs along sequences. When selecting graphs, you choose to display the hydrophobicity scores underneath the sequence. This can be done either by a line-plot or bar-plot, or by coloring. The latter option offers you the same possibilities of amplifying the scores as applies for coloring of letters. The different ways to display the scores when choosing 'graphs' are displayed in figure 13.43. Notice that you can choose the height of the graphs underneath the sequence.

13.11.3 Bioinformatics explained: Protein hydrophobicity

Calculation of hydrophobicity is important to the identification of various protein features. This can be membrane spanning regions, antigenic sites, exposed loops or buried residues. Usually, these calculations are shown as a plot along the protein sequence, making it easy to identify the location of potential protein features.

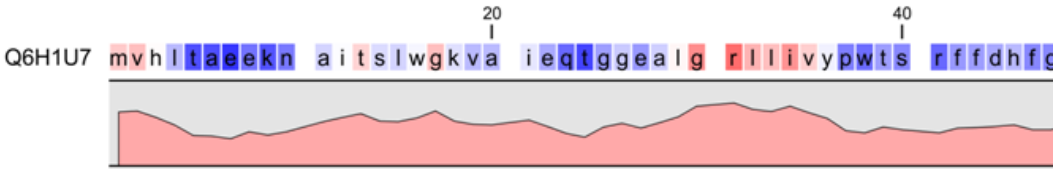


Figure 13.44: *Plot of hydrophobicity along the amino acid sequence. Hydrophobic regions on the sequence have higher numbers according to the graph below the sequence, furthermore hydrophobic regions are colored on the sequence. Red indicates regions with high hydrophobicity and blue indicates regions with low hydrophobicity.*

The hydrophobicity is calculated by sliding a fixed size window (of an odd number) over the protein sequence. At the central position of the window, the average hydrophobicity of the entire window is plotted (see figure 13.44).

Hydrophobicity scales

Several hydrophobicity scales have been published for various uses. Many of the commonly used hydrophobicity scales are described below.

Kyte-Doolittle scale. The Kyte-Doolittle scale is widely used for detecting hydrophobic regions in proteins. Regions with a positive value are hydrophobic. This scale can be used for identifying both surface-exposed regions as well as transmembrane regions, depending on the window size used. Short window sizes of 5-7 generally work well for predicting putative surface-exposed regions. Large window sizes of 19-21 are well suited for finding transmembrane domains if the values calculated are above 1.6 [Kyte and Doolittle, 1982]. These values should be used as a rule of thumb and deviations from the rule may occur.

Engelman scale. The Engelman hydrophobicity scale, also known as the GES-scale, is another scale which can be used for prediction of protein hydrophobicity [Engelman et al., 1986]. As the Kyte-Doolittle scale, this scale is useful for predicting transmembrane regions in proteins.

Eisenberg scale. The Eisenberg scale is a normalized consensus hydrophobicity scale which shares many features with the other hydrophobicity scales [Eisenberg et al., 1984].

Hopp-Woods scale. Hopp and Woods developed their hydrophobicity scale for identification of potentially antigenic sites in proteins. This scale is basically a hydrophilic index where apolar residues have been assigned negative values. Antigenic sites are likely to be predicted when using a window size of 7 [Hopp and Woods, 1983].

Cornette scale. Cornette *et al.* computed an optimal hydrophobicity scale based on 28 published scales [Cornette et al., 1987]. This optimized scale is also suitable for prediction of alpha-helices in proteins.

Rose scale. The hydrophobicity scale by Rose *et al.* is correlated to the average area of buried amino acids in globular proteins [Rose et al., 1985]. This results in a scale which is not showing the helices of a protein, but rather the surface accessibility.

Janin scale. This scale also provides information about the accessible and buried amino acid residues of globular proteins [Janin, 1979].

Welling scale. Welling et al. used information on the relative occurrence of amino acids in antigenic regions to make a scale which is useful for prediction of antigenic regions. This method is better than the Hopp-Woods scale of hydrophobicity which is also used to identify antigenic regions.

Kolaskar-Tongaonkar. A semi-empirical method for prediction of antigenic regions has been developed [Kolaskar and Tongaonkar, 1990]. This method also includes information of surface accessibility and flexibility and at the time of publication the method was able to predict antigenic determinants with an accuracy of 75%.

Surface Probability. Display of surface probability based on the algorithm by [Emini et al., 1985]. This algorithm has been used to identify antigenic determinants on the surface of proteins.

Chain Flexibility. Display of backbone chain flexibility based on the algorithm by [Karplus and Schulz, 1985]. It is known that chain flexibility is an indication of a putative antigenic determinant.

Many more scales have been published throughout the last three decades. Even though more advanced methods have been developed for prediction of membrane spanning regions, the simple and very fast calculations are still highly used.

Other useful resources

AAindex: Amino acid index database
http://www.genome.ad.jp/dbget/aaindex.html

13.12 Pfam domain search

With *CLC Drug Discovery Workbench* you can perform a search for domains in protein sequences using the Pfam database. The Pfam database [Bateman et al., 2004] at http://pfam.sanger.ac.uk/ was initially developed to aid the annotation of the *C. elegans* genome. The database is a large collection of multiple sequence alignments that cover 14831 protein domains and protein families as of March 2014. The database contains profile hidden Markov models (HMMs) for individual domain alignments, which can be used to quickly identify domains in protein sequences.

aa	aa	Kyte-Doolittle	Hopp-Woods	Cornette	Eisenberg	Rose	Janin	Engelman (GES)
A	Alanine	1.80	-0.50	0.20	0.62	0.74	0.30	1.60
C	Cysteine	2.50	-1.00	4.10	0.29	0.91	0.90	2.00
D	Aspartic acid	-3.50	3.00	-3.10	-0.90	0.62	-0.60	-9.20
E	Glutamic acid	-3.50	3.00	-1.80	-0.74	0.62	-0.70	-8.20
F	Phenylalanine	2.80	-2.50	4.40	1.19	0.88	0.50	3.70
G	Glycine	-0.40	0.00	0.00	0.48	0.72	0.30	1.00
H	Histidine	-3.20	-0.50	0.50	-0.40	0.78	-0.10	-3.00
I	Isoleucine	4.50	-1.80	4.80	1.38	0.88	0.70	3.10
K	Lysine	-3.90	3.00	-3.10	-1.50	0.52	-1.80	-8.80
L	Leucine	3.80	-1.80	5.70	1.06	0.85	0.50	2.80
M	Methionine	1.90	-1.30	4.20	0.64	0.85	0.40	3.40
N	Asparagine	-3.50	0.20	-0.50	-0.78	0.63	-0.50	-4.80
P	Proline	-1.60	0.00	-2.20	0.12	0.64	-0.30	-0.20
Q	Glutamine	-3.50	0.20	-2.80	-0.85	0.62	-0.70	-4.10
R	Arginine	-4.50	3.00	1.40	-2.53	0.64	-1.40	-12.3
S	Serine	-0.80	0.30	-0.50	-0.18	0.66	-0.10	0.60
T	Threonine	-0.70	-0.40	-1.90	-0.05	0.70	-0.20	1.20
V	Valine	4.20	-1.50	4.70	1.08	0.86	0.60	2.60
W	Tryptophan	-0.90	-3.40	1.00	0.81	0.85	0.30	1.90
Y	Tyrosine	-1.30	-2.30	3.20	0.26	0.76	-0.40	-0.70

Table 13.3: *Hydrophobicity scales. This table shows seven different hydrophobicity scales which are generally used for prediction of e.g. transmembrane regions and antigenicity.*

Many proteins have a unique combination of domains, which can be responsible for e.g. the catalytic activities of enzymes. Annotating sequences based on pairwise alignment methods by simply transferring annotation from a known protein to the unknown partner does not take domain organization into account [Galperin and Koonin, 1998]. For example, a protein may be annotated incorrectly as an enzyme if the pairwise alignment only finds a regulatory domain.

Using the **Pfam Domain Search** tool in *CLC Drug Discovery Workbench*, you can search for domains in sequence data which otherwise do not carry any annotation information. The domain search is performed using the hmmsearch tool from the HMMER3 package version 3.1b1 (`http://hmmer.janelia.org/`). The Pfam search tool annotates protein sequences with all domains in the Pfam database that have a significant match. It is possible to lower the significance cutoff thresholds in the hmmsearch algorithm, which will reduce the number of domain annotations. Individual domain annotations can be removed manually as described in section 10.3.4.

13.12.1 Download of Pfam database

To be able to run the **Pfam Domain Search** tool you must first download the Pfam database. The Pfam database can be downloaded using:

Toolbox | **Sequence Analysis** ( )| **Download Pfam Database** ( )

Specify where you would like to save the downloaded Pfam database. The output of the **Download Pfam Database** tool is a database object, which can be selected as a parameter for the Pfam Domain Search tool. It doesn't really make sense to try to open the database object directly from the **Navigation Area** as all you can see directly is the element history (which version of the Workbench that has been used and the name of the downloaded files) and the element info, which in this case only provides information about the database name.

13.12.2 Running Pfam Domain Search

When you have downloaded the Pfam database you are ready to perform a Pfam domain search. To do this start the Pfam search tool:

Toolbox | **Sequence Analysis (**  **)**| **Pfam Domain Search (****)**

If a sequence was selected before choosing the Toolbox action, this sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the selected elements.

You can perform the analysis on several protein sequences at a time. This will add annotations to all the sequences. Click **Next** to adjust parameters (see figure 13.45).

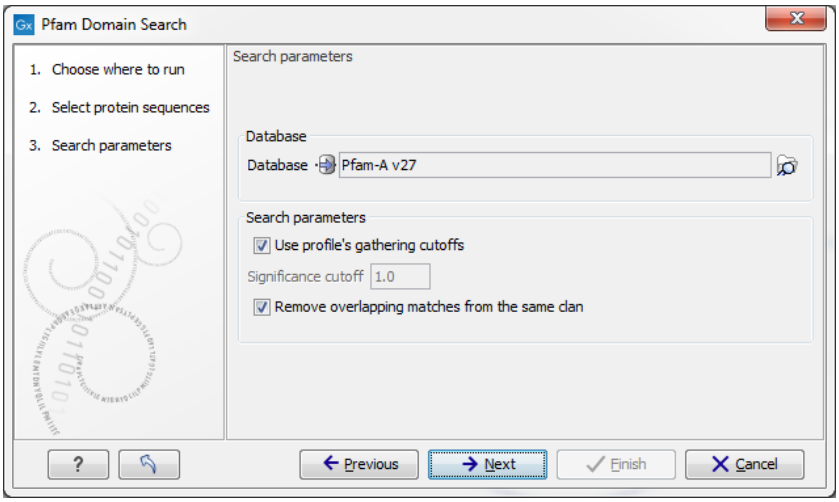


Figure 13.45: *Setting parameters for Pfam Domain Search.*

- **Database.** Choose which database to use when searching for Pfam domains. For information on how to download a Pfam database see section 13.12.1
- Significance cutoff
  - **Use profile's gathering cutoffs.** Use cutoffs specifically assigned to each family by the curator instead of manually assigning the **Significance cutoff**.
  - **Significance cutoff.** The E-value (expectation value) describes the number of hits one would expect to see by chance when searching a database of a particular size. Essentially, a hit with a low E-value is more significant compared to a hit with a high E-value. By lowering the significance threshold the domain search will become more specific and less sensitive, i.e. fewer hits will be reported but the reported hits will be more significant on average.
- **Remove overlapping matches from the same clan.** Perform post-processing of the results where overlaps between hits are resolved by keeping the hit with the smallest e-value.

Click **Next** to adjust the output of the tool. The Pfam search tool can produce two types of output. It can add annotations on the input sequences that show the domains found (see figure 13.46) and it can output a table with all the domains found.

Click **Next** if you wish to adjust how to handle the results (see section 7.2). If not, click **Finish**.

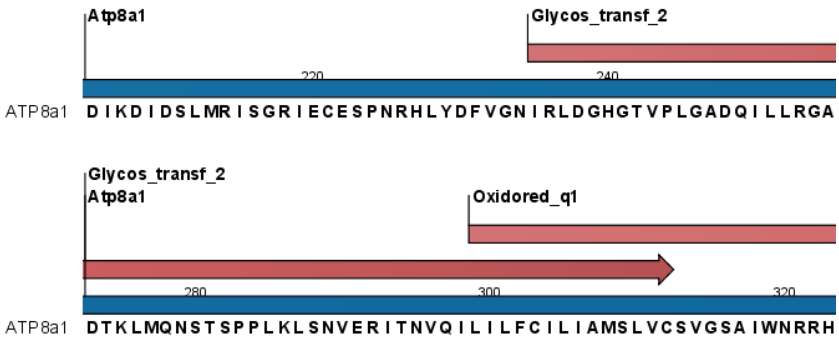


Figure 13.46: *Annotations (in red) that were added by the Pfam search tool.*

Domain annotations added by the Pfam search tool have the type **Region**. If the annotations are not visible they have to be enabled in the side panel. Detailed information for each domain annotation, such as the bit score which is the basis for the prediction of domains, is available through the annotation tool tip.

A more detailed description of the scores provided in the annotation tool tips can be found here: http://pfam.sanger.ac.uk/help#tabview=tab5.

13.13 Secondary structure prediction

An important issue when trying to understand protein function is to know the actual structure of the protein. Many questions that are raised by molecular biologists are directly targeted at protein structure. The alpha-helix forms a coiled rod like structure whereas a beta-sheet show an extended sheet-like structure. Some proteins are almost devoid of alpha-helices such as chymotrypsin (PDB_ID: 1AB9) whereas others like myoglobin (PDB_ID: 101M) have a very high content of alpha-helices.

With *CLC Drug Discovery Workbench* one can predict the secondary structure of proteins very fast. Predicted elements are alpha-helix, beta-sheet (same as beta-strand) and other regions.

Based on extracted protein sequences from the protein databank (http://www.rcsb.org/pdb/) a hidden Markov model (HMM) was trained and evaluated for performance. Machine learning methods have shown superior when it comes to prediction of secondary structure of proteins [Rost, 2001]. By far the most common structures are Alpha-helices and beta-sheets which can be predicted, and predicted structures are automatically added to the query as annotation which later can be edited.

In order to predict the secondary structure of proteins:

Toolbox | **Sequence Analysis** ( )| **Predict secondary structure** ( )

This opens the dialog displayed in figure 13.47:

If a sequence was selected before choosing the Toolbox action, this sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the selected elements.

You can perform the analysis on several protein sequences at a time. This will add annotations to all the sequences and open a view for each sequence.

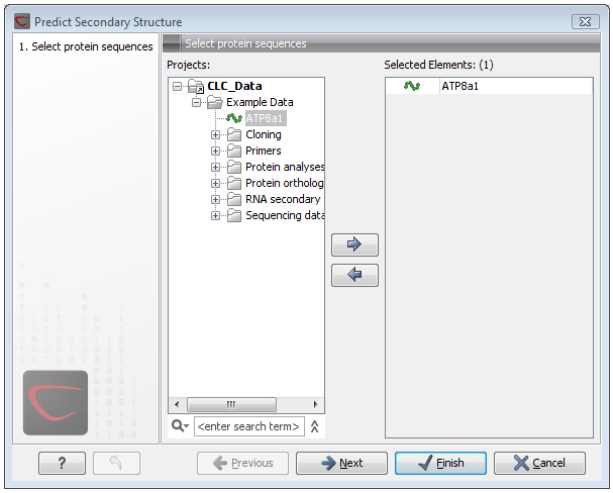


Figure 13.47: *Choosing one or more protein sequences for secondary structure prediction.*

Click **Next** if you wish to adjust how to handle the results (see section 7.2). If not, click **Finish**.

After running the prediction as described above, the protein sequence will show predicted alpha-helices and beta-sheets as annotations on the original sequence (see figure 13.48).

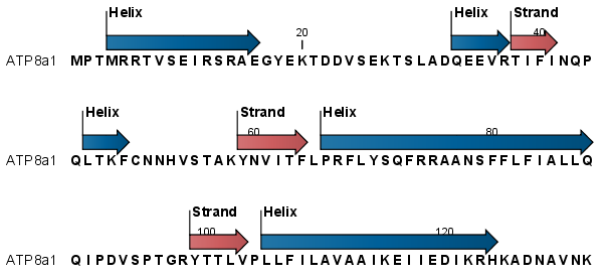


Figure 13.48: *Alpha-helices and beta-strands shown as annotations on the sequence.*

Each annotation will carry a tooltip note saying that the corresponding annotation is predicted with *CLC Drug Discovery Workbench*. Additional notes can be added through the **Edit Annotation** ( ) right-click mouse menu. See section 10.3.2.

Undesired alpha-helices or beta-sheets can be removed through the **Delete Annotation** ( ) right-click mouse menu. See section 10.3.4.

Chapter 14

Sequence alignment

Contents

CLC Drug Discovery Workbench can align nucleotides and proteins using a *progressive alignment* algorithm (see section 14.5 or read the White paper on alignments in the **Science** section of http://www.clcbio.com).
This chapter describes how to use the program to align sequences. The chapter also describes alignment algorithms in more general terms.

14.1 Create an alignment

Alignments can be created from sequences, sequence lists (see section 10.6), existing alignments and from any combination of the three.

To create an alignment in *CLC Drug Discovery Workbench*:

Toolbox | **Sequence Alignment ( )**| **Create Alignment ( )**

This opens the dialog shown in figure 14.1.

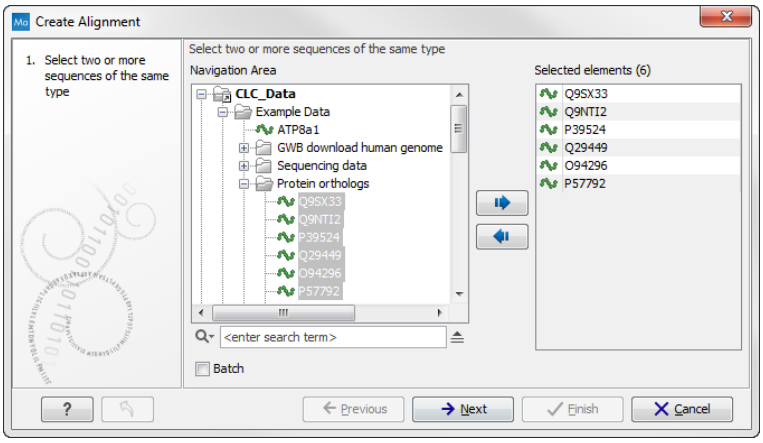


Figure 14.1: *Creating an alignment.*

If you have selected some elements before choosing the Toolbox action, they are now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences, sequence lists or alignments from the selected elements. Click **Next** to adjust alignment algorithm parameters. Clicking **Next** opens the dialog shown in figure 14.2.

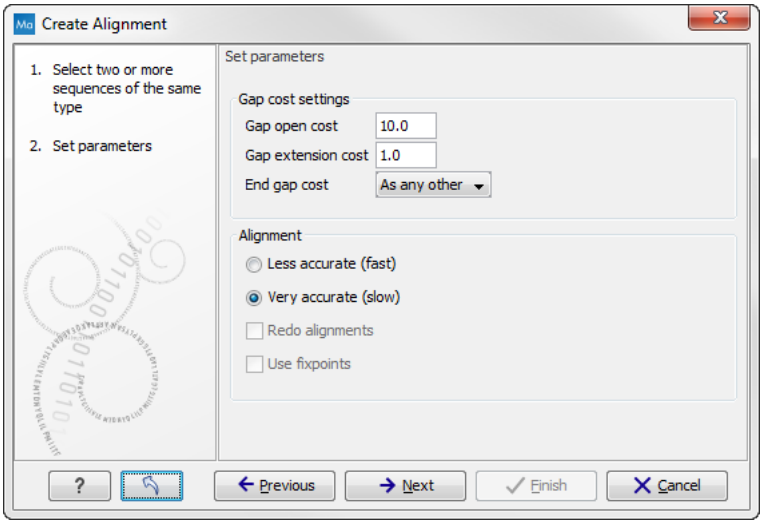


Figure 14.2: *Adjusting alignment algorithm parameters.*

14.1.1 Gap costs

The alignment algorithm has three parameters concerning gap costs: Gap open cost, Gap extension cost and End gap cost. The precision of these parameters is to one place of decimal.

- **Gap open cost**. The price for introducing gaps in an alignment.
- **Gap extension cost**. The price for every extension past the initial gap.

If you expect a lot of small gaps in your alignment, the Gap open cost should equal the Gap extension cost. On the other hand, if you expect few but large gaps, the Gap open cost should be set significantly higher than the Gap extension cost.

However, for most alignments it is a good idea to make the Gap open cost quite a bit higher than the Gap extension cost. The default values are 10.0 and 1.0 for the two parameters, respectively.

- **End gap cost**. The price of gaps at the beginning or the end of the alignment. One of the advantages of the *CLC Drug Discovery Workbench* alignment method is that it provides flexibility in the treatment of gaps at the ends of the sequences. There are three possibilities:
  - **Free end gaps**. Any number of gaps can be inserted in the ends of the sequences without any cost.
  - **Cheap end gaps**. All end gaps are treated as gap extensions and any gaps past 10 are free.
  - **End gaps as any other**. Gaps at the ends of sequences are treated like gaps in any other place in the sequences.

When aligning a long sequence with a short partial sequence, it is ideal to use free end gaps, since this will be the best approximation to the situation. The many gaps inserted at the ends are not due to evolutionary events, but rather to partial data.

Many homologous proteins have quite different ends, often with large insertions or deletions. This confuses alignment algorithms, but using the **Cheap end gaps** option, large gaps will generally be tolerated at the sequence ends, improving the overall alignment. This is the default setting of the algorithm.

Finally, treating end gaps like any other gaps is the best option when you know that there are no biologically distinct effects at the ends of the sequences.

Figures 14.3 and 14.4 illustrate the differences between the different gap scores at the sequence ends.

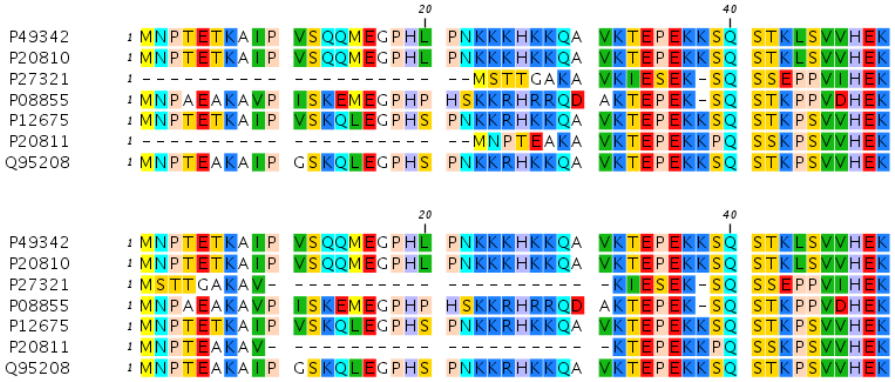


Figure 14.3: *The first 50 positions of two different alignments of seven calpastatin sequences. The top alignment is made with cheap end gaps, while the bottom alignment is made with end gaps having the same price as any other gaps. In this case it seems that the latter scoring scheme gives the best result.*

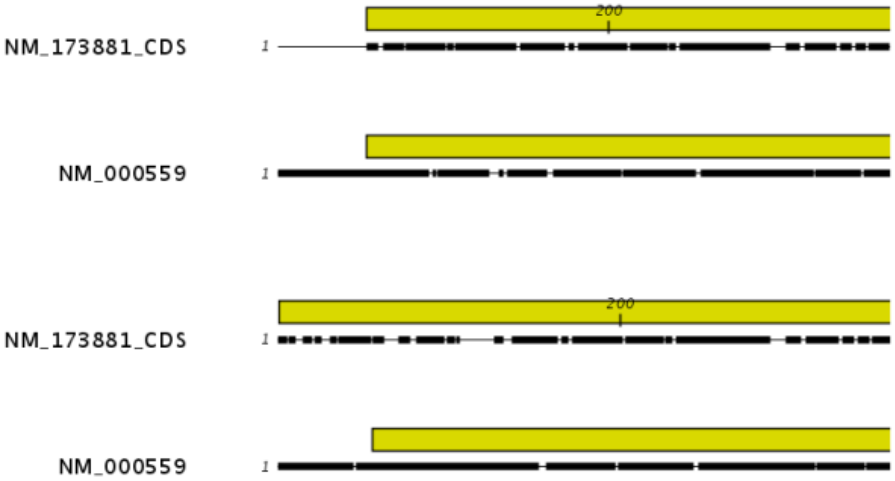


Figure 14.4: *The alignment of the coding sequence of bovine myoglobin with the full mRNA of human gamma globin. The top alignment is made with free end gaps, while the bottom alignment is made with end gaps treated as any other. The yellow annotation is the coding sequence in both sequences. It is evident that free end gaps are ideal in this situation as the start codons are aligned correctly in the top alignment. Treating end gaps as any other gaps in the case of aligning distant homologs where one sequence is partial leads to a spreading out of the short sequence as in the bottom alignment.*

14.1.2 Fast or accurate alignment algorithm

CLC Drug Discovery Workbench has two algorithms for calculating alignments:

- **Fast (less accurate).** This allows for use of an optimized alignment algorithm which is very fast. The fast option is particularly useful for data sets with very long sequences.
- **Slow (very accurate).** This is the recommended choice unless you find the processing time too long.

Both algorithms use progressive alignment. The faster algorithm builds the initial tree by doing more approximate pairwise alignments than the slower option.

14.1.3 Aligning alignments

If you have selected an existing alignment in the first step (14.1), you have to decide how this alignment should be treated.

- **Redo alignment.** The original alignment will be realigned if this checkbox is checked. Otherwise, the original alignment is kept in its original form except for possible extra equally sized gaps in all sequences of the original alignment. This is visualized in figure 14.5.

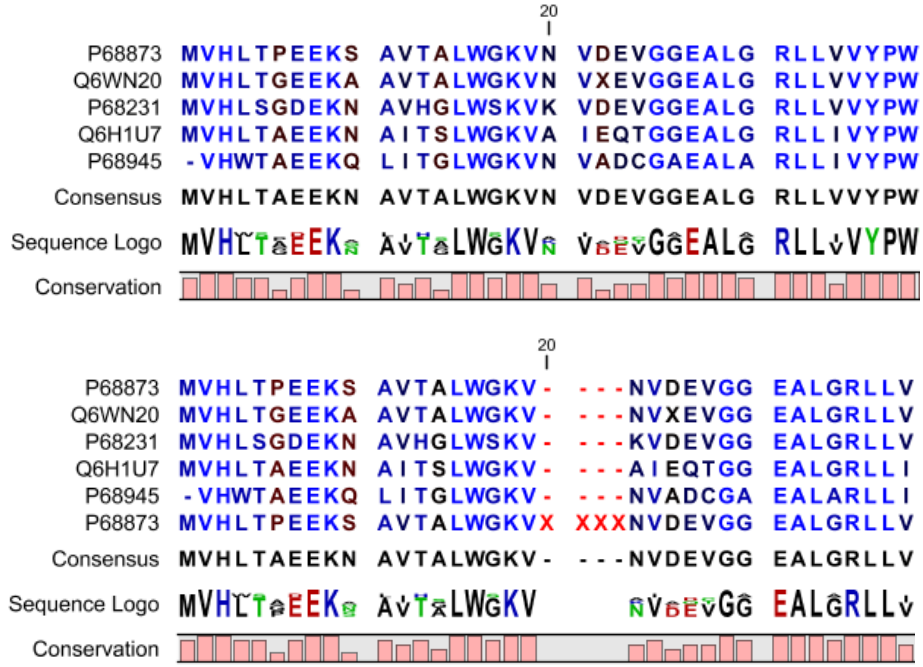


Figure 14.5: *The top figures shows the original alignment. In the bottom panel a single sequence with four inserted X's are aligned to the original alignment. This introduces gaps in all sequences of the original alignment. All other positions in the original alignment are fixed.*

This feature is useful if you wish to add extra sequences to an existing alignment, in which case you just select the alignment and the extra sequences and choose not to redo the alignment.

It is also useful if you have created an alignment where the gaps are not placed correctly. In this case, you can realign the alignment with different gap cost parameters.

14.1.4 Fixpoints

With fixpoints, you can get full control over the alignment algorithm. The fixpoints are points on the sequences that are forced to align to each other.

To add a fixpoint, open the sequence or alignment and:

> **Select the region you want to use as a fixpoint | right-click the selection | Set alignment fixpoint here**

This will add an annotation labeled "Fixpoint" to the sequence (see figure 14.6). Use this procedure to add fixpoints to the other sequence(s) that should be forced to align to each other.

Figure 14.6: *Adding a fixpoint to a sequence in an existing alignment. At the top you can see a fixpoint that has already been added.*

When you click "Create alignment" and go to **Step 2**, check **Use fixpoints** in order to force the alignment algorithm to align the fixpoints in the selected sequences to each other.

In figure 14.7 the result of an alignment using fixpoints is illustrated.

You can add multiple fixpoints, e.g. adding two fixpoints to the sequences that are aligned will force their first fixpoints to be aligned to each other, and their second fixpoints will also be aligned to each other.

Advanced use of fixpoints

Fixpoints with the same names will be aligned to each other, which gives the opportunity for great control over the alignment process. It is only necessary to change any fixpoint names in very special cases.

One example would be three sequences A, B and C where sequences A and B has one copy of a domain while sequence C has two copies of the domain. You can now force sequence A to align to the first copy and sequence B to align to the second copy of the domains in sequence C. This is done by inserting fixpoints in sequence C for each domain, and naming them 'fp1' and 'fp2' (for example). Now, you can insert a fixpoint in each of sequences A and B, naming them 'fp1' and 'fp2', respectively. Now, when aligning the three sequences using fixpoints, sequence A will

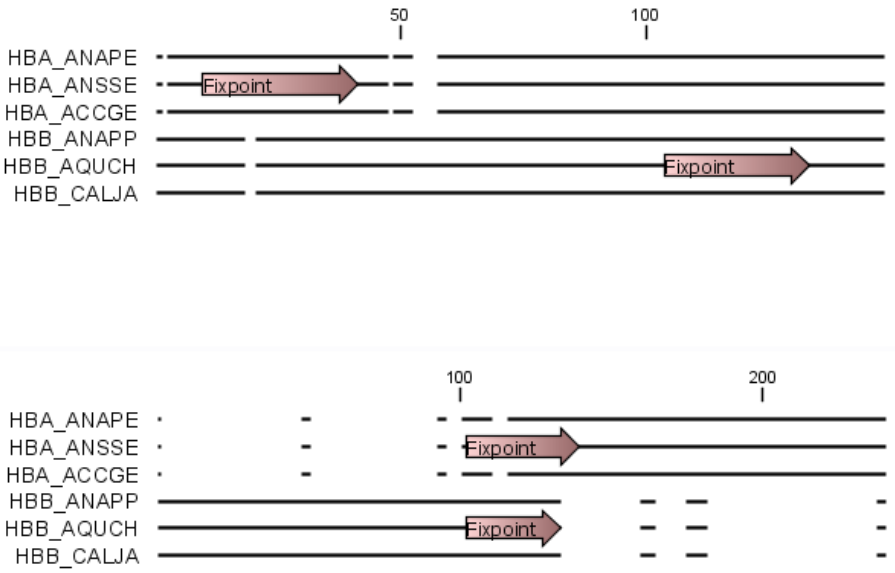


Figure 14.7: *Realigning using fixpoints. In the top view, fixpoints have been added to two of the sequences. In the view below, the alignment has been realigned using the fixpoints. The three top sequences are very similar, and therefore they follow the one sequence (number two from the top) that has a fixpoint.*

align to the first copy of the domain in sequence C, while sequence B would align to the second copy of the domain in sequence C.

You can name fixpoints by:

> **right-click the Fixpoint annotation | Edit Annotation ( ) | type the name in the 'Name' field**

14.2 View alignments

Since an alignment is a display of several sequences arranged in rows, the basic options for viewing alignments are the same as for viewing sequences. Therefore we refer to section 10.1 for an explanation of these basic options.

However, there are a number of alignment-specific view options in the **Alignment info** and the **Nucleotide info** in the **Side Panel** to the right of the view. Below is more information on these view options.

Under **Translation** in the **Nucleotide info**, there is an extra checkbox: **Relative to top sequence**. Checking this box will make the reading frames for the translation align with the top sequence so that you can compare the effect of nucleotide differences on the protein level.

The options in the **Alignment info** relate to each column in the alignment:

- **Consensus.** Shows a consensus sequence at the bottom of the alignment. The consensus sequence is based on every single position in the alignment and reflects an artificial sequence which resembles the sequence information of the alignment, but only as one single sequence. If all sequences of the alignment is 100% identical the consensus sequence will be identical to all sequences found in the alignment. If the sequences of the alignment differ the consensus sequence will reflect the most common sequences in the

alignment. Parameters for adjusting the consensus sequences are described below.

- **Limit.** This option determines how conserved the sequences must be in order to agree on a consensus. Here you can also choose **IUPAC** which will display the ambiguity code when there are differences between the sequences. E.g. an alignment with **A** and a **G** at the same position will display an **R** in the consensus line if the **IUPAC** option is selected. (The IUPAC codes can be found in section E and D.) Please note that the IUPAC codes are only available for nucleotide alignments.
- **No gaps.** Checking this option will not show gaps in the consensus.
- **Ambiguous symbol.** Select how ambiguities should be displayed in the consensus line (as **N**, **?**, *****, **.** or **-**). This option has no effect if **IUPAC** is selected in the **Limit** list above.

The **Consensus Sequence** can be opened in a new view, simply by right-clicking the **Consensus Sequence** and click **Open Consensus in New View**.

- **Conservation.** Displays the level of conservation at each position in the alignment. The conservation shows the conservation of all sequence positions. The height of the bar, or the gradient of the color reflect how conserved that particular position is in the alignment. If one position is 100% conserved the bar will be shown in full height, and it is colored in the color specified at the right side of the gradient slider.
  - **Foreground color.** Colors the letters using a gradient, where the right side color is used for highly conserved positions and the left side color is used for positions that are less conserved.
  - **Background color.** Sets a background color of the residues using a gradient in the same way as described above.
  - **Graph.** Displays the conservation level as a graph at the bottom of the alignment. The bar (default view) show the conservation of all sequence positions. The height of the graph reflects how conserved that particular position is in the alignment. If one position is 100% conserved the graph will be shown in full height. Learn how to export the data behind the graph in section 6.5.
    - **Height.** Specifies the height of the graph.
    - **Type.** The type of the graph.
      - **Line plot.** Displays the graph as a line plot.
      - **Bar plot.** Displays the graph as a bar plot.
      - **Colors.** Displays the graph as a color bar using a gradient like the foreground and background colors.
    - **Color box.** Specifies the color of the graph for line and bar plots, and specifies a gradient for colors.
- **Gap fraction.** Which fraction of the sequences in the alignment that have gaps. The gap fraction is only relevant if there are gaps in the alignment.
  - **Foreground color.** Colors the letter using a gradient, where the left side color is used if there are relatively few gaps, and the right side color is used if there are relatively many gaps.
  - **Background color.** Sets a background color of the residues using a gradient in the same way as described above.

  - **Graph.** Displays the gap fraction as a graph at the bottom of the alignment (Learn how to export the data behind the graph in section 6.5).
    * **Height.** Specifies the height of the graph.
    * **Type.** The type of the graph.
      · **Line plot.** Displays the graph as a line plot.
      · **Bar plot.** Displays the graph as a line plot.
      · **Colors.** Displays the graph as a color bar using a gradient like the foreground and background colors.
    * **Color box.** Specifies the color of the graph for line and bar plots, and specifies a gradient for colors.

- **Color different residues.** Indicates differences in aligned residues.
  - **Foreground color.** Colors the letter.
  - **Background color.** Sets a background color of the residues.

- **Sequence logo.** A sequence logo displays the frequencies of residues at each position in an alignment. This is presented as the relative heights of letters, along with the degree of sequence conservation as the total height of a stack of letters, measured in bits of information. The vertical scale is in bits, with a maximum of 2 bits for nucleotides and approximately 4.32 bits for amino acid residues. See section 14.2.1 for more details.
  - **Foreground color.** Color the residues using a gradient according to the information content of the alignment column. Low values indicate columns with high variability whereas high values indicate columns with similar residues.
  - **Background color.** Sets a background color of the residues using a gradient in the same way as described above.
  - **Logo.** Displays sequence logo at the bottom of the alignment.
    * **Height.** Specifies the height of the sequence logo graph.
    * **Color.** The sequence logo can be displayed in black or Rasmol colors. For protein alignments, a polarity color scheme is also available, where hydrophobic residues are shown in black color, hydrophilic residues as green, acidic residues as red and basic residues as blue.

14.2.1 Bioinformatics explained: Sequence logo

In the search for homologous sequences, researchers are often interested in conserved sites/residues or positions in a sequence which tend to differ a lot. Most researches use alignments (see Bioinformatics explained*: multiple alignments*) for visualization of homology on a given set of either DNA or protein sequences. In proteins, active sites in a given protein family are often highly conserved. Thus, in an alignment these positions (which are not necessarily located in proximity) are fully or nearly fully conserved. On the other hand, antigen binding sites in the F_{ab} unit of immunoglobulins tend to differ quite a lot, whereas the rest of the protein remains relatively unchanged.

In DNA, promoter sites or other DNA binding sites are highly conserved (see figure 14.8). This is also the case for repressor sites as seen for the Cro repressor of bacteriophage λ.

When aligning such sequences, regardless of whether they are highly variable or highly conserved at specific sites, it is very difficult to generate a consensus sequence which covers the actual variability of a given position. In order to better understand the information content or significance of certain positions, a sequence logo can be used. The sequence logo displays the information content of all positions in an alignment as residues or nucleotides stacked on top of each other (see figure 14.8). The sequence logo provides a far more detailed view of the entire alignment than a simple consensus sequence. Sequence logos can aid to identify protein binding sites on DNA sequences and can also aid to identify conserved residues in aligned domains of protein sequences and a wide range of other applications.

Each position of the alignment and consequently the sequence logo shows the sequence information in a computed score based on Shannon entropy [Schneider and Stephens, 1990]. The height of the individual letters represent the sequence information content in that particular position of the alignment.

A sequence logo is a much better visualization tool than a simple consensus sequence. An example hereof is an alignment where in one position a particular residue is found in 70% of the sequences. If a consensus sequence is used, it typically only displays the single residue with 70% coverage. In figure 14.8 an un-gapped alignment of 11 *E. coli* start codons including flanking regions are shown. In this example, a consensus sequence would only display ATG as the start codon in position 1, but when looking at the sequence logo it is seen that a GTG is also allowed as a start codon.

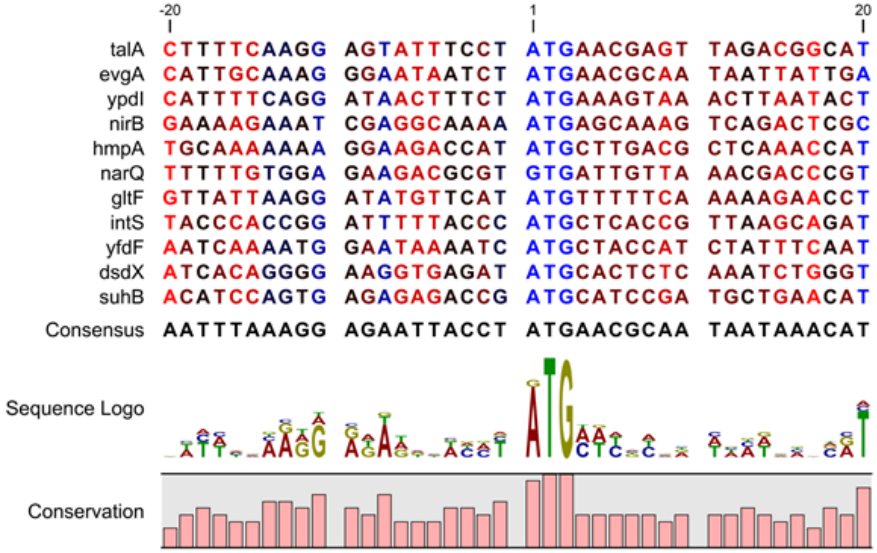


Figure 14.8: *Ungapped sequence alignment of eleven* E. coli *sequences defining a start codon. The start codons start at position 1. Below the alignment is shown the corresponding sequence logo. As seen, a GTG start codon and the usual ATG start codons are present in the alignment. This can also be visualized in the logo at position 1.*

Calculation of sequence logos

A comprehensive walk-through of the calculation of the information content in sequence logos is beyond the scope of this document but can be found in the original paper by [Schneider and Stephens, 1990]. Nevertheless, the conservation of every position is defined as R_{seq} which is the difference between the maximal entropy (S_{max}) and the observed entropy for the residue

distribution (S_{obs}),

$$R_{seq} = S_{max} - S_{obs} = \log_2 N - \left(-\sum_{n=1}^{N} p_n \log_2 p_n \right)$$

p_n is the observed frequency of a amino acid residue or nucleotide of symbol n at a particular position and N is the number of distinct symbols for the sequence alphabet, either 20 for proteins or four for DNA/RNA. This means that the maximal sequence information content per position is $\log_2 4 = 2\ bits$ for DNA/RNA and $\log_2 20 \approx 4.32\ bits$ for proteins.

The original implementation by Schneider does not handle sequence gaps.

We have slightly modified the algorithm so an estimated logo is presented in areas with sequence gaps.

If amino acid residues or nucleotides of one sequence are found in an area containing gaps, we have chosen to show the particular residue as the fraction of the sequences. Example; if one position in the alignment contain 9 gaps and only one alanine (A) the A represented in the logo has a hight of 0.1.

Other useful resources
The website of Tom Schneider
http://www-lmmb.ncifcrf.gov/~toms/

WebLogo
http://weblogo.berkeley.edu/

[Crooks et al., 2004]

14.3 Edit alignments

14.3.1 Move residues and gaps

The placement of gaps in the alignment can be changed by modifying the parameters when creating the alignment (see section 14.1). However, gaps and residues can also be moved after the alignment is created:

select one or more gaps or residues in the alignment | drag the selection to move

This can be done both for single sequences, but also for multiple sequences by making a selection covering more than one sequence. When you have made the selection, the mouse pointer turns into a horizontal arrow indicating that the selection can be moved (see figure 14.9).

Note! Residues can only be moved when they are next to a gap.

14.3.2 Insert gaps

The placement of gaps in the alignment can be changed by modifying the parameters when creating the alignment. However, gaps can also be added manually after the alignment is created.

To insert extra gaps:

select a part of the alignment | right-click the selection | Add gaps before/after

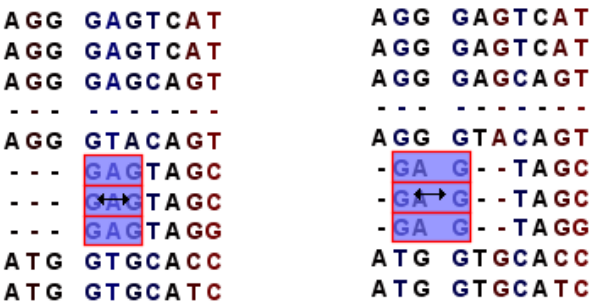


Figure 14.9: *Moving a part of an alignment. Notice the change of mouse pointer to a horizontal arrow.*

If you have made a selection covering e.g. five residues, a gap of five will be inserted. In this way you can easily control the number of gaps to insert. Gaps will be inserted in the sequences that you selected. If you make a selection in two sequences in an alignment, gaps will be inserted into these two sequences. This means that these two sequences will be displaced compared to the other sequences in the alignment.

14.3.3 Delete residues and gaps

Residues or gaps can be deleted for individual sequences or for the whole alignment. For individual sequences:

> **select the part of the sequence you want to delete** | **right-click the selection** | **Edit Selection (**  **)** | **Delete the text in the dialog** | **Replace**

The selection shown in the dialog will be replaced by the text you enter. If you delete the text, the selection will be replaced by an empty text, i.e. deleted.

In order to delete entire columns:

> **manually select the columns to delete** | **right-click the selection** | **click 'Delete Selection'**

14.3.4 Copy annotations to other sequences

Annotations on one sequence can be transferred to other sequences in the alignment:

> **right-click the annotation** | **Copy Annotation to other Sequences**

This will display a dialog listing all the sequences in the alignment. Next to each sequence is a checkbox which is used for selecting which sequences, the annotation should be copied to. Click **Copy** to copy the annotation.

If you wish to copy all annotations on the sequence, click the **Copy All Annotations to other Sequences**.

Copied/transferred annotations will contain the same qualifier text as the original. That is, the text is not updated. As an example, if the annotation contains 'translation' as qualifier text this translation will be copied to the new sequence and will thus reflect the translation of the original sequence, not the new sequence, which may differ.

14.3.5 Move sequences up and down

Sequences can be moved up and down in the alignment:

drag the name of the sequence up or down

When you move the mouse pointer over the label, the pointer will turn into a vertical arrow indicating that the sequence can be moved.

The sequences can also be sorted automatically to let you save time moving the sequences around. To sort the sequences alphabetically:

Right-click the name of a sequence | **Sort Sequences Alphabetically**

If you change the Sequence name (in the **Sequence Layout** view preferences), you will have to ask the program to sort the sequences again.

If you have one particular sequence that you would like to use as a reference sequence, it can be useful to move this to the top. This can be done manually, but it can also be done automatically:

Right-click the name of a sequence | **Move Sequence to Top**

The sequences can also be sorted by similarity, grouping similar sequences together:

Right-click the name of a sequence | **Sort Sequences by Similarity**

14.3.6 Delete and rename sequences

14.3.7 Delete, rename and add sequences

Sequences can be removed from the alignment by right-clicking the label of a sequence:

right-click label | **Delete Sequence**

This can be undone by clicking **Undo** () in the Toolbar.

If you wish to delete several sequences, you can check all the sequences, right-click and choose **Delete Marked Sequences**. To show the checkboxes, you first have to click the **Show Selection Boxes** in the **Side Panel**.

A sequence can also be renamed:

right-click label | **Rename Sequence**

This will show a dialog, letting you rename the sequence. This will not affect the sequence that the alignment is based on.

Extra sequences can be added to the alignment by creating a new alignment where you select the current alignment and the extra sequences (see section 14.1).

The same procedure can be used for joining two alignments.

14.3.8 Realign selection

If you have created an alignment, it is possible to realign a part of it, leaving the rest of the alignment unchanged:

select a part of the alignment to realign | **right-click the selection** | **Realign selection**

This will open **Step 2** in the "Create alignment" dialog, allowing you to set the parameters for the realignment (see section 14.1).

It is possible for an alignment to become shorter or longer as a result of the realignment of a region. This is because gaps may have to be inserted in, or deleted from, the sequences not selected for realignment. This will only occur for entire columns of gaps in these sequences, ensuring that their relative alignment is unchanged.

Realigning a selection is a very powerful tool for editing alignments in several situations:

- **Removing changes.** If you change the alignment in a specific region by hand, you may end up being unhappy with the result. In this case you may of course undo your edits, but another option is to select the region and realign it.
- **Adjusting the number of gaps.** If you have a region in an alignment which has too many gaps in your opinion, you can select the region and realign it. By choosing a relatively high gap cost you will be able to reduce the number of gaps.
- **Combine with fixpoints.** If you have an alignment where two residues are not aligned, but you know that they should have been. You can now set an alignment fixpoint on each of the two residues, select the region and realign it using the fixpoints. Now, the two residues are aligned with each other and everything in the selected region around them is adjusted to accommodate this change.

14.4 Pairwise comparison

For a given set of aligned sequences it is possible to make a pairwise comparison in which each pair of sequences are compared to each other. This provides an overview of the diversity among the sequences in the alignment.

In *CLC Drug Discovery Workbench* this is done by creating a comparison table:

Toolbox | **Sequence Alignment** ( )| **Create Pairwise Comparison** ( )

This opens the dialog displayed in figure 14.10:

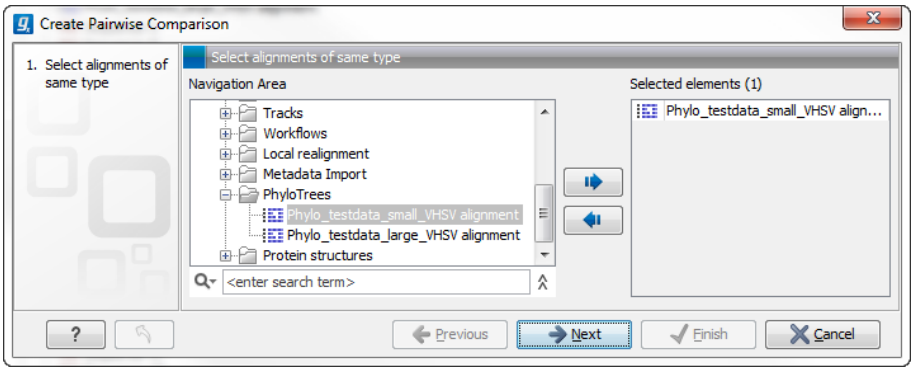


Figure 14.10: *Creating a pairwise comparison table.*

If an alignment was selected before choosing the Toolbox action, this alignment is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove elements from the **Navigation Area**. Click **Next** to adjust parameters.

14.4.1 Pairwise comparison on alignment selection

A pairwise comparison can also be performed for a selected part of an alignment:

right-click on an alignment selection | Pairwise Comparison ()

This leads directly to the dialog described in the next section.

14.4.2 Pairwise comparison parameters

There are five kinds of comparison that can be made between the sequences in the alignment, as shown in figure 14.11.

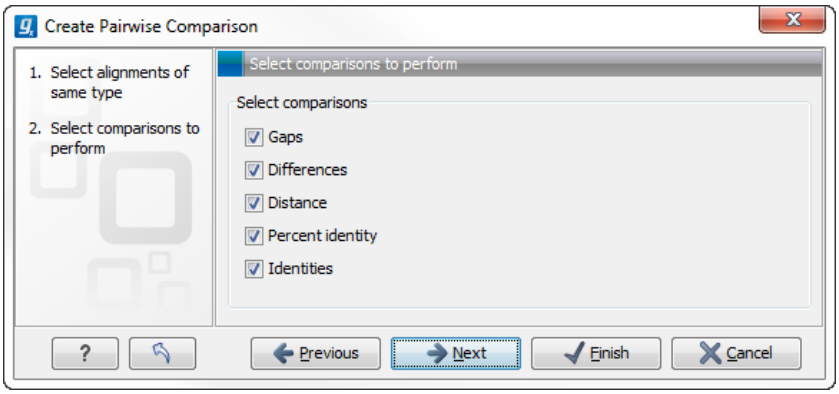


Figure 14.11: *Adjusting parameters for pairwise comparison.*

- **Gaps** Calculates the number of alignment positions where one sequence has a gap and the other does not.
- **Identities** Calculates the number of identical alignment positions to overlapping alignment positions between the two sequences. An overlapping alignment position is a position where at least one residue is present, rather than only gaps.
- **Differences** Calculates the number of alignment positions where one sequence is different from the other. This includes gap differences as in the Gaps comparison.
- **Distance** Calculates the Jukes-Cantor distance between the two sequences. This number is given as the Jukes-Cantor correction of the proportion between identical and overlapping alignment positions between the two sequences.
- **Percent identity** Calculates the percentage of identical residues in alignment positions to overlapping alignment positions between the two sequences.

14.4.3 The pairwise comparison table

The table shows the results of selected comparisons (see an example in figure 14.12). Since comparisons are often symmetric, the table can show the results of two comparisons at the same time, one in the upper-right and one in the lower-left triangle.

Note that you can change the minimum and maximum values of the gradient coloring by sliding the corresponding cursor along the gradient in the right side panel of the comparison table. The

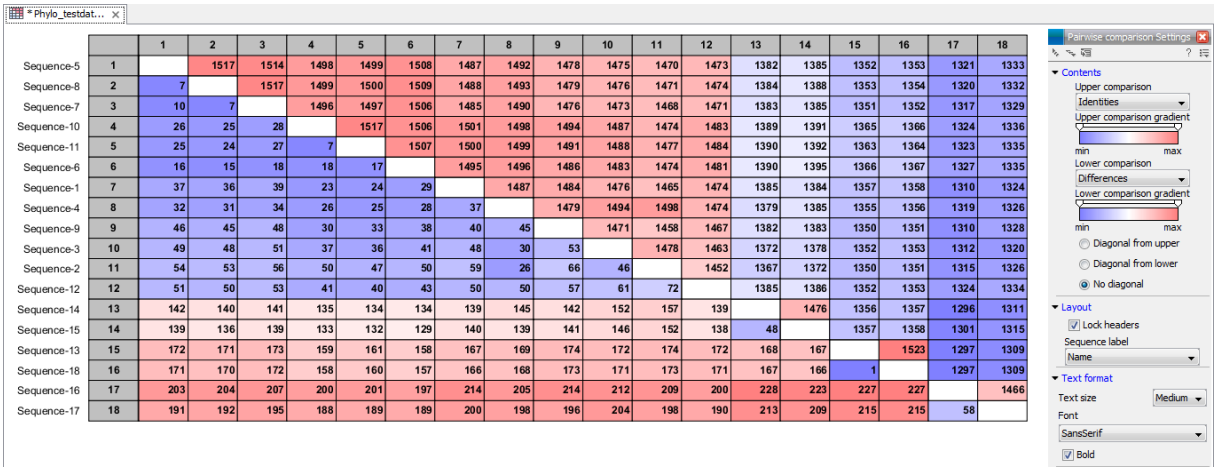


Figure 14.12: *A pairwise comparison table.*

values that appears when you slide the cursor reflect the percentage of the range of values in the table, and not absolute values.

The following settings are present in the side panel:

- **Contents**
  - **Upper comparison** Selects the comparison to show in the upper triangle of the table.
  - **Upper comparison gradient** Selects the color gradient to use for the upper triangle.
  - **Lower comparison** Selects the comparison to show in the lower triangle. Choose the same comparison as in the upper triangle to show all the results of an asymmetric comparison.
  - **Lower comparison gradient** Selects the color gradient to use for the lower triangle.
  - **Diagonal from upper** Use this setting to show the diagonal results from the upper comparison.
  - **Diagonal from lower** Use this setting to show the diagonal results from the lower comparison.
  - **No Diagonal.** Leaves the diagonal table entries blank.
- **Layout**
  - **Lock headers** Locks the sequence labels and table headers when scrolling the table.
  - **Sequence label** Changes the sequence labels.
- **Text format**
  - **Text size** Changes the size of the table and the text within it.
  - **Font** Changes the font in the table.
  - **Bold** Toggles the use of boldface in the table.

14.5 Bioinformatics explained: Multiple alignments

Multiple alignments are at the core of bioinformatical analysis. Often the first step in a chain of bioinformatical analyses is to construct a multiple alignment of a number of homologs DNA or protein sequences. However, despite their frequent use, the development of multiple alignment algorithms remains one of the algorithmically most challenging areas in bioinformatical research.

Constructing a multiple alignment corresponds to developing a hypothesis of how a number of sequences have evolved through the processes of character substitution, insertion and deletion. The input to multiple alignment algorithms is a number of homologous sequences i.e. sequences that share a common ancestor and most often also share molecular function. The generated alignment is a table (see figure 14.13) where each row corresponds to an input sequence and each column corresponds to a position in the alignment. An individual column in this table represents residues that have all diverged from a common ancestral residue. Gaps in the table (commonly represented by a '-') represent positions where residues have been inserted or deleted and thus do not have ancestral counterparts in all sequences.

14.5.1 Use of multiple alignments

Once a multiple alignment is constructed it can form the basis for a number of analyses:

- The phylogenetic relationship of the sequences can be investigated by tree-building methods based on the alignment.
- Annotation of functional domains, which may only be known for a subset of the sequences, can be transferred to aligned positions in other un-annotated sequences.
- Conserved regions in the alignment can be found which are prime candidates for holding functionally important sites.
- Comparative bioinformatical analysis can be performed to identify functionally important regions.

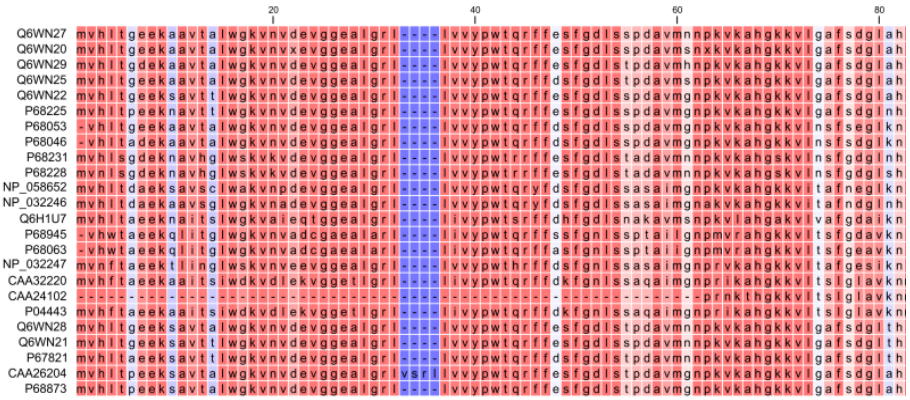


Figure 14.13: *The tabular format of a multiple alignment of 24 Hemoglobin protein sequences. Sequence names appear at the beginning of each row and the residue position is indicated by the numbers at the top of the alignment columns. The level of sequence conservation is shown on a color scale with blue residues being the least conserved and red residues being the most conserved.*

14.5.2 Constructing multiple alignments

Whereas the optimal solution to the pairwise alignment problem can be found in reasonable time, the problem of constructing a multiple alignment is much harder.

The first major challenge in the multiple alignment procedure is how to rank different alignments i.e. which *scoring function* to use. Since the sequences have a shared history they are correlated through their *phylogeny* and the scoring function should ideally take this into account. Doing so is, however, not straightforward as it increases the number of model parameters considerably. It is therefore commonplace to either ignore this complication and assume sequences to be unrelated, or to use heuristic corrections for shared ancestry.

The second challenge is to find the optimal alignment given a scoring function. For pairs of sequences this can be done by *dynamic programming* algorithms, but for more than three sequences this approach demands too much computer time and memory to be feasible.

A commonly used approach is therefore to do *progressive alignment* [Feng and Doolittle, 1987] where multiple alignments are built through the successive construction of pairwise alignments. These algorithms provide a good compromise between time spent and the quality of the resulting alignment

Presently, the most exciting development in multiple alignment methodology is the construction of *statistical alignment* algorithms [Hein, 2001], [Hein et al., 2000]. These algorithms employ a scoring function which incorporates the underlying phylogeny and use an explicit stochastic model of molecular evolution which makes it possible to compare different solutions in a statistically rigorous way. The optimization step, however, still relies on dynamic programming and practical use of these algorithms thus awaits further developments.

14.6 Phylogenetic tree features

Phylogenetics describes the taxonomical classification of organisms based on their evolutionary history i.e. their phylogeny. Phylogenetics is therefore an integral part of the science of systematics that aims to establish the phylogeny of organisms based on their characteristics. Furthermore, phylogenetics is central to evolutionary biology as a whole as it is the condensation of the overall paradigm of how life arose and developed on earth. The focus of this module is the reconstruction and visualization of phylogenetic trees. Phylogenetic trees illustrate the inferred evolutionary history of a set of organisms, and makes it possible to e.g. identify groups of closely related organisms and observe clustering of organisms with common traits. See 14.7.2 for a more detailed introduction to phylogenetic trees.

The viewer for visualizing and working with phylogenetic trees allows the user to create high-quality, publication-ready figures of phylogenetic trees. Large trees can be explored in two alternative tree layouts; circular and radial. The viewer supports importing, editing and visualization of metadata associated with nodes in phylogenetic trees.

Below is an overview of the main features of the phylogenetic tree editor. Further details can be found in the subsequent sections.

Main features of the phylogenetic tree editor:

- Circular and radial layouts.

- Import of metadata in Excel and CSV format.
- Tabular view of metadata with support for editing.
- Options for collapsing nodes based on bootstrap values.
- Re-ordering of tree nodes.
- Legends describing metadata.
- Visualization of metadata though e.g. node color, node shape, branch color, etc.
- Minimap navigation.
- Coloring and labeling of subtrees.
- Curved edges.
- Editable node sizes and line width.
- Intelligent visualization of overlapping labels and nodes.

The viewer for visualizing and working with phylogenetic trees allows the user to create high-quality, publication-ready figures of phylogenetic trees. Large trees can be explored in two alternative tree layouts; circular and radial.

Below is an overview of the main features of the phylogenetic tree editor. Further details can be found in the subsequent sections.

Main features of the phylogenetic tree editor:

- Circular and radial layouts.
- Options for collapsing nodes based on bootstrap values.
- Re-ordering of tree nodes.
- Minimap navigation.
- Coloring and labeling of subtrees.
- Curved edges.
- Editable node sizes and line width.
- Intelligent visualization of overlapping labels and nodes.

14.7 Create Trees

For a given set of aligned sequences (see section 14.1) it is possible to infer their evolutionary relationships. In *CLC Drug Discovery Workbench* this may be done using one of two distance based methods (see "Bioinformatics explained" in section 14.7.2).

For a given set of aligned sequences (see section 14.1) it is possible to infer their evolutionary relationships. In *CLC Drug Discovery Workbench* this may be done using a distance based method to generate a phylogenetic tree:

- **Create Tree** ( ) Is a tool that uses distance estimates computed from multiple alignments to create trees. The user can select whether to use Jukes-Cantor distance correction or Kimura distance correction (Kimura 80 for nucleotides/Kimura protein for proteins) in combination with either the neighbor joining or UPGMA method (see section 14.7.2).

14.7.1 Create tree

The "Create tree" tool can be used to generate a distance-based phylogenetic tree with multiple alignments as input:

Toolbox | **Sequence Alignment** ( )| **Create Tree** ( )

This will open the dialog displayed in figure 14.14:

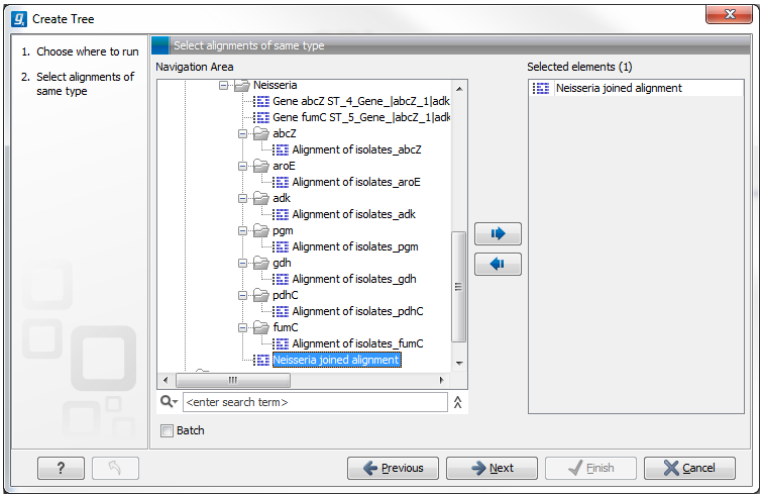


Figure 14.14: *Creating a tree.*

If an alignment was selected before choosing the Toolbox action, this alignment is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove elements from the **Navigation Area**. Click **Next** to adjust parameters.

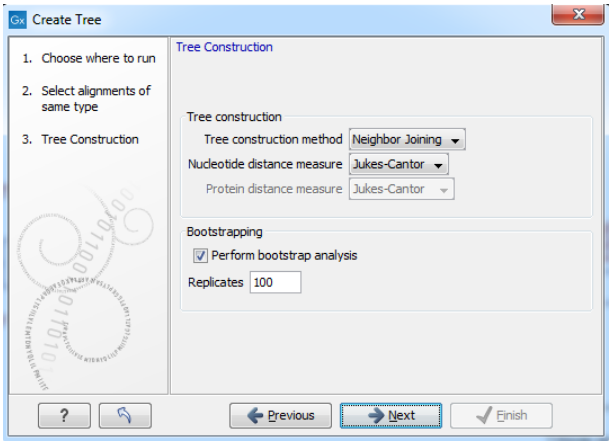


Figure 14.15: *Adjusting parameters for distance-based methods.*

Figure 14.15 shows the parameters that can be set for this distance-based tree creation:

- Tree construction (see section 14.7.2)

  - Tree construction method
    - The **UPGMA** method. Assumes constant rate of evolution.
    - The **Neighbor Joining** method. Well suited for trees with varying rates of evolution.
  - Nucleotide distance measure
    - **Jukes-Cantor**. Assumes equal base frequencies and equal substitution rates.
    - **Kimura 80**. Assumes equal base frequencies but distinguishes between transitions and transversions.
  - Protein distance measure
    - **Jukes-Cantor**. Assumes equal amino acid frequency and equal substitution rates.
    - **Kimura protein**. Assumes equal amino acid frequency and equal substitution rates. Includes a small correction term in the distance formula that is intended to give better distance estimates than Jukes-Cantor.

- Bootstrapping.
  - Perform bootstrap analysis. To evaluate the reliability of the inferred trees, *CLC Drug Discovery Workbench* allows the option of doing a **bootstrap** analysis (see section 14.7.2). A bootstrap value will be attached to each node, and this value is a measure of the confidence in the subtree rooted at the node. The number of replicates used in the bootstrap analysis can be adjusted in the wizard. The default value is 100 replicates which is usually enough to distinguish between reliable and unreliable nodes in the tree. The bootstrap value assigned to each inner node in the output tree is the percentage (0-100) of replicates which contained the same subtree as the one rooted at the inner node.

For a more detailed explanation, see "Bioinformatics explained" in section 14.7.2.

14.7.2 Bioinformatics explained

The phylogenetic tree

The evolutionary hypothesis of a phylogeny can be graphically represented by a phylogenetic tree.

Figure 14.16 shows a proposed phylogeny for the great apes, *Hominidae*, taken in part from Purvis [Purvis, 1995]. The tree consists of a number of nodes (also termed vertices) and branches (also termed edges). These nodes can represent either an individual, a species, or a higher grouping and are thus broadly termed taxonomical units. In this case, the terminal nodes (also called leaves or tips of the tree) represent extant species of *Hominidae* and are the *operational taxonomical units* (OTUs). The internal nodes, which here represent extinct common ancestors of the great apes, are termed *hypothetical taxonomical units* since they are not directly observable.

The ordering of the nodes determine the tree *topology* and describes how lineages have diverged over the course of evolution. The branches of the tree represent the amount of evolutionary divergence between two nodes in the tree and can be based on different measurements. A tree is completely specified by its topology and the set of all edge lengths.

The phylogenetic tree in figure 14.16 is rooted at the most recent common ancestor of all *Hominidae* species, and therefore represents a hypothesis of the direction of evolution e.g. that

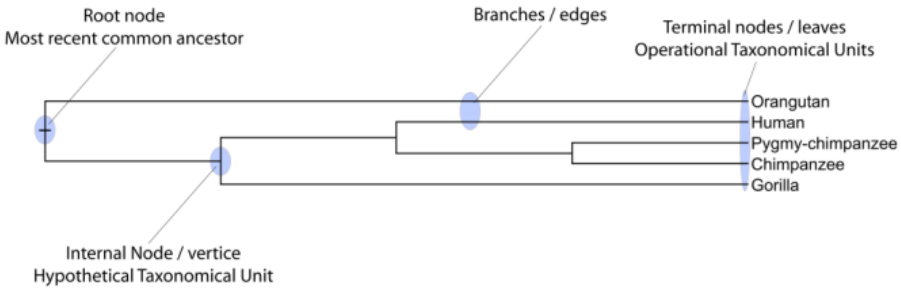


Figure 14.16: *A proposed phylogeny of the great apes (Hominidae). Different components of the tree are marked, see text for description.*

the common ancestor of gorilla, chimpanzee and man existed before the common ancestor of chimpanzee and man. In contrast, an unrooted tree would represent relationships without assumptions about ancestry.

Modern usage of phylogenies

Besides evolutionary biology and systematics the inference of phylogenies is central to other areas of research.

As more and more genetic diversity is being revealed through the completion of multiple genomes, an active area of research within bioinformatics is the development of comparative machine learning algorithms that can simultaneously process data from multiple species [Siepel and Haussler, 2004]. Through the comparative approach, valuable evolutionary information can be obtained about which amino acid substitutions are functionally tolerant to the organism and which are not. This information can be used to identify substitutions that affect protein function and stability, and is of major importance to the study of proteins [Knudsen and Miyamoto, 2001]. Knowledge of the underlying phylogeny is, however, paramount to comparative methods of inference as the phylogeny describes the underlying correlation from shared history that exists between data from different species.

In molecular epidemiology of infectious diseases, phylogenetic inference is also an important tool. The very fast substitution rate of microorganisms, especially the RNA viruses, means that these show substantial genetic divergence over the time-scale of months and years. Therefore, the phylogenetic relationship between the pathogens from individuals in an epidemic can be resolved and contribute valuable epidemiological information about transmission chains and epidemiologically significant events [Leitner and Albert, 1999], [Forsberg et al., 2001].

Distance based reconstruction methods

Distance based phylogenetic reconstruction methods use a pairwise distance estimate between the input organisms to reconstruct trees. The distances are an estimate of the evolutionary distance between each pair of organisms which are usually computed from DNA or amino acid sequences. Given two homologous sequences a distance estimate can be computed by aligning the sequences and then counting the number of positions where the sequences differ. The number of differences is called the observed number of substitutions and is usually an underestimate of the real distance as multiple mutations could have occurred at any position.

To correct for these hidden substitutions a substitution model, such as Jukes-Cantor or Kimura

80, can be used to get a more precise distance estimate.

Alternatively, k-mer based methods or SNP based methods can be used to get a distance estimate without the use of substitution models.

After distance estimates have been computed, a phylogenetic tree can be reconstructed using a distance based reconstruction method. Most distance based methods perform a bottom up reconstruction using a greedy clustering algorithm. Initially, each input organism is put in its own cluster which corresponds to a leaf node in the resulting tree. Next, pairs of clusters are iteratively joined into higher level clusters, which correspond to connecting two nodes in the tree with a new parent node. When a single node remains, the tree is reconstructed.

The *CLC Drug Discovery Workbench* provides two of the most widely used distance based reconstruction methods:

- The **UPGMA** method [Michener and Sokal, 1957] which assumes a constant rate of evolution (molecular clock hypothesis) in the different lineages. This method reconstruct trees by iteratively joining the two nearest clusters until there is only one cluster left. The result of the UPGMA method is a rooted bifurcating tree annotated with branch lengths.
- The **Neighbor Joining** method [Saitou and Nei, 1987] attempts to reconstruct a minimum evolution tree (a tree where the sum of all branch lengths is minimized). Opposite to the UPGMA method, the neighbor joining method is well suited for trees with varying rates of evolution in different lineages. A tree is reconstructed by iteratively joining clusters which are close to each other but at the same time far from all other clusters. The resulting tree is a bifurcating tree with branch lenghts. Since no particular biological hypothesis is made about the placement of the root in this method, the resulting tree is unrooted.

Bootstrap tests

Bootstrap tests [Felsenstein, 1985] is one of the most common ways to evaluate the reliability of the topology of a phylogenetic tree. In a bootstrap test, trees are evaluated using Efron's re-sampling technique [Efron, 1982], which samples nucleotides from the original set of sequences as follows:

Given an alignment of n sequences (rows) of length l (columns), we randomly choose l columns in the alignment with replacement and use them to create a new alignment. The new alignment has n rows and l columns just like the original alignment but it may contain duplicate columns and some columns in the original alignment may not be included in the new alignment. From this new alignment we reconstruct the corresponding tree and compare it to the original tree. For each subtree in the original tree we search for the same subtree in the new tree and add a score of one to the node at the root of the subtree if the subtree is present in the new tree. This procedure is repeated a number of times (usually around 100 times). The result is a counter for each interior node of the original tree, which indicate how likely it is to observe the exact same subtree when the input sequences are sampled. A bootstrap value is then computed for each interior node as the percentage of resampled trees that contained the same subtree as that rooted at the node.

Bootstrap values can be seen as a measure of how reliably we can reconstruct a tree, given the sequence data available. If all trees reconstructed from resampled sequence data have very different topologies, then most bootstrap values will be low, which is a strong indication that the

topology of the original tree cannot be trusted.

Scale bar

The scale bar unit depends on the distance measure used and the tree construction algorithm used. The trees produced using the Maximum Likelihood Phylogeny tool has a very specific interpretation: A distance of x means that the expected number of substitutions/changes per nucleotide (amino acid for protein sequences) is x. i.e. if the distance between two taxa is 0.01, you expected a change in each nucleotide independently with probability 1 %. For the remaining algorithms, there is not as nice an interpretation. The distance depends on the weight given to different mutations as specified by the distance measure.

14.8 Tree Settings

The Tree Settings Side Panel found in the left side of the view area can be used to adjust the tree layout and to visualize metadata that is associated with the tree nodes . The following section describes the visualization options available from the Tree Settings side panel. Note however that editing legend boxes related to metadata can be done directly from editing the metadata table (see section 14.9).

The preferred tree layout settings (user defined tree settings) can be saved and applied via the top right **Save Tree Settings** (figure 14.17). Settings can either be saved **For This Tree Only** or for all saved phylogenetic trees (**For Tree View in General**). The first option will save the layout of the tree for that tree only and it ensures that the layout is preserved even if it is exported and opened by a different user. The second option stores the layout globally in the Workbench and makes it available to other trees through the **Apply Saved Settings** option.

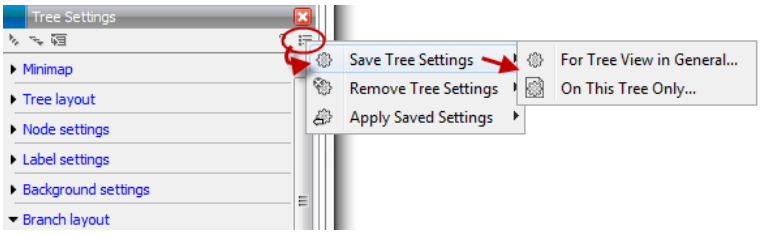


Figure 14.17: *Save, remove or apply preferred layout settings.*

Tree Settings contains the following categories:

- Minimap
- Tree layout
- Node settings
- Label settings
- Background settings
- Branch layout
- Bootstrap settings
- Metadata

14.8.1 Minimap

The Minimap is a navigation tool that shows a small version of the tree. A grey square indicates the specific part of the tree that is visible in the View Area (figure 14.18). To navigate the tree using the Minimap, click on the Minimap with the mouse and move the grey square around within the Minimap.

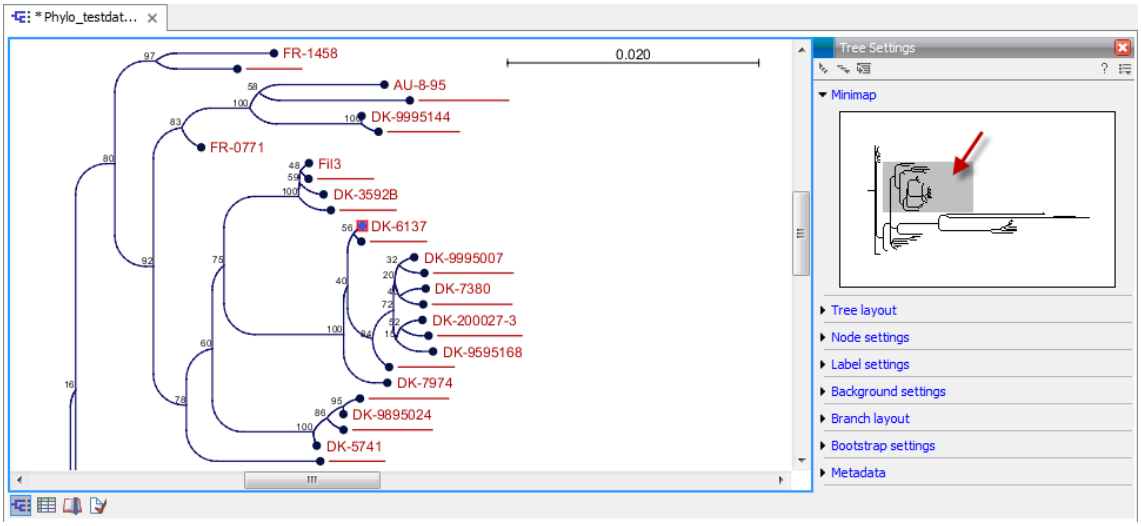


Figure 14.18: *Visualization of a phylogenetic tree. The grey square in the Minimap shows the part of the tree that is shown in the View Area.*

14.8.2 Tree layout

The **Tree Layout** can be adjusted in the Side Panel (figure 14.20).

- **Layout** Selects one of the five layout types: Phylogram, Cladogram, Circular Phylogram, Circular Cladogram or Radial. Note that only the Cladogram layouts are available if all branches in the tree have zero length.
  - **Phylogram** is a rooted tree where the edges have "lengths", usually proportional to the inferred amount of evolutionary change to have occurred along each branch.
  - **Cladogram** is a rooted tree without branch lengths which is useful for visualizing the topology of trees.
  - **Circular Phylogram** is also a phylogram but with the leaves in a circular layout.
  - **Circular Cladogram** is also a cladogram but with the leaves in a circular layout.
  - **Radial** is an unrooted tree that has the same topology and branch lengths as the rooted styles, but lacks any indication of evolutionary direction.
- **Ordering** The nodes can be ordered after the branch length; either **Increasing** (shown in figure 14.20) or **Decreasing**.
- **Reset Tree Topology** Resets to the default tree topology and node order (see figure 14.20).
- **Fixed width on zoom** Locks the horizontal size of the tree to the size of the main window. Zoom is therefore only performed on the vertical axis when this option is enabled.
- **Show as unrooted tree** The tree can be shown with or without a root.

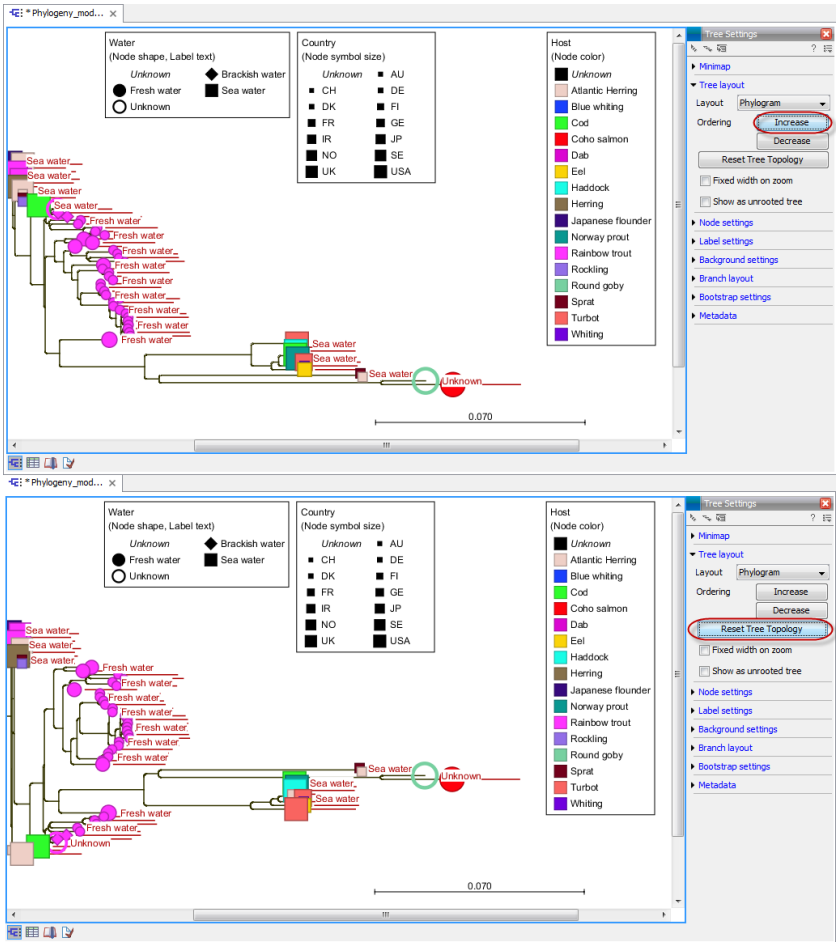


Figure 14.19: *The tree layout can be adjusted in the Side Panel. The top part of the figure shows a tree with increasing node order. In the bottom part of the figure the tree has been reverted to the original tree topology.*

14.8.3 Node settings

The nodes can be manipulated in several ways.

- **Leaf node symbol** Leaf nodes can be shown as a range of different symbols (Dot, Box, Circle, etc.).
- **Internal node symbols** The internal nodes can also be shown with a range of different symbols (Dot, Box, Circle, etc.).
- **Max. symbol size** The size of leaf- and internal node symbols can be adjusted.
- **Avoid overlapping symbols** The symbol size will be automatically limited to avoid overlaps between symbols in the current view.
- **Node color** Specify a fixed color for all nodes in the tree.

The node layout settings in the Side Panel are shown in figure 14.21.

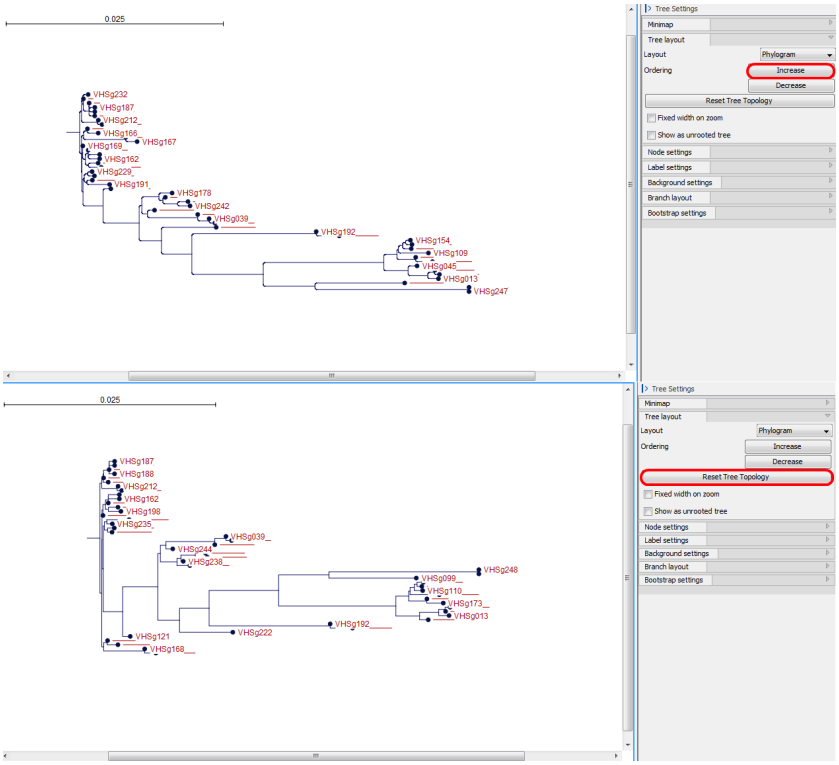


Figure 14.20: *The tree layout can be adjusted in the Side Panel. The top part of the figure shows a tree with increasing node order. In the bottom part of the figure the tree has been reverted to the original tree topology.*

14.8.4 Label settings

- **Label font settings** Can be used to specify/adjust font type, size and typography (Bold, Italic or normal).
- **Hide overlapping labels** Disable automatic hiding of overlapping labels and display all labels even if they overlap.
- **Show internal node labels** Labels for internal nodes of the tree (if any) can be displayed. Please note that subtrees and nodes can be labeled with a custom text. This is done by right clicking the node and selecting **Edit Label** (see figure 14.23).
- **Show leaf node labels** Leaf node labels can be shown or hidden.
- **Rotate Subtree labels** Subtree labels can be shown horizontally or vertically. Labels are shown vertically when "Rotate subtree labels" has been selected. Subtree labels can be added with the right click option "Set Subtree Label" that is enabled from "Decorate subtree" (see section 14.8.9).
- **Align labels** Align labels to the node furthest from the center of the tree so that all labels are positioned next to each other. The exact behavior depends on the selected tree layout.
- **Connect labels to nodes** Adds a thin line from the leaf node to the aligned label. Only possible when Align labels option is selected.

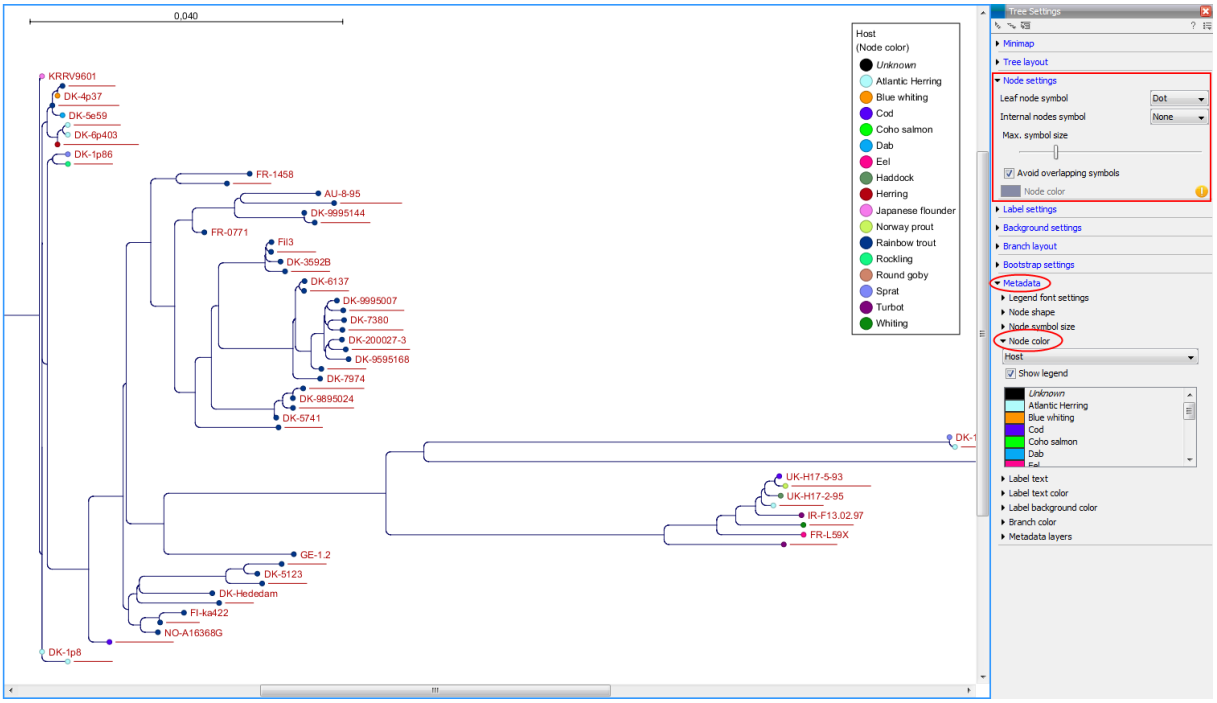


Figure 14.21: *The Node Layout settings. Node color is specified by metadata and is therefore inactive in this example.*

When working with big trees there is typically not enough space to show all labels. As illustrated in figure 14.23, only some of the labels are shown. The hidden labels are illustrated with thin horizontal lines (figure 14.24).

There are different ways of showing more labels. One way is to reduce the font size of the labels, which can be done under **Label font settings** in the Side Panel. Another option is to zoom in on specific areas of the tree (figure 14.24 and figure 14.25). The last option is to disable **Hide overlapping labels** under "Label settings" in the right side panel. When this option is unchecked all labels are shown even if the text overlaps. When allowing overlapping labels it is usually a good idea to disable **Show label background** under "Background settings" (see section 14.8.5).

Note! When working with a tree with hidden labels, it is possible to make the hidden label text appear by moving the mouse over the node with the hidden label. **Note!** The text within labels can be edited by editing the metadata table values directly.

14.8.5 Background settings

- **Show label background** Show a background color for each label. Once ticked, it is possible to specify whether to use a fixed color or to use the color that is associated with the selected metadata category.

- **Show label background** Show a background color for each label. Once ticked, it is possible to specify a background color.

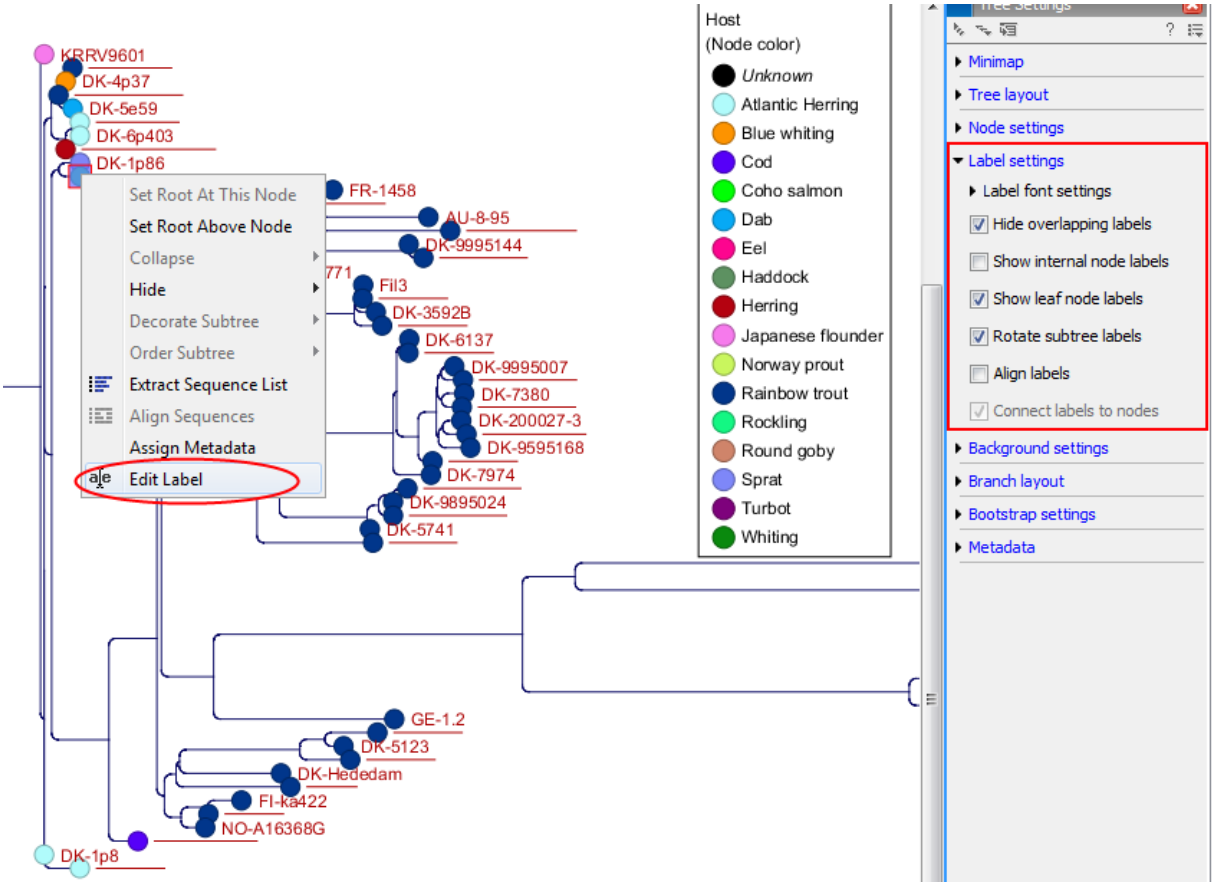


Figure 14.22: *"Edit label" in the right click menu can be used to customize the label text. The way node labels are displayed can be controlled through the labels settings in the right side panel.*

14.8.6 Branch layout

- **Branch length font settings** Specify/adjust font type, size and typography (Bold, Italic or normal).
- **Line color** Select the default line color.
- **Line width** Select the width of branches (1.0-3.0 pixels).
- **Curvature** Adjust the degree of branch curvature to get branches with round corners.
- **Min. length** Select a minimum branch length. This option can be used to prevent nodes connected with a short branch to cluster at the parent node.
- **Show branch lengths** Show or hide the branch lengths.

The branch layout settings in the Side Panel are shown in figure 14.27.

14.8.7 Bootstrap settings

Bootstrap values can be shown on the internal nodes. The bootstrap values are shown in percent and can be interpreted as confidence levels where a bootstrap value close to 100 indicate a clade, which is strongly supported by the data from which the tree was reconstructed. Bootstrap

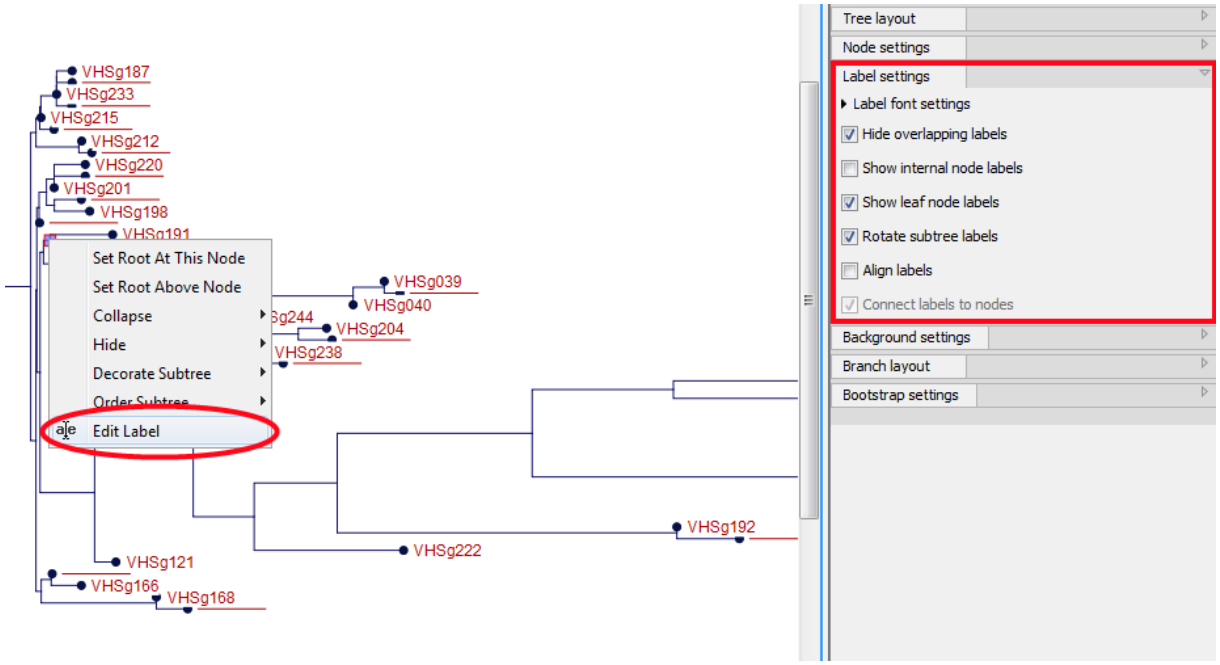


Figure 14.23: *"Edit label" in the right click menu can be used to customize the label text. The way node labels are displayed can be controlled through the labels settings in the right side panel.*

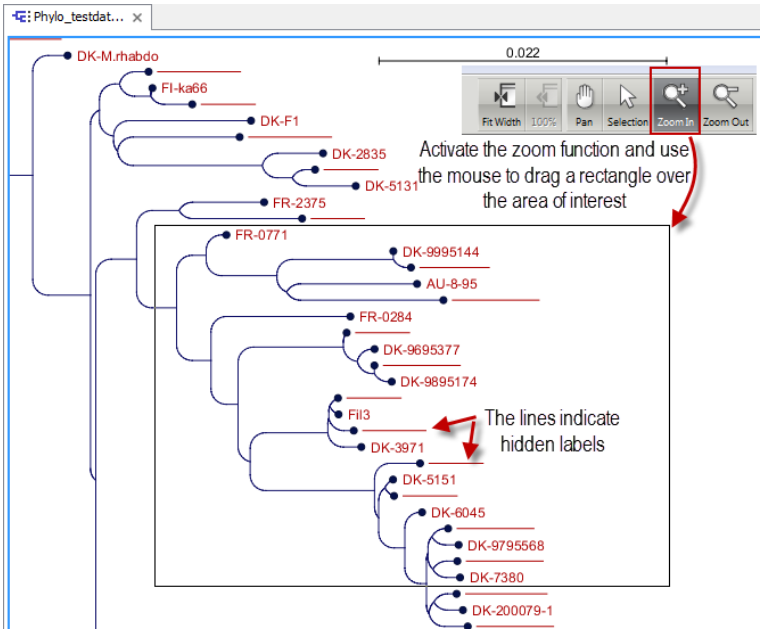


Figure 14.24: *The zoom function in the upper right corner of CLC Genomics Workbench can be used to zoom in on a particular region of the tree. When the zoom function has been activated, use the mouse to drag a rectangle over the area that you wish to zoom in at.*

values are useful for identifying clades in the tree where the topology (and branch lengths) should not be trusted.

Some branches in rooted trees may not have bootstrap values. Trees constructed with neighbour joining are unrooted and to correctly visualize them, the "Radial" view is required. In all other tree views we need a root to visualize the tree. An "artificial node" and therefore an extra branch are created for such visualization to achieve this, which makes it look like a bootstrap value is

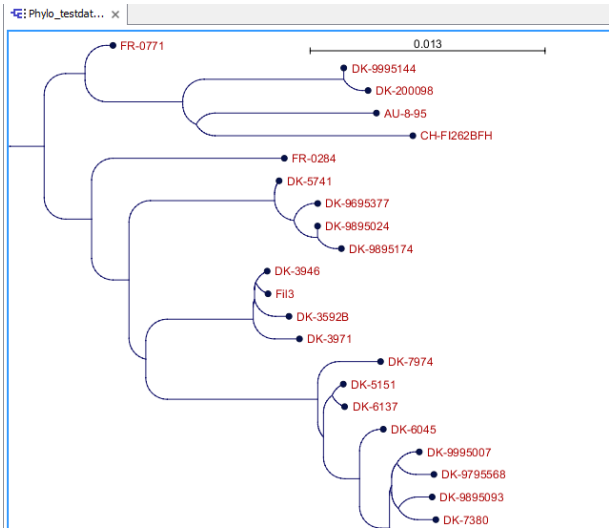


Figure 14.25: *After zooming in on a region of interest more labels become visible. In this example all labels are now visible.*

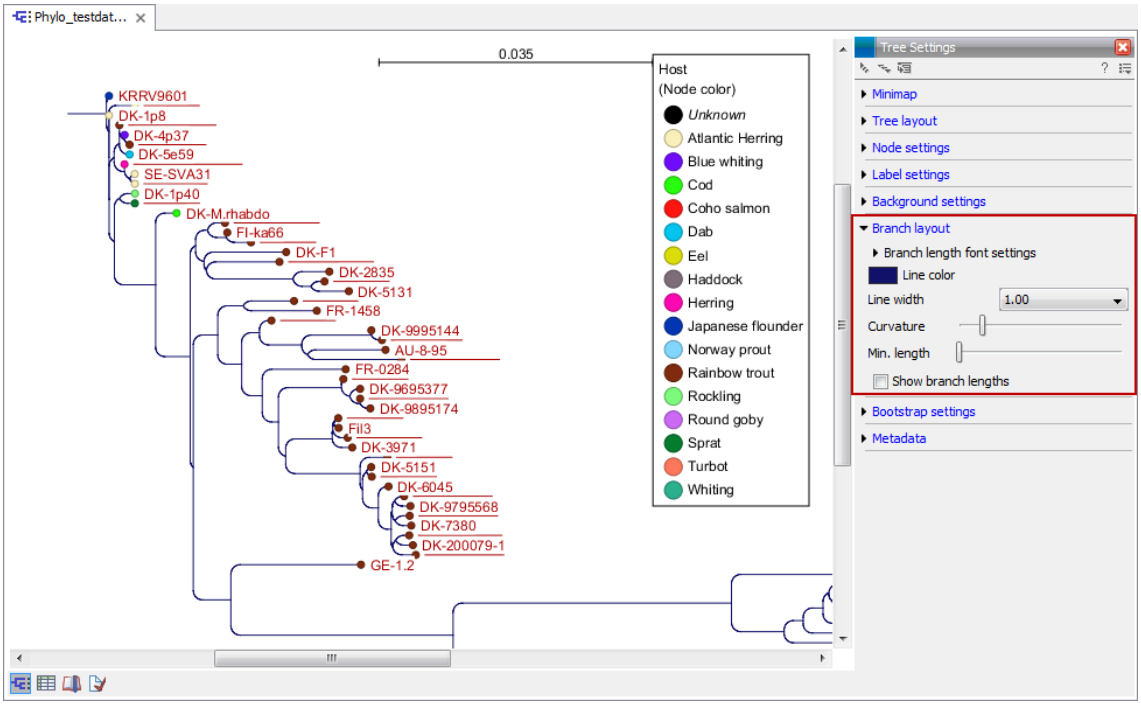


Figure 14.26: *Branch Layout settings.*

missing

- **Bootstrap value font settings** Specify/adjust font type, size and typography (Bold, Italic or normal).
- **Show bootstrap values (%)** Show or hide bootstrap values. When selected, the bootstrap values (in percent) will be displayed on internal nodes if these have been computed during the reconstruction of the tree.
- **Bootstrap threshold (%)** When specifying a bootstrap threshold, the branch lengths can be controlled manually by collapsing internal nodes with bootstrap values under a certain

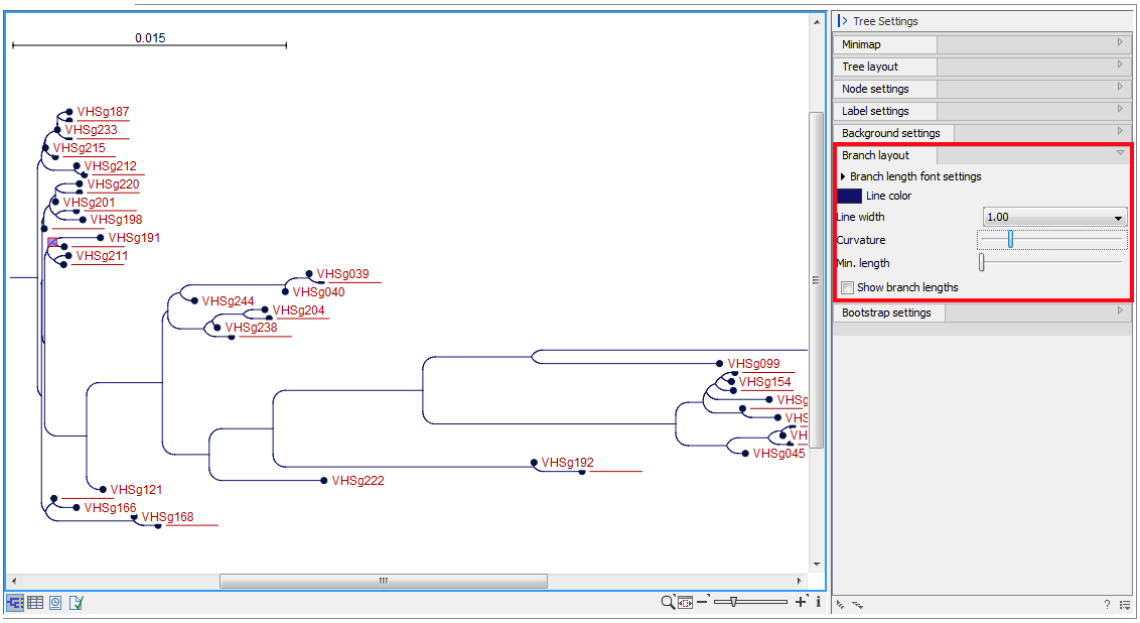


Figure 14.27: *Branch Layout settings.*

threshold.

- **Highlight bootstrap $\geq$ (%)** Highlights branches where the bootstrap value is above the user defined threshold.

14.8.8 Visualizing metadata

Metadata associated with a phylogenetic tree (described in detail in section 14.9) can be visualized in a number of different ways:

- **Node shape** Different node shapes are available to visualize metadata.
- **Node symbol size** Change the node symbol size to visualize metadata.
- **Node color** Change the node color to visualize metadata.
- **Label text** The metadata can be shown directly as text labels as shown in figure 14.28.
- **Label text color** The label text can be colored and used to visualize metadata (see figure 14.28).
- **Label background color** The background color of node text labels can be used to visualize metadata.
- **Branch color** Branch colors can be changed according to metadata.
- **Metadata layers** Color coded layers shown next to leaf nodes.

Please note that when visualizing metadata through a tree property that can be adjusted in the right side panel (such as node color or node size), an exclamation mark will appear next to the control for that property to indicate that the setting is inactive because it is defined by metadata (see figure 14.21).

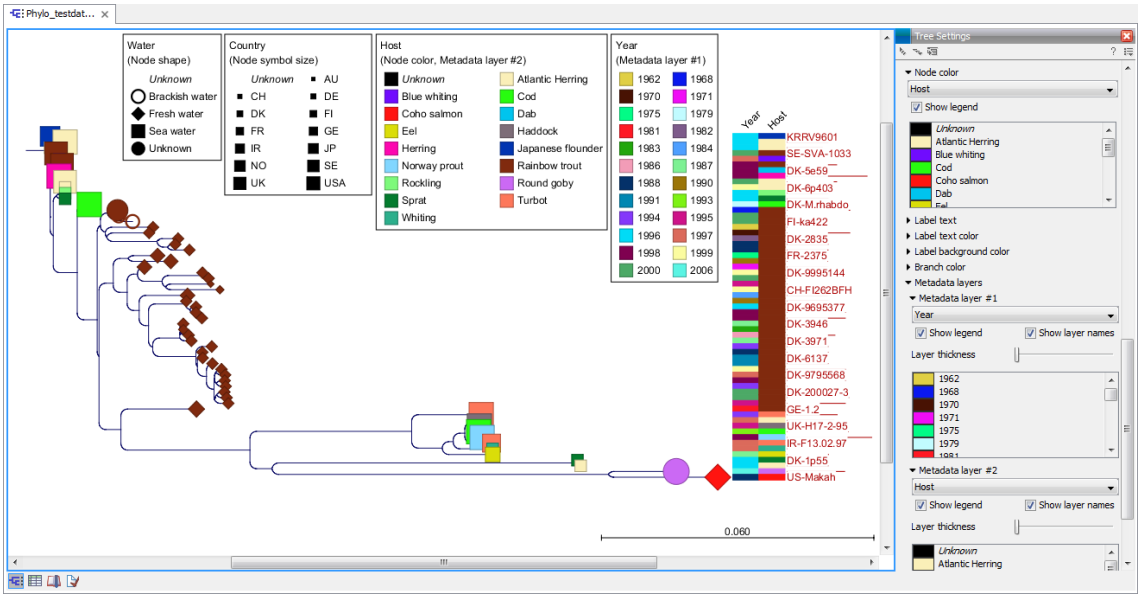


Figure 14.28: *Different types of metadata kan be visualized by adjusting node size, shape, and color. Two color-code metadata layers (Year and Host) are shown in the right side of the tree.*

14.8.9 Node right click menu

Additional options for layout and extraction of subtree data are available when right clicking the nodes (figure 14.23):

- **Set Root At This Node** Re-root the tree using the selected node as root. Please note that re-rooting will change the tree topology.
- **Set Root Above Node** Re-root the tree by inserting a node between the selected node and its parent. Useful for rooting trees using an outgroup.
- **Collapse** Branches associated with a selected node can be collapsed with or without the associated labels. Collapsed branches can be uncollapsed using the *Uncollapse* option in the same menu.
- **Hide** Can be used to hide a node or a subtree. Hidden nodes or subtrees can be shown again using the *Show Hidden Subtree* function on a node which is root in a subtree containing hidden nodes (see figure 14.30). When hiding nodes, a new button appears labeled "Show X hidden nodes" in the Side Panel under "Tree Layout" (figure 14.31). When pressing this button, all hidden nodes are shown again.
- **Decorate Subtree** A subtree can be labeled with a customized name, and the subtree lines and/or background can be colored. To save the decoration, see figure 14.17 and use option: **Save/Restore Settings** | **Save Tree View Settings On This Tree View only**.
- **Order Subtree** Rearrange leaves and branches in a subtree by Increasing/Decreasing depth, respectively. Alternatively, change the order of a node's children by left clicking and dragging one of the node's children.
- **Extract Sequence List** Sequences associated with selected leaf nodes are extracted to a new sequence list.

Figure 14.29: *A subtree can be hidden by selecting "Hide Subtree" and is shown again when selecting "Show Hidden Subtree" on a parent node.*

- **Align Sequences** Sequences associated with selected leaf nodes are extracted and used as input to the *Create Alignment* tool.

- **Assign Metadata** Metadata can be added, deleted or modified. To add new metadata categories a new "Name" must be assigned. (This will be the column header in the metadata table). To add a new metadata category, enter a value in the "Value" field. To delete values, highlight the relevant nodes and right click on the selected nodes. In the dialog that appears, use the drop-down list to select the name of the desired metadata category and leave the value field empty. When pressing "Add" the values for the selected metadata category will be deleted from the selected nodes. Metadata can be modified in the same way, but instead of leaving the value field empty, the new value should be entered.

- **Edit label** Edit the text in the selected node label. Labels can be shown or hidden by using the Side Panel: **Label settings | Show internal node labels**

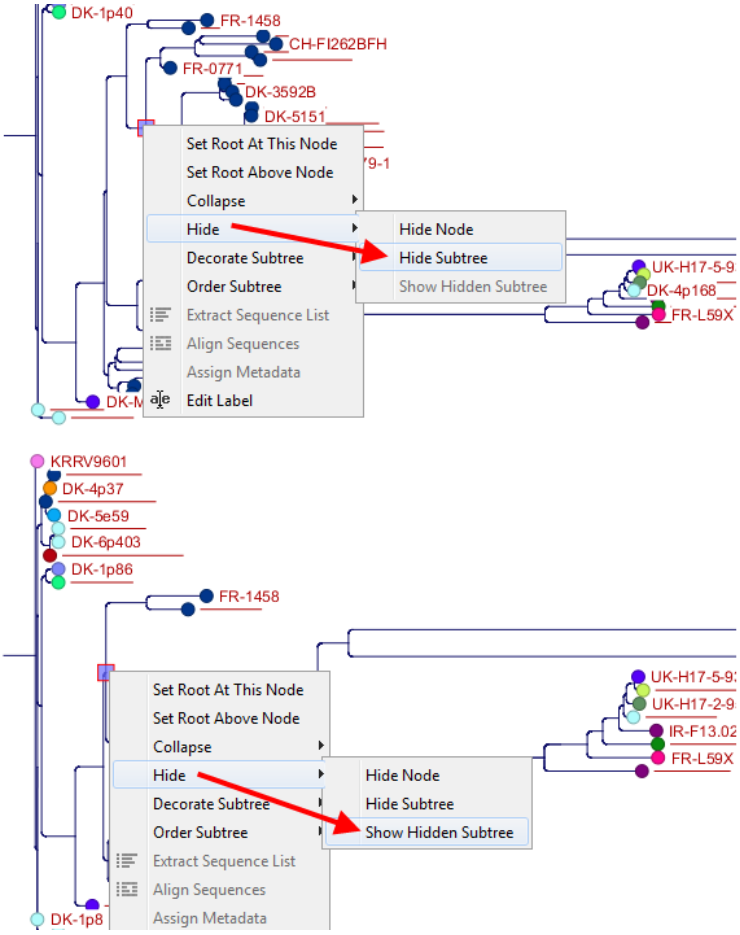


Figure 14.30: *A subtree can be hidden by selecting "Hide Subtree" and is shown again when selecting "Show Hidden Subtree" on a parent node.*

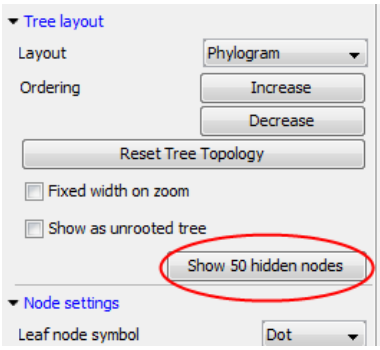


Figure 14.31: *When hiding nodes, a new button labeled "Show X hidden nodes" appears in the Side Panel under "Tree Layout". When pressing this button, all hidden nodes are brought back.*

14.9 Metadata and phylogenetic trees

When a tree is reconstructed, some mandatory metadata will be added to nodes in the tree. These metadata are special in the sense that the tree viewer has specialized features for visualizing the data and some of them cannot be edited. The mandatory metadata include:

- **Node name** The node name.
- **Branch length** The length of the branch, which connects a node to the parent node.

- **Bootstrap value** The bootstrap value for internal nodes.
- **Size** The length of the sequence which corresponds to each leaf node. This only applies to leaf nodes.
- **Start of sequence** The first 50bp of the sequence corresponding to each leaf node.

To view metadata associated with a phylogenetic tree, click on the table icon ( ) at the bottom of the tree. If you hold down the Ctrl key (or ⌘ on Mac) while clicking on the table icon ( ), you will be able to see both the tree and the table in a split view (figure 14.32).

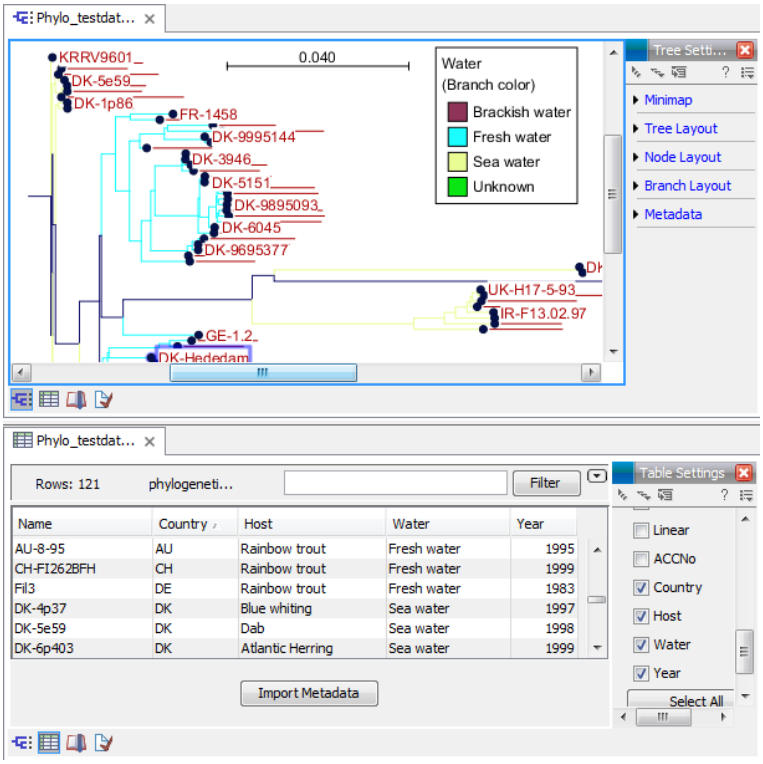


Figure 14.32: *Tabular metadata that is associated with an existing tree shown in a split view.*

Additional metadata can be associated with a tree by clicking the **Import Metadata** button. This will open up the dialog shown in figure 14.33.

To associate metadata with an existing tree a common denominator is required. This is achieved by mapping the node names in the "Name" column of the metadata table to the names that have been used in the metadata table to be imported. In this example the "Strain" column holds the names of the nodes and this column must be assigned "Name" to allow the importer to associate metadata with nodes in the tree.

It is possible to import a subset of the columns in a set of metadata. An example is given in figure 14.33. The column "H" is not relevant to import and can be excluded simply by leaving the text field at the top row of the column empty.

14.9.1 Table Settings and Filtering

How to use the metadata table (see figure 14.34):

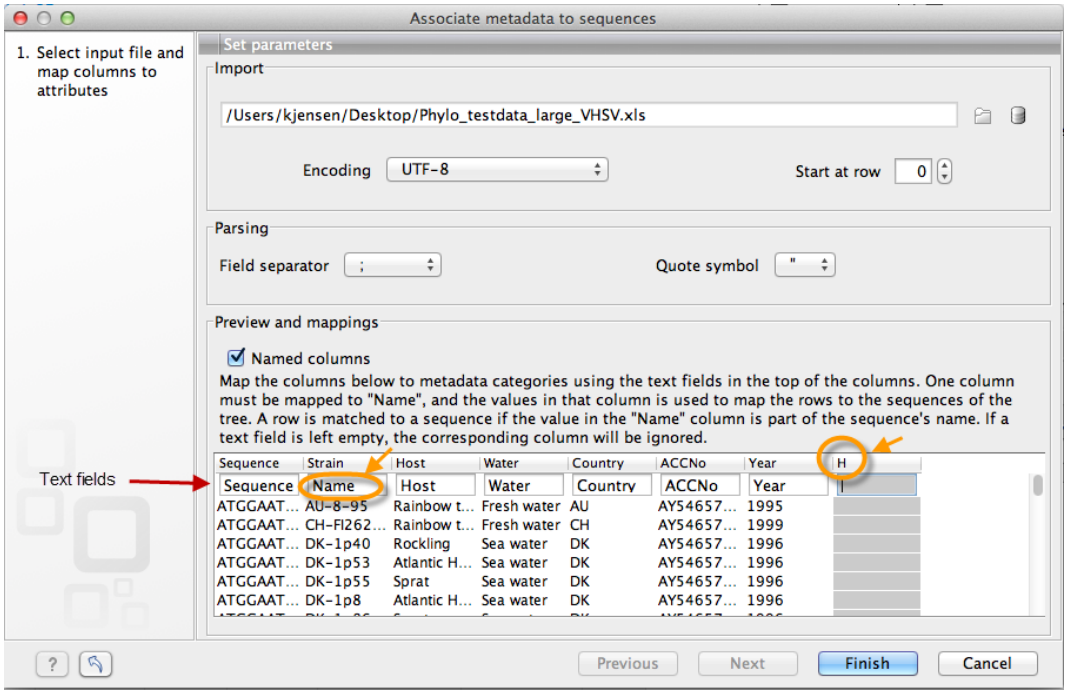


Figure 14.33: *Import of metadata for a tree. The second column named "Strain" is choosen as the common denominator by entering "Name" in the text field of the column. The column labeled "H" is ignored by not assigning a column heading to this column.*

- **Column width** The column width can be adjusted in two ways; *Manually* or *Automatically*.
- **Show column** Selects which metadata categories that are shown in the table layout.
- **Filtering Metadata information** Metadata information in a table can be filtered by a simple- or advanced mode (this is described in the CLC Genomics Workbench manual, Appendix D, Filtering tables).

14.9.2 Add or modify metadata on a tree

It is possible to add and modify metadata from both the tree view and the table view.

Metadata can be added and edited in the metadata table by using the following right click options (see figure 14.35):

- **Assign Metadata** The right click option "Assign Metadata" can be used for four purposes.
  - To add new metadata categories (columns). In this case, a new "Name" must be assigned, which will be the column header. To add a new column requires that a value is entered in the "Value" field. This can be done by right clicking anywhere in the table.
  - To add values to one or more rows in an existing column. In this case, highlight the relevant rows and right click on the selected rows. In the dialog that appears, use the drop-down list to select the name of the desired column and enter a value.
  - To delete values from an existing column. This is done in the same way as when adding a new value, with the only exception that the value field should be left empty.

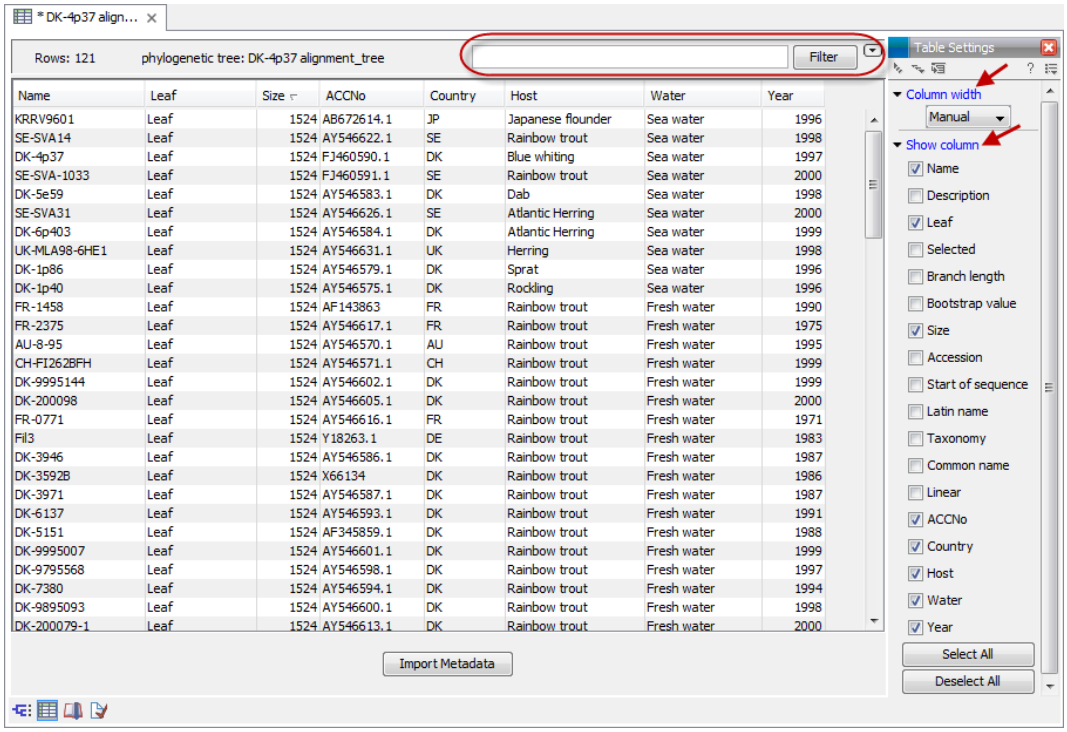


Figure 14.34: *Metadata table. The column width can be adjusted manually or automatically. Under "Show column" it is possible to select which columns should be shown in the table. Filtering using specific criteria can be performed (this is described in the CLC Genomics Workbench manual, Appendix D, Filtering tables).*

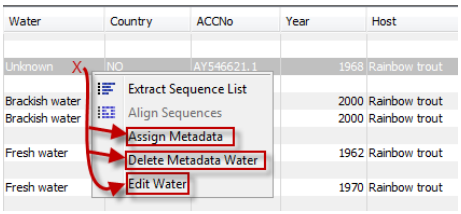


Figure 14.35: *Right click options in the metadata table.*

  - To delete metadata columns. This is done by selecting all rows in the table followed by a right click anywhere in the table. Select the name of the column to delete from the drop down menu and leave the value field blank. When pressing "Add", the selected column will disappear.

- **Delete Metadata "column header"** This is the most simple way of deleting a metadata column. Click on one of the rows in the column to delete and select "Delete *column header*".

- **Edit "column header"** To modify existing metadata point, right click on a cell in the table and select the "Edit *column header*" (see an example in figure 14.36). To edit multiple entries at once, select multiple rows in the table, right click a selected cell in the column you want to edit and choose "Edit *column header*". This will change values in all selected rows in the column that was clicked.

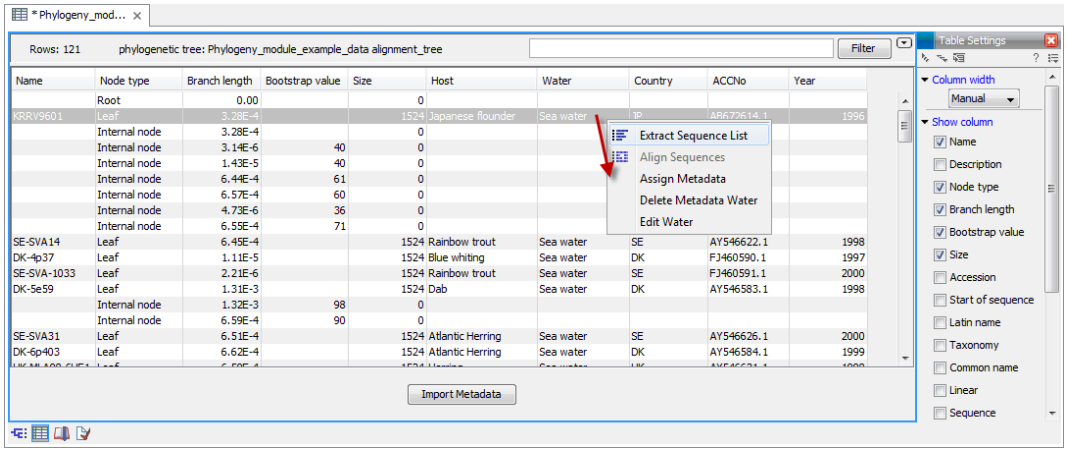


Figure 14.36: *To include an extra metadata column, use the right click option "Assign Metadata", provide "Name" (the column header) and "Value". To modify existing metadata, click on the specific field, select "Edit column header" and provide new value.*

14.9.3 Undefined metadata values on a tree

When visualizing a metadata category where one or more nodes in the tree have undefined values, these nodes will be visualized using a default value. This value will always be shown in italics in the top of the legend (see the entry "*Unknown*" in figure 14.37). To remove this entry in the legend, all nodes must have a value assigned in the corresponding metadata category.

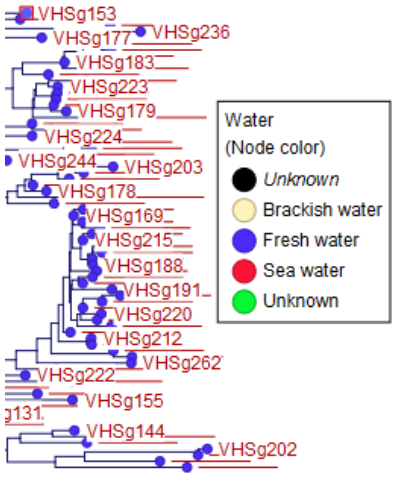


Figure 14.37: *A legend for a metadata category where one or more values are undefined. Fill your metadata table with a value of your choice to edit the mention of* "unknown" *in the legend.*

14.9.4 Selection of specific nodes

Selection of nodes in a tree is automatically synchronized to the metadata table and the other way around. Nodes in a tree can be selected in three ways:

- *Selection of a single node* Click once on a single node. Additional nodes can be added by holding down Ctrl (or ⌘ for Mac) and clicking on them (see figure 14.38).

- *Selecting all nodes in a subtree* Double clicking on a inner node results in the selection of all nodes in the subtree rooted at the node.
- *Selection via the Metadata table* Select one or more entries in the table. The corresponding nodes will now be selected in the tree.

It is possible to extract a subset of the underlying sequence data directly through either the tree viewer or the metadata table as follows. Select one or more nodes in the tree where at least one node has a sequence attached. Right click one of the selected nodes and choose **Extract Sequence List**. This will generate a new sequence list containing all sequences attached to the selected nodes. The same functionality is available in the metadata table where sequences can be extracted from selected rows using the right click menu. Please note that all extracted sequences are copies and any changes to these sequences will not be reflected in the tree.

When analyzing a phylogenetic tree it is often convenient to have a multiple alignment of sequences from e.g. a specific clade in the tree. A quick way to generate such an alignment is to first select one or more nodes in the tree (or the corresponding entries in the metadata table) and then select **Align Sequences** in the right click menu. This will extract the sequences corresponding to the selected elements and use a copy of them as input to the multiple alignment tool (see section 14.5). Next, change relevant option in the multiple alignment wizard that pops up and click **Finish**. The multiple alignment will now be generated.

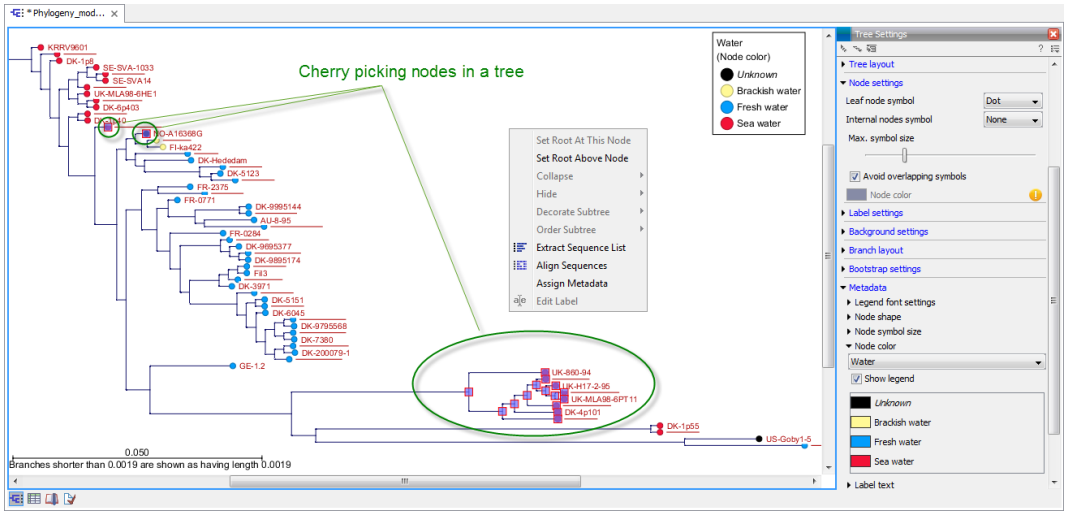


Figure 14.38: *Cherry picking nodes in a tree. The selected leaf sequences can be extracted by right clicking on one of the selected nodes and selecting "Extract Sequence List". It is also possible to Align Sequences directly by right clicking on the nodes or leaves.*

Part IV

Appendix

Appendix A

Use of multi-core computers

Many tools in CLC Workbenches and Servers can make use of multi-core CPUs. This does not necessarily mean that all available CPU cores are used throughout the analysis. It means that these tools benefit from running on computers with multiple CPU cores.

Tools available differ between CLC Workbenches. In the table, the availability of these tools in different CLC Workbench Toolbox menus is indicated with an X.

Use of multi-core computers	Genomics	Drug Discovery	Biomedical Genomics
Basic Variant Detection	X		X
BLAST (will not scale well on many cores)	X		
Create Alignment	X	X	X
Create Detailed Mapping Report	X		X
Create Sequencing QC Report (will not scale well on more than four cores)	X		
De Novo Assembly	X		
Dock Ligands		X	
Download Reference Genome Data	X		
Extract and Count	X		X
Fixed Ploidy Variant Detection	X		X
Import Molecules from SMILES or 2D		X	
K-mer Based Tree Construction	X		
Large Gap Read Mapper (currently in beta)	X		
Local Realignment	X		X
Low Frequency Variant Detection	X		X
Map Reads to Contigs	X		
Map Reads to Reference	X		X
Maximum Likelihood Phylogeny	X		
Model Testing	X		
Probabilistic Variant Detection (legacy)	X		X
QC for Sequencing Reads (will not scale well on more than four cores)			X
Quality-based Variant Detection (legacy)	X		X
RNA-Seq Analysis	X		X
Screen Ligands		X	
Trim Sequences	X		X

Please note that a static license has a limitation on the maximum number of cores, see section 1.3.1.

Appendix B

Graph preferences

This section explains the view settings of graphs. The **Graph preferences** at the top of the **Side Panel** includes the following settings:

- **Lock axes** This will always show the axes even though the plot is zoomed to a detailed level.
- **Frame** Shows a frame around the graph.
- **Show legends** Shows the data legends.
- **Tick type** Determine whether tick lines should be shown outside or inside the frame.
  - Outside
  - Inside
- **Tick lines at** Choosing Major ticks will show a grid behind the graph.
  - None
  - Major ticks
- **Horizontal axis range** Sets the range of the horizontal axis (x axis). Enter a value in **Min** and **Max**, and press Enter. This will update the view. If you wait a few seconds without pressing Enter, the view will also be updated.
- **Vertical axis range** Sets the range of the vertical axis (y axis). Enter a value in **Min** and **Max**, and press Enter. This will update the view. If you wait a few seconds without pressing Enter, the view will also be updated.
- **X-axis at zero**. This will draw the x axis at y = 0. Note that the axis range will not be changed.
- **Y-axis at zero**. This will draw the y axis at x = 0. Note that the axis range will not be changed.
- **Show as histogram**. For some data-series it is possible to see the graph as a histogram rather than a line plot.

The **Lines and plots** below contains the following settings:

- **Dot type**
  - None
  - Cross
  - Plus
  - Square
  - Diamond
  - Circle
  - Triangle
  - Reverse triangle
  - Dot
- **Dot color.** Allows you to choose between many different colors. Click the color box to select a color.
- **Line width**
  - Thin
  - Medium
  - Wide
- **Line type**
  - None
  - Line
  - Long dash
  - Short dash
- **Line color.** Allows you to choose between many different colors. Click the color box to select a color.

For graphs with multiple data series, you can select which curve the dot and line preferences should apply to. This setting is at the top of the **Side Panel** group.

Note that the graph title and the axes titles can be edited simply by clicking with the mouse. These changes will be saved when you **Save** () the graph - whereas the changes in the **Side Panel** need to be saved explicitly (see section 4.5).

Appendix C

BLAST databases

Several databases are available at NCBI, which can be selected to narrow down the possible BLAST hits.

C.1 Peptide sequence databases

- **nr.** Non-redundant GenBank CDS translations + PDB + SwissProt + PIR + PRF, excluding those in env_nr.
- **refseq.** Protein sequences from NCBI Reference Sequence project http://www.ncbi.nlm.nih.gov/RefSeq/.
- **swissprot.** Last major release of the SWISS-PROT protein sequence database (no incremental updates).
- **pat.** Proteins from the Patent division of GenBank.
- **pdb.** Sequences derived from the 3-dimensional structure records from the Protein Data Bank http://www.rcsb.org/pdb/.
- **env_nr.** Non-redundant CDS translations from env_nt entries.
- **month.** All new or revised GenBank CDS translations + PDB + SwissProt + PIR + PRF released in the last 30 days..

C.2 Nucleotide sequence databases

- **nr.** All GenBank + EMBL + DDBJ + PDB sequences (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences). No longer "non-redundant" due to computational cost.
- **refseq_rna.** mRNA sequences from NCBI Reference Sequence Project.
- **refseq_genomic.** Genomic sequences from NCBI Reference Sequence Project.
- **est.** Database of GenBank + EMBL + DDBJ sequences from EST division.
- **est_human.** Human subset of est.

- **est_mouse.** Mouse subset of est.
- **est_others.** Subset of est other than human or mouse.
- **gss.** Genome Survey Sequence, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences.
- **htgs.** Unfinished High Throughput Genomic Sequences: phases 0, 1 and 2. Finished, phase 3 HTG sequences are in nr.
- **pat.** Nucleotides from the Patent division of GenBank.
- **pdb.** Sequences derived from the 3-dimensional structure records from Protein Data Bank. They are NOT the coding sequences for the corresponding proteins found in the same PDB record.
- **month.** All new or revised GenBank+EMBL+DDBJ+PDB sequences released in the last 30 days.
- **alu.** Select Alu repeats from REPBASE, suitable for masking Alu repeats from query sequences. See "Alu alert" by Claverie and Makalowski, Nature 371: 752 (1994).
- **dbsts.** Database of Sequence Tag Site entries from the STS division of GenBank + EMBL + DDBJ.
- **chromosome.** Complete genomes and complete chromosomes from the NCBI Reference Sequence project. It overlaps with refseq_genomic.
- **wgs.** Assemblies of Whole Genome Shotgun sequences.
- **env_nt.** Sequences from environmental samples, such as uncultured bacterial samples isolated from soil or marine samples. The largest single source is Sagarsso Sea project. This does overlap with nucleotide nr.

C.3 Adding more databases

Besides the databases that are part of the default configuration, you can add more databases located at NCBI by configuring files in the Workbench installation directory.

The list of databases that can be added is here: https://web.archive.org/web/20120409025527/http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/remote_blastdblist.html.

In order to add a new database, find the `settings` folder in the Workbench installation directory (e.g. `C:\Program files\CLC Genomics Workbench 4`). Download unzip and place the following files in this directory to replace the built-in list of databases:

- Nucleotide databases: http://www.clcbio.com/wbsettings/NCBI_BlastNucleotideDataba zip
- Protein databases: http://www.clcbio.com/wbsettings/NCBI_BlastProteinDatabases.zip

Open the file you have downloaded into the `settings` folder, e.g. `NCBI_BlastProteinDatabases.proper` in a text editor and you will see the contents look like this:

```
nr[clcdefault] = Non-redundant protein sequences
refseq_protein = Reference proteins
swissprot = Swiss-Prot protein sequences
pat = Patented protein sequences
pdb = Protein Data Bank proteins
env_nr = Environmental samples
month = New or revised GenBank sequences
```

Simply add another database as a new line with the first item being the database name taken from https://web.archive.org/web/20120409025527/http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/remote_blastdblist.html and the second part is the name to display in the Workbench. Restart the Workbench, and the new database will be visible in the BLAST dialog.

Appendix D

IUPAC codes for amino acids

(Single-letter codes based on International Union of Pure and Applied Chemistry)

The information is gathered from: http://www.insdc.org/documents/feature_table.html

One-letter abbreviation	Three-letter abbreviation	Description
A	Ala	Alanine
R	Arg	Arginine
N	Asn	Asparagine
D	Asp	Aspartic acid
C	Cys	Cysteine
Q	Gln	Glutamine
E	Glu	Glutamic acid
G	Gly	Glycine
H	His	Histidine
J	Xle	Leucine or Isoleucineucine
L	Leu	Leucine
I	ILe	Isoleucine
K	Lys	Lysine
M	Met	Methionine
F	Phe	Phenylalanine
P	Pro	Proline
O	Pyl	Pyrrolysine
U	Sec	Selenocysteine
S	Ser	Serine
T	Thr	Threonine
W	Trp	Tryptophan
Y	Tyr	Tyrosine
V	Val	Valine
B	Asx	Aspartic acid or Asparagine Asparagine
Z	Glx	Glutamic acid or Glutamine Glutamine
X	Xaa	Any amino acid

Appendix E

IUPAC codes for nucleotides

(Single-letter codes based on International Union of Pure and Applied Chemistry)

The information is gathered from: http://www.iupac.org and http://www.insdc.org/documents/feature_table.html.

Code	Description
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil
R	Purine (A or G)
Y	Pyrimidine (C, T, or U)
M	C or A
K	T, U, or G
W	T, U, or A
S	C or G
B	C, T, U, or G (not A)
D	A, T, U, or G (not C)
H	A, T, U, or C (not G)
V	A, C, or G (not T, not U)
N	Any base (A, C, G, T, or U)

Appendix F

Formats for import and export

F.1 List of bioinformatic data formats

Below is a list of bioinformatic data formats, i.e. formats for importing and exporting molecule structures, sequences, alignments and trees.

F.1.1 Molecule structure formats

File type	Suffix	Import	Export	Description
PDB	.pdb	X	X	
Tripos Mol2	.mol2	X	X	
MDL Mol	.sdf	X		
CLC	.clc	X	X	Rich format including all information

F.1.2 Sequence data formats

File type	Suffix	Import	Export	Description
AB1	.ab1	X		Including chromatograms
ABI	.abi	X		Including chromatograms
CLC	.clc	X	X	Rich format including all information
Clone manager	.cm5	X		Clone manager sequence format
FASTA	.fsa/.fasta	X	X	Simple format, name & description
GCG sequence	.gcg	X	X	Rich information incl. annotations
Raw sequence	any	X		Only sequence (no name)
Sequence CSV	.csv	X	X	Simple format. One seq per line: name, description(optional), sequence
Tab delimited text	.txt		X	Annotations in tab delimited text format
Phred	.phd	X		Including chromatograms
PIR(NBRF)	.pir	X	X	Simple format, name and description
SCF2	.scf	X		Including chromatograms
SCF3	.scf	X	X	Including chromatograms
Staden	.sdn	X		
Swiss-Prot	.swp	X		Rich information incl. annotations (only peptides)

F.1.3 Alignment formats

File type	Suffix	Import	Export	Description
Aligned fasta	.fa	X	X	Simple fasta-based format with – for gaps
CLC	.clc	X	X	Rich format including all information
ClustalW	.aln	X	X	
GCG Alignment	.msf	X	X	
Phylip Alignment	.phy	X		

F.1.4 Tree formats

File type	Suffix	Import	Export	Description
CLC	.clc	X	X	Rich format including all information
Newick	.nwk	X		

F.1.5 Table and text formats

File type	Suffix	Import	Export	Description
Excel	.xls/.xlsx	X	X	All tables and reports
Table CSV	.csv	X	X	All tables
Tab delimited	.txt		X	All tables
Text	.txt	X	X	All data in a textual format
CLC	.clc	X	X	Rich format including all information
HTML	.html		X	All tables
PDF	.pdf		X	Export reports in Portable Document Format

F.1.6 File compression formats

File type	Suffix	Import	Export	Description
Zip export	.zip		X	Selected files in CLC format
Zip import	.zip/.gz/.tar	X		Contained files/folder structure

Note! It is possible to import 'external' files into the Workbench and view these in the **Navigation Area**, but it is only the above mentioned formats whose *contents* can be shown in the Workbench.

F.2 List of graphics data formats

Below is a list of formats for exporting graphics. All data displayed in a graphical format can be exported using these formats. Data represented in lists and tables can only be exported in .pdf format (see section 6.4 for further details).

Format	Suffix	Type
Portable Network Graphics	.png	bitmap
JPEG	.jpg	bitmap
Tagged Image File	.tif	bitmap
PostScript	.ps	vector graphics
Encapsulated PostScript	.eps	vector graphics
Portable Document Format	.pdf	vector graphics
Scalable Vector Graphics	.svg	vector graphics

Bibliography

[Altschul and Gish, 1996] Altschul, S. F. and Gish, W. (1996). Local alignment statistics. *Methods Enzymol*, 266:460–480.

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–410.

[Andrade et al., 1998] Andrade, M. A., O'Donoghue, S. I., and Rost, B. (1998). Adaptation of protein surfaces to subcellular location. *J Mol Biol*, 276(2):517–525.

[Bachmair et al., 1986] Bachmair, A., Finley, D., and Varshavsky, A. (1986). In vivo half-life of a protein is a function of its amino-terminal residue. *Science*, 234(4773):179–186.

[Bateman et al., 2004] Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C., and Eddy, S. R. (2004). The Pfam protein families database. *Nucleic Acids Res.*, 32(Database issue):D138–D141.

[Bendtsen et al., 2004a] Bendtsen, J. D., Jensen, L. J., Blom, N., Heijne, G. V., and Brunak, S. (2004a). Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng Des Sel*, 17(4):349–356.

[Bendtsen et al., 2005] Bendtsen, J. D., Kiemer, L., Fausbøll, A., and Brunak, S. (2005). Non-classical protein secretion in bacteria. *BMC Microbiol*, 5:58.

[Bendtsen et al., 2004b] Bendtsen, J. D., Nielsen, H., von Heijne, G., and Brunak, S. (2004b). Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*, 340(4):783–795.

[Berman et al., 2003] Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide protein data bank. *Nat Struct Biol*, 10(12):980.

[Blobel, 2000] Blobel, G. (2000). Protein targeting (Nobel lecture). *Chembiochem.*, 1:86–102.

[Cheng et al., 2007a] Cheng, A. C., Coleman, R. G., Smyth, K. T., Cao, Q., Soulard, P., Caffrey, D. R., Salzberg, A. C., and Huang, E. S. (2007a). Structure-based maximal affinity model predicts small-molecule druggability. *Nat Biotechnol*, 25(1):71–75.

[Cheng et al., 2007b] Cheng, T., Zhao, Y., Li, X., Lin, F., Xu, Y., Zhang, X., Li, Y., and Wang, R. (2007b). Computation of octanol-water partition coefficients by guiding an additive model with knowledge. *J. Chem. Inf. Model.*, 47:2140–2148.

[Cornette et al., 1987] Cornette, J. L., Cease, K. B., Margalit, H., Spouge, J. L., Berzofsky, J. A., and DeLisi, C. (1987). Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J Mol Biol*, 195(3):659–685.

[Crooks et al., 2004] Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res*, 14(6):1188–1190.

[Dalby et al., 1992] Dalby, A., Nourse, J. G., Hounshell, W. D., Gushurst, A. K. I., Grier, D. L., Leland, B. A., and Laufer, J. (1992). Description of several chemical structure file formats used by computer programs developed at molecular design limited. *Journal of Chemical Information and Computer Sciences*, 32 (3):244–255.

[Dayhoff and Schwartz, 1978] Dayhoff, M. O. and Schwartz, R. M. (1978). *Atlas of Protein Sequence and Structure*, volume 3 of *5 suppl.*, pages 353–358. Nat. Biomed. Res. Found., Washington D.C.

[Eddy, 2004] Eddy, S. R. (2004). Where did the BLOSUM62 alignment score matrix come from? *Nat Biotechnol*, 22(8):1035–1036.

[Efron, 1982] Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*, volume 38. SIAM.

[Eisenberg et al., 1984] Eisenberg, D., Schwarz, E., Komaromy, M., and Wall, R. (1984). Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J Mol Biol*, 179(1):125–142.

[Emini et al., 1985] Emini, E. A., Hughes, J. V., Perlow, D. S., and Boger, J. (1985). Induction of hepatitis a virus-neutralizing antibody by a virus-specific synthetic peptide. *J Virol*, 55(3):836–839.

[Engelman et al., 1986] Engelman, D. M., Steitz, T. A., and Goldman, A. (1986). Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu Rev Biophys Biophys Chem*, 15:321–353.

[Felsenstein, 1985] Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Journal of Molecular Evolution*, 39:783–791.

[Feng and Doolittle, 1987] Feng, D. F. and Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol*, 25(4):351–360.

[Forsberg et al., 2001] Forsberg, R., Oleksiewicz, M. B., Petersen, A. M., Hein, J., Bøtner, A., and Storgaard, T. (2001). A molecular clock dates the common ancestor of European-type porcine reproductive and respiratory syndrome virus at more than 10 years before the emergence of disease. *Virology*, 289(2):174–179.

[Galperin and Koonin, 1998] Galperin, M. Y. and Koonin, E. V. (1998). Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biol*, 1(1):55–67.

[Gill and von Hippel, 1989] Gill, S. C. and von Hippel, P. H. (1989). Calculation of protein extinction coefficients from amino acid sequence data. *Anal Biochem*, 182(2):319–326.

[Gonda et al., 1989] Gonda, D. K., Bachmair, A., Wünning, I., Tobias, J. W., Lane, W. S., and Varshavsky, A. (1989). Universality and structure of the N-end rule. *J Biol Chem*, 264(28):16700–16712.

[Halgren, 1996] Halgren, T. A. (1996). Merck molecular force field. i. basis, form, scope, parameterization, and performance of mmff94. *Journal of Computational Chemistry*, 17(5-6):490–519.

[Hein, 2001] Hein, J. (2001). An algorithm for statistical alignment of sequences related by a binary tree. In *Pacific Symposium on Biocomputing*, page 179.

[Hein et al., 2000] Hein, J., Wiuf, C., Knudsen, B., Møller, M. B., and Wibling, G. (2000). Statistical alignment: computational properties, homology testing and goodness-of-fit. *J Mol Biol*, 302(1):265–279.

[Henikoff and Henikoff, 1992] Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–10919.

[Hopp and Woods, 1983] Hopp, T. P. and Woods, K. R. (1983). A computer program for predicting protein antigenic determinants. *Mol Immunol*, 20(4):483–489.

[Ikai, 1980] Ikai, A. (1980). Thermostability and aliphatic index of globular proteins. *J Biochem (Tokyo)*, 88(6):1895–1898.

[Janin, 1979] Janin, J. (1979). Surface and inside volumes in globular proteins. *Nature*, 277(5696):491–492.

[Karplus and Schulz, 1985] Karplus, P. A. and Schulz, G. E. (1985). Prediction of chain flexibility in proteins. *Naturwissenschaften*, 72:212–213.

[Klee and Ellis, 2005] Klee, E. W. and Ellis, L. B. M. (2005). Evaluating eukaryotic secreted protein prediction. *BMC Bioinformatics*, 6:256.

[Knudsen and Miyamoto, 2001] Knudsen, B. and Miyamoto, M. M. (2001). A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proc Natl Acad Sci U S A*, 98(25):14512–14517.

[Knudsen and Miyamoto, 2003] Knudsen, B. and Miyamoto, M. M. (2003). Sequence alignments and pair hidden markov models using evolutionary history. *Journal of Molecular Biology*, 333(2):453 – 460.

[Kolaskar and Tongaonkar, 1990] Kolaskar, A. S. and Tongaonkar, P. C. (1990). A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Lett*, 276(1-2):172--174.

[Korb et al., 2009] Korb, O., Stützle, T., and Exner, T. E. (2009). Empirical scoring functions for advanced protein-ligand docking with plants. *J Chem Inf Model*, 49(1):84–96.

[Krogh et al., 2001] Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*, 305(3):567–580.

[Kyte and Doolittle, 1982] Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1):105–132.

[Leitner and Albert, 1999] Leitner, T. and Albert, J. (1999). The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proc Natl Acad Sci U S A*, 96(19):10752–10757.

[Li et al., 2008] Li, B., Turuvekere, S., Agrawal, M., La, D., Ramani, K., and Kihara, D. (2008). Characterization of local geometry of protein surfaces with the visibility criterion. *Proteins*, 71(2):670–683.

[Lipinski et al., 2001] Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 46:3–26.

[Lu et al., 2008] Lu, M., Dousis, A. D., and Ma, J. (2008). Opus-rota: A fast and accurate method for side-chain modeling. *Protein Science*, 17(9):1576–1585.

[Maizel and Lenk, 1981] Maizel, J. V. and Lenk, R. P. (1981). Enhanced graphic matrix analysis of nucleic acid and protein sequences. *Proc Natl Acad Sci U S A*, 78(12):7665–7669.

[McGinnis and Madden, 2004] McGinnis, S. and Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*, 32(Web Server issue):W20–W25.

[Menne et al., 2000] Menne, K. M., Hermjakob, H., and Apweiler, R. (2000). A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics*, 16(8):741–742.

[Miao et al., 2011] Miao, Z., Cao, Y., and Jiang, T. (2011). Rasp: rapid modeling of protein side chain conformations. *Bioinformatics*, 27(22):3117–3122.

[Michener and Sokal, 1957] Michener, C. and Sokal, R. (1957). A quantitative approach to a problem in classification. *Evolution*, 11:130–162.

[Mukherjee and Zhang, 2009] Mukherjee, S. and Zhang, Y. (2009). MM-align: A quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic Acids Res.*, 37.

[Nelder and Mead, 1965] Nelder and Mead (1965). A simplex method for function minimization. *Computer Journal*, 7:308–313.

[Nielsen et al., 1997] Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng*, 10(1):1–6.

[Purvis, 1995] Purvis, A. (1995). A composite estimate of primate phylogeny. *Philos Trans R Soc Lond B Biol Sci*, 348(1326):405–421.

[Reinhardt and Hubbard, 1998] Reinhardt, A. and Hubbard, T. (1998). Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res*, 26(9):2230–2236.

[Rose et al., 1985] Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H., and Zehfus, M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science*, 229(4716):834–838.

[Rost, 2001] Rost, B. (2001). Review: protein secondary structure prediction continues to rise. *J Struct Biol*, 134(2-3):204–218.

[Saitou and Nei, 1987] Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425.

[Schneider and Stephens, 1990] Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, 18(20):6097–6100.

[Siepel and Haussler, 2004] Siepel, A. and Haussler, D. (2004). Combining phylogenetic and hidden Markov models in biosequence analysis. *J Comput Biol*, 11(2-3):413–428.

[Smith and Waterman, 1981] Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197.

[Tobias et al., 1991] Tobias, J. W., Shrader, T. E., Rocap, G., and Varshavsky, A. (1991). The N-end rule in bacteria. *Science*, 254(5036):1374–1377.

[Vainio and Johnson, 2007] Vainio, M. J. and Johnson, M. S. (2007). Generating conformer ensembles using a multiobjective genetic algorithm. *J Chem Inf Model*, 47(6):2462–2474.

[von Heijne, 1986] von Heijne, G. (1986). A new method for predicting signal sequence cleavage sites. *Nucl. Acids Res.*, 14:4683–4690.

[Welling et al., 1985] Welling, G. W., Weijer, W. J., van der Zee, R., and Welling-Wester, S. (1985). Prediction of sequential antigenic regions in proteins. *FEBS Lett*, 188(2):215–218.

[Wootton and Federhen, 1993] Wootton, J. C. and Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Computers in Chemistry*, 17:149–163.

[Xu and Zhang, 2010] Xu, J. and Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, 26(7):889–95.

[Zhang and Skolnick, 2004] Zhang, Y. and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702–10.

Part V

Index

Appendix G

Index