Tutorial

BLAST Searches

June 27, 2019

Sample to Insight

QIAGEN Aarhus · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
Telephone: +45 70 22 32 44 · www.qiagenbioinformatics.com · ts-bioinformatics@qiagen.com

BLAST Searches

Here, you will learn how to:

- Use BLAST to find the gene coding for a protein in a genomic sequence.
- Find primer binding sites on genomic sequences.

A valuable source of information about BLAST can be found at http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=ProgSelectionGuide.

Remember that BLAST is a heuristic method. This means that certain assumptions are made to allow searches to be done in a reasonable amount of time. Thus, you cannot trust BLAST search results to present with an accurate alignment. For more accurate results you should consider using other algorithms, such as Smith-Waterman for a multiple sequence alignment. You can read more in "Bioinformatics explained: BLAST" found here: http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=BE_BLAST.html.

Locate a protein sequence on the chromosome

It is easy to to see the actual location of a protein sequence on a chromosome using BLAST.

In this example we wish to map the protein sequence of the human arrestin domain-containing protein 5 (ARRDC5) protein to a chromosome. We know in advance that the ARRDC5 is located somewhere on chromosome 19.

1. Data used in this example can be downloaded from GenBank:

   Download | Search for Sequences at NCBI ( )

   Search and click on the "Download and Save" button for the following:

   - Homo sapiens chromosome 19 (NC_000019) using the Nucleotide database (see figure 1)
   - The human arrestin domain-containing protein 5 ARRDC5 (NP_001073992) in the Protein database (highlighted in red in figure 2)

   Save them in a new folder you can create for this tutorial in the Navigation Area.

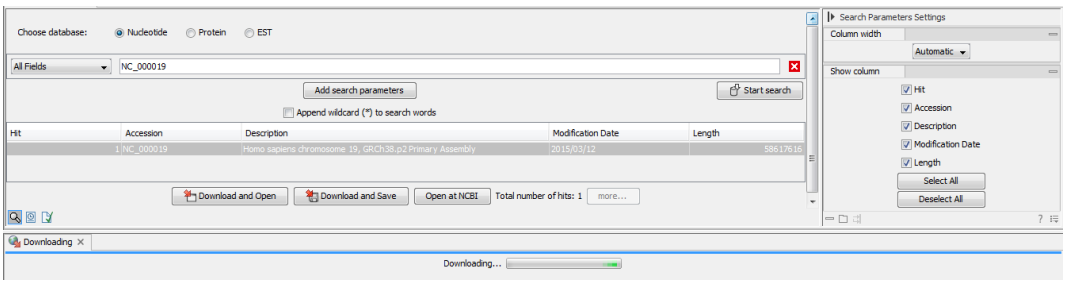

Figure 1: *Search for Homo sapiens chromosome 19.*

2. Next, conduct a local BLAST search: **Toolbox | BLAST ( ) | BLAST ( )**
3. Select the protein sequence as query sequence (figure 3) and click **Next**.

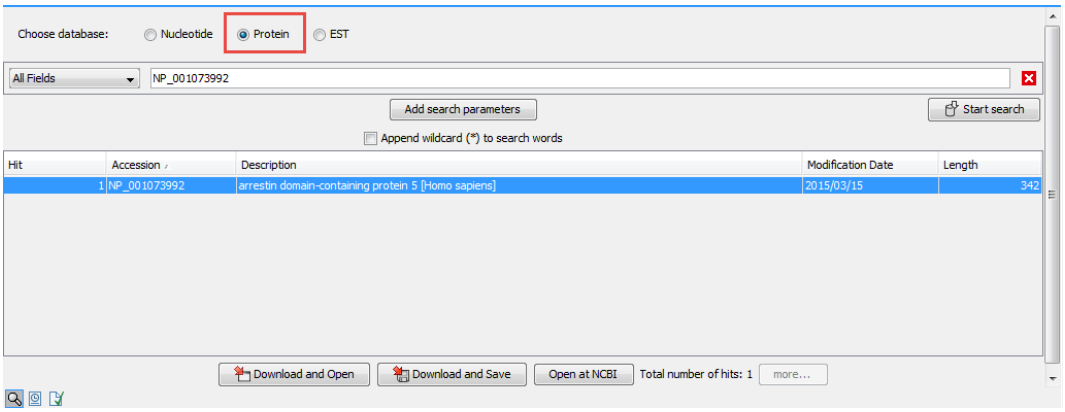

Figure 2: *Search for ARRDC5.*

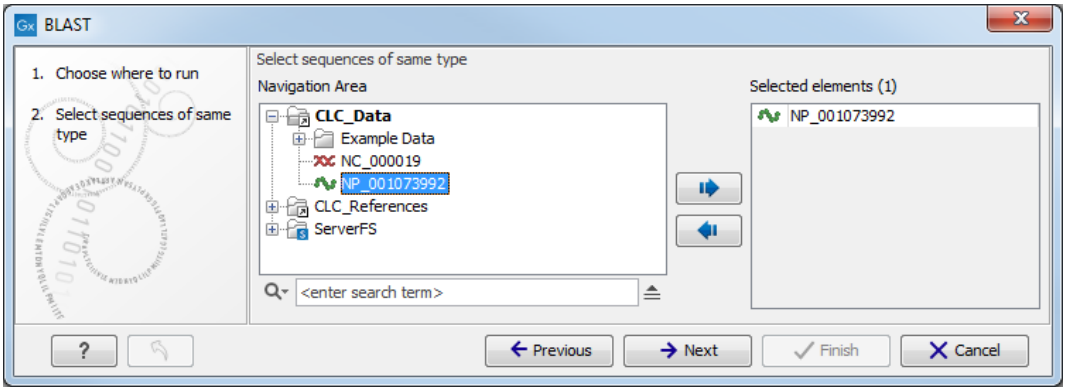

Figure 3: *Select the sequence(s) you want to BLAST.*

4. Since you wish to BLAST a protein sequence against a nucleotide sequence, select **tblastn**, which will automatically translate the selected nucleotide sequence to database format.

   Select the just downloaded chr19 sequence *NC_000019* as target. If you are used to BLAST, you will know that you usually have to create a BLAST database before BLASTing. However,the workbench does this "on the fly" when one or more sequences have been selected (figure 4). Click **Next**.

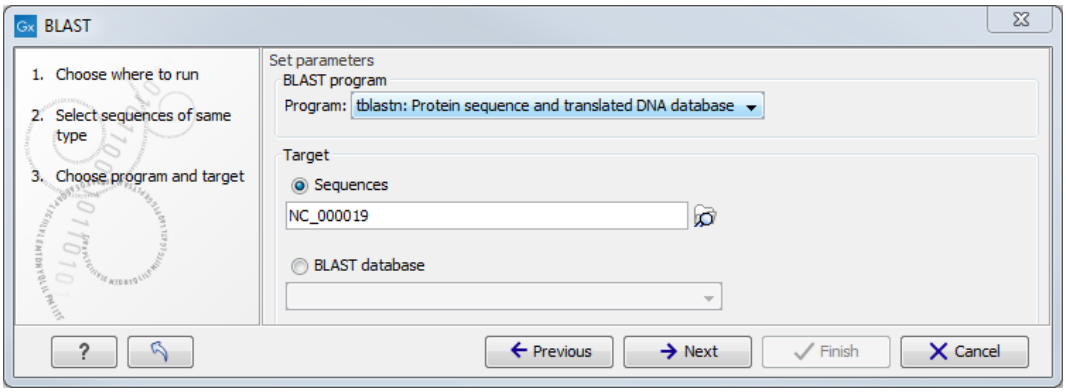

Figure 4: *Select the human chromosome 19 sequence and the blastn option.*

5. In the parameters dialog, leave all options at their default values (figure 5). Click **Next**.
6. Choose to **Open** your results and then **Finish**.

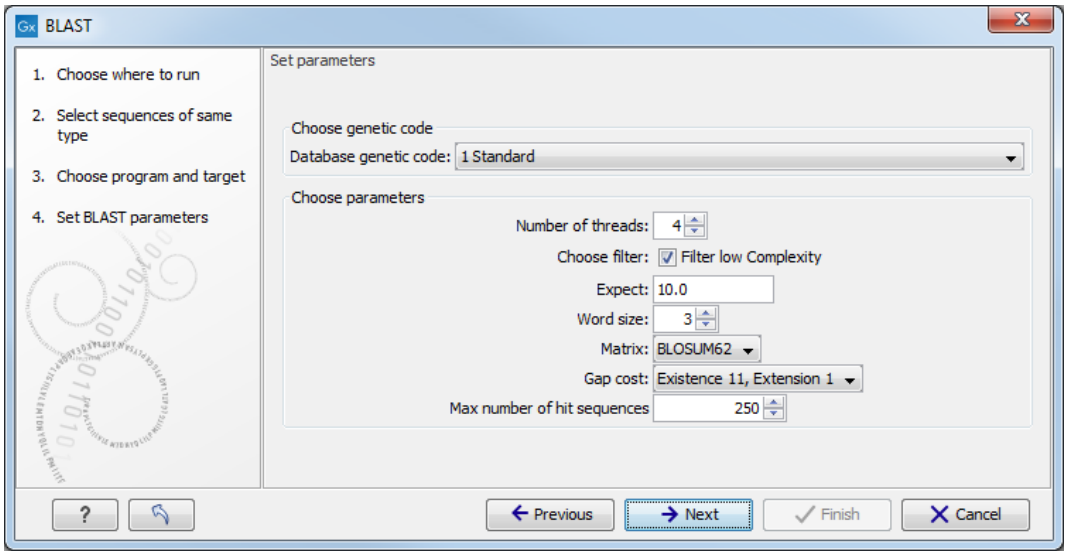

Figure 5: *Leave BLAST parameters as default.*

7. When the BLAST result appears make a split view so that both the table and graphical view is visible (see figure 6). This is done by pressing Ctrl (⌘ on Mac) while clicking the table view ( ) at the bottom of the view.

8. In the side panel under BLAST table settings, select the same settings as shown in figure 6 by checking the relevant parameters under "Show column" (Hit, E-value, HSP start, HSP end, Query start, Query end, %Identity, %Positive).

9. Now, sort the BLAST table view by clicking twice on the column header "% Positive". Then, press and hold the Ctrl button (⌘ on Mac) and click the header "Query start". You have now sorted the table first on % Positive hits and then on the start position of the query sequence. In the table you can see three hits with 97 - 100% positive (=similar residues) at different locations on the chromosome sequence (see figure 6).

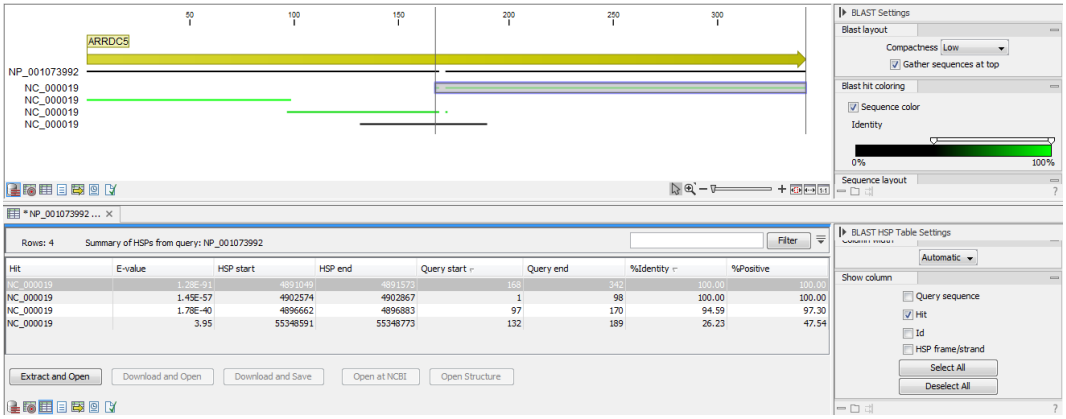

Figure 6: *Placement of translated nucleotide sequence hits on the Human ARRDC5 protein.*

Why did we find, on the protein level, three identical regions between our query protein sequence and nucleotide database?

The ARRDC5 gene is known to have three exons, which is in agreement with the three hits in the BLAST search. Each translated exon will hit the corresponding sequence on the chromosome.

If you place the mouse cursor on the sequence hits in the graphical view, you can see the reading frame, which is -1, -3 and -3 for the three hits, respectively.

Verify the result Open NC_0000019 in a view, and go to Hit position 4,902,879 (use the right hand side panel to find it) and zoom to see the blue gene annotation. You can now see the exon structure of the ARRDC5 gene showing the three exons on the reverse strand (see figure 7).

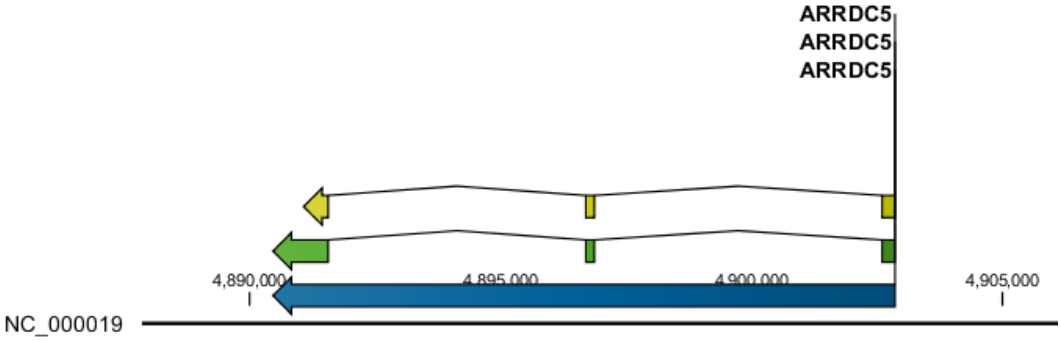

Figure 7: *ARRDC5 exon view.*

If you wish to verify the result:

1. Make a selection covering the gene region and open it in a new view:
   right-click | **Open Selection in New View (****)** | **Save (****)**
2. Save the sequence and perform a new BLAST search: **Toolbox** | **BLAST** () | **BLAST (****)**
3. Use the new sequence as query(figure 8) and click **Next**.

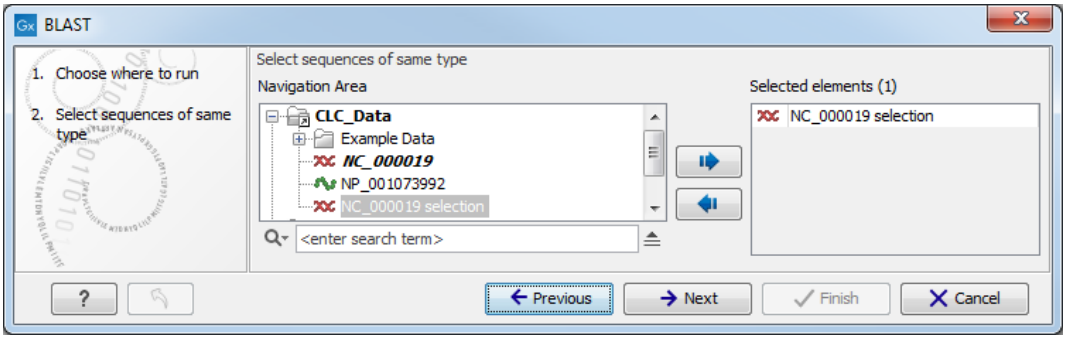

Figure 8: *Select the sequence selection you want to verify.*

4. Use the program blastx and the protein sequence NP_001073992 as target sequence (figure 9). Click **Next**.
5. Leave the BLAST parameters as default and click **Next**.
6. Choose to **Open** the results and click **Finish**.

Using the genomic sequence as query, the mapping of the protein sequence to the exons is visually very clear as shown in figure 10.

In theory you could use the chromosome sequence as query, but the performance would not be optimal: it would take a long time, and the computer might run out of memory.

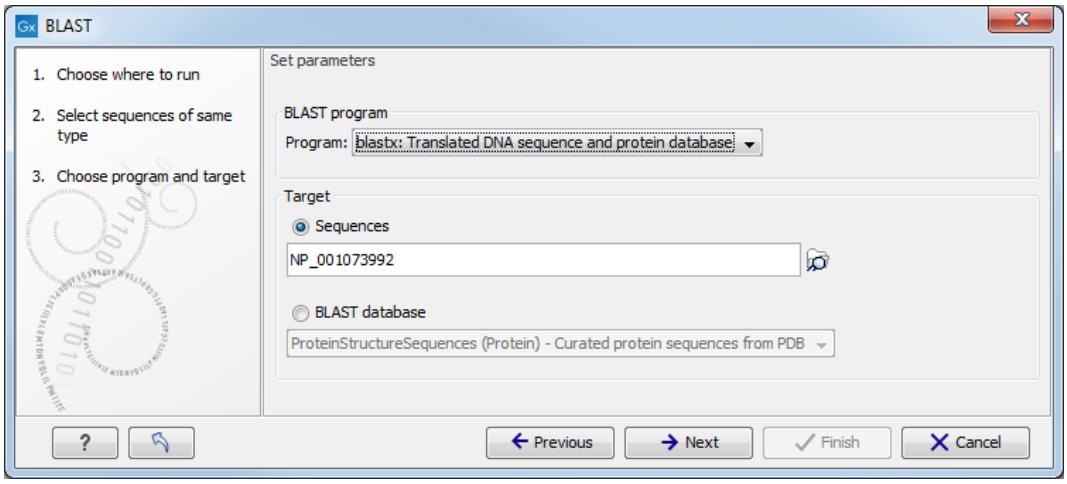

Figure 9: *Select the protein sequence NP_001073992 and the blastx option.*

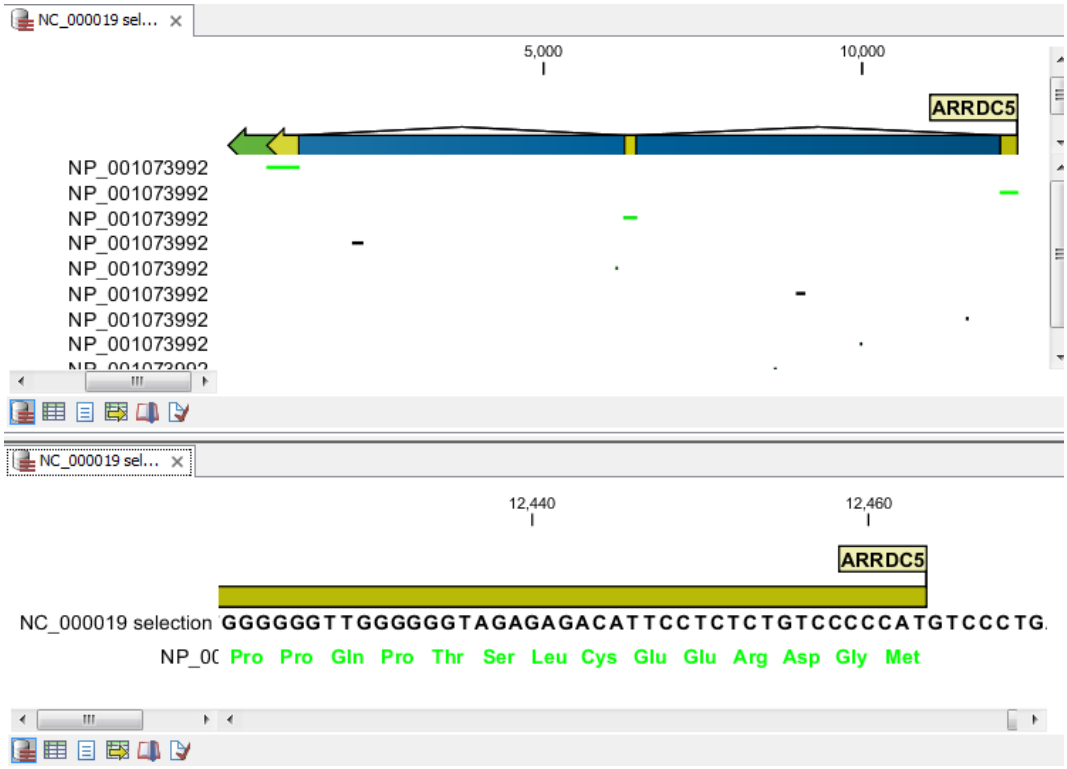

Figure 10: *Verification of the result: at the top a view of the whole BLAST result. At the bottom the same view zoomed in on exon 3 to show the amino acids.*

In this example you have used well-annotated sequences where you could have searched for the name of the gene instead of using BLAST. However, there are other situations where you either do not know the name of the gene, or the genomic sequence is poorly annotated. In these cases, the approach described in this tutorial can be very productive.

Additional BLAST tools

Create BLAST Database Instead of using the protein sequence NP_001073992 as target sequence you can also use it as database. To do this, you first have to create a database. BLAST databases can be created from DNA, RNA, and protein sequences located in the **Navigation**

Area. Any given BLAST database can only include one molecule type. If you wish to use a pre-formatted BLAST database instead, see above section on Download BLAST Database.

To create a BLAST database, go to:

Toolbox | **BLAST (**) | **Create BLAST Database (**)

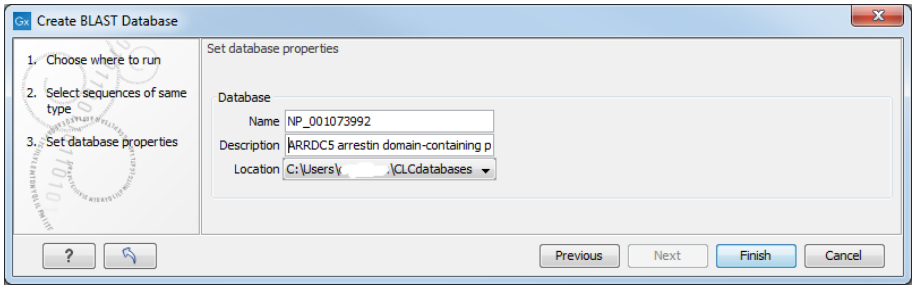

Figure 11: *When using Create BLAST Database, a database of the specified sequences files (and/or sequence lists) will be created.*

Download BLAST Database The Download BLAST Database allows you to fetch copies of specific databases stored at NCBI FTP-site. Examples include 16S microbial sequences, environmental nucleotide or protein sequences, EST- or genomic data data from mouse, human, other organisms, ref sequences etc. (figure 12).

The download location can be specified, and downloaded databases can be managed subsequently using the Manage BLAST Databases (see following section).

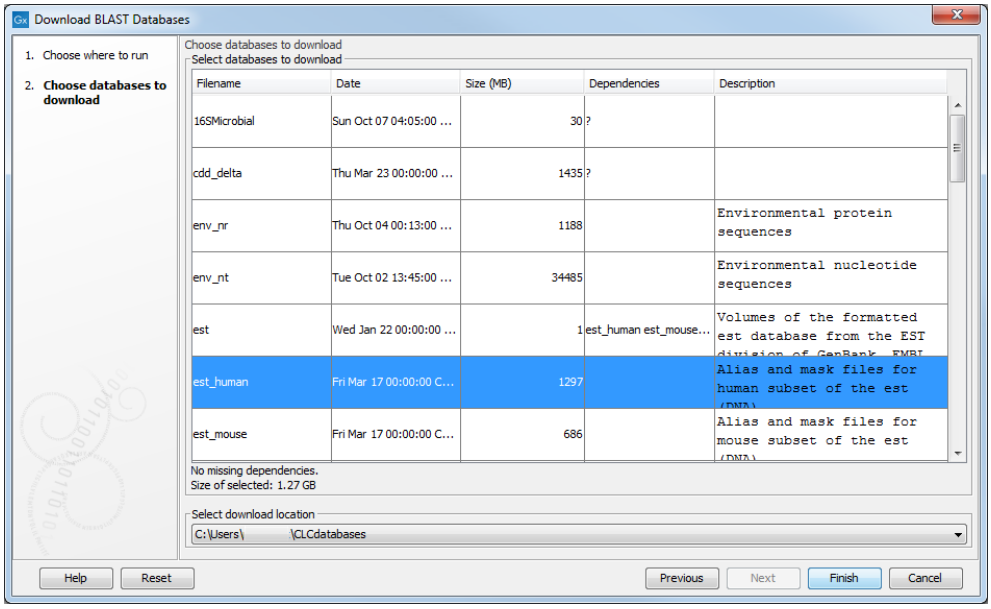

Figure 12: *Topic specific databases can be downloaded from NCBI's FTP server, and stored locally.*

BLAST Database Manager 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. Once downloaded, it is easy to get an overview of the local databases as well as to modify them (add- and refresh location as well as remove database, see figure 13).

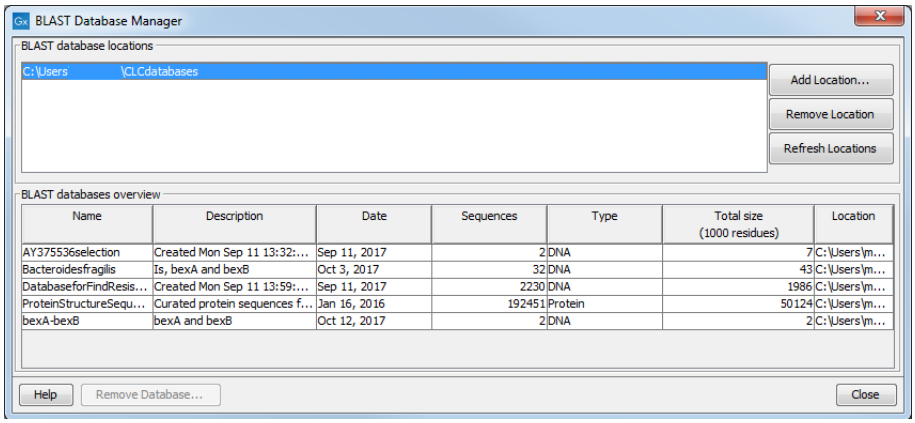

Figure 13: *The BLAST Database Manager provides an overview of available BLAST databases, their description, size and location, as well as the option to change location.*

BLAST for primer binding sites You can adjust the BLAST parameters so it becomes possible to match short primer sequences against a larger sequence. Then it is easy to examine whether already existing lab primers can be reused for other purposes, or if the primers you designed are specific.

Purpose	Program	Word size	Low complexity filter	Expect value
Standard BLAST	blastn	11	On	10
Primer search	blastn	7	Off	1000

These settings are shown in figure 14.

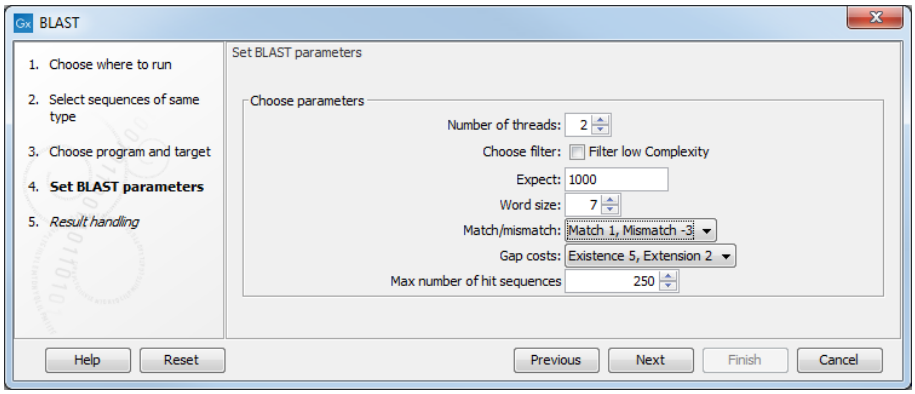

Figure 14: *Settings for searching for primer binding sites.*

Performing a BLAST at NCBI search

It is possible to download a folder of "Example Data" in your Navigation Area by going to

Help | **Import Example Data** .

This folder contains - among other - the ATP8a1 protein sequence, which is a phospholipid-transporting ATPase expressed in the adult house mouse, *Mus musculus*

To obtain more information about this molecule:

1. Go to: **Toolbox** | **BLAST (**) | **BLAST at NCBI (**)

2. select protein ATP8a1 to use as query sequence (figure 15). Click **Next**.

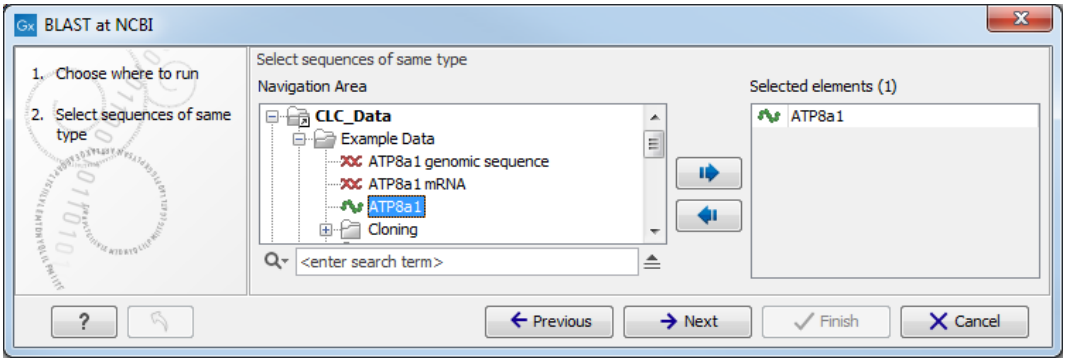

Figure 15: *Select the ATP8a1 protein sequence.*

3. Choose the default BLAST program: **blastp: Protein sequence and database** (figure 16 and select the **Swiss-Prot** database in the **Database** drop down menu. Click **Next**.

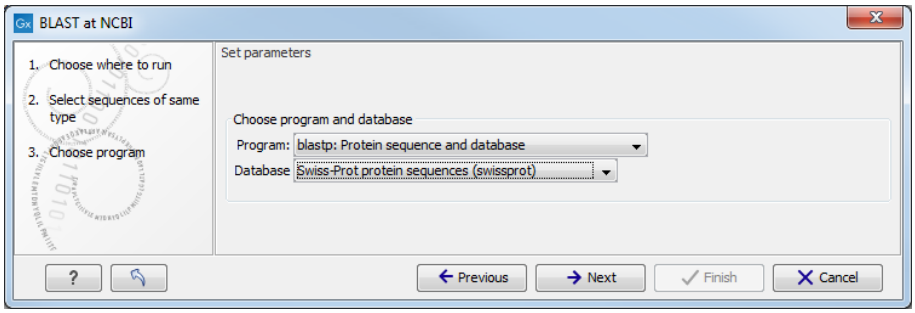

Figure 16: *Choosing BLAST program and database.*

4. In the BLAST parameters wizard window, set **Limit by Entrez query** to **Homo sapiens[ORGN]** from the drop down menu (figure 17). Including this term will limit the query to proteins of human origin. Click **Next**.

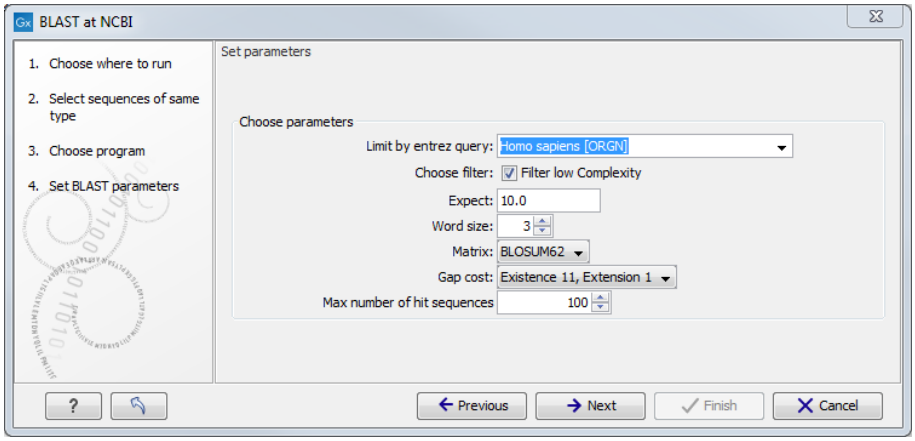

Figure 17: *The BLAST search is limited to "homo sapiens [ORGN]". The remaining parameters are left as default.*

5. Choose to **Open** your results and click **Finish** to accept the parameter settings and begin the BLAST search.

The computer now contacts NCBI and places your query in the BLAST search queue. After a short while the result should be received and opened in a new view.

Inspecting the results The output is shown in figure 18 and consists of a list of potential homologs that are sorted by their BLAST match-score and shown in descending order below the query sequence.

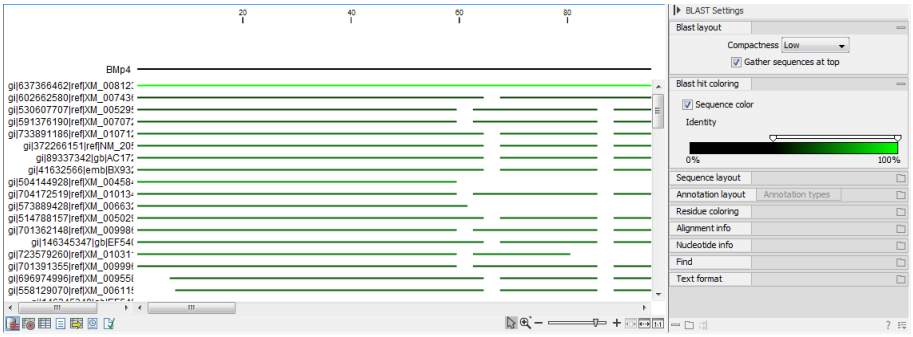

Figure 18: *Output of a BLAST search. By holding the mouse pointer over the lines you can get information about the sequence.*

Try placing your mouse cursor over a potential homologous sequence. You will see that a context box appears containing information about the sequence and the match-scores obtained from the BLAST algorithm.

The lines in the BLAST view are the actual sequences that have been downloaded. This means that you can zoom in and see the actual alignment. To zoom in, use the zoom to base level button ( ).

Now we will focus our attention on sequence Q9Y2Q0 - the BLAST hit that is at the top of the list. To download the full sequence, right-click the line representing sequence Q9Y2Q0 and click on **Download and Open**.

This opens the sequence. However, the sequence is not saved yet. Drag and drop the sequence into the **Navigation Area** to save it. This homologous sequence is now stored in the *CLC Main Workbench* and you can use it to gain information about the query sequence by using the various tools of the workbench, e.g., by studying its textual information, by studying its annotation or by aligning it to the query sequence.

Using the BLAST table view As an alternative to the graphic BLAST view, you can click the Table View ( ) at the bottom. This will display a tabular view of the BLAST hits as shown in figure 19.

This view provides more statistics about the hits, and you can use the filter to search for a specific type of protein for example. If you wish to download several of the hit sequences, this is easily done in this view. Simply select the relevant sequences and drag them into a folder in the **Navigation Area**.

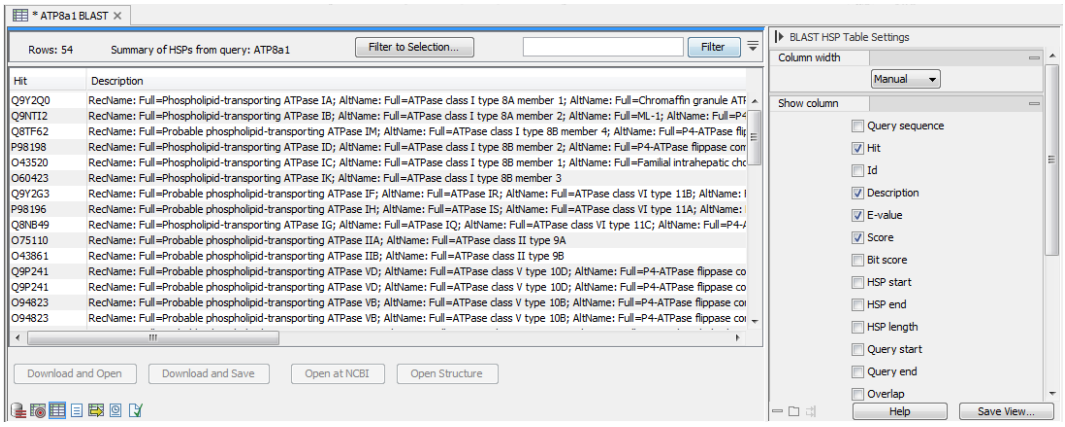

Figure 19: *Output of a BLAST search shown in a table.*