(P)

SignalP
and TMHMM
Plugin

USER MANUAL

User manual for

SignalP and TMHMM 24.0

Windows, macOS and Linux

January 5, 2024

This software is for research purposes only.

QIAGEN Aarhus
Silkeborgvej 2
Prismet
DK-8000 Aarhus C
Denmark

Contents

Chapter 1

Introduction

The SignalP and TMHMM plugin installs the Signal Peptide Prediction and Transmembrane Helix Prediction tools in the Toolbox of your CLC Workbench (figure 1.1).
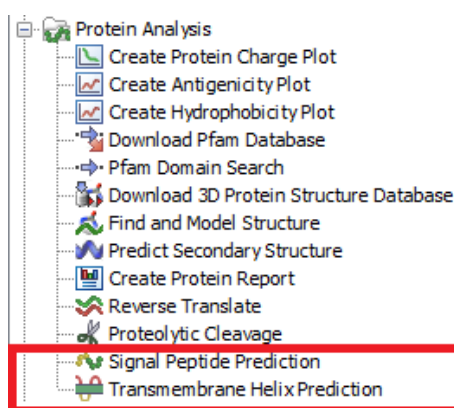

Figure 1.1: *Tools from the SignalP and TMHMM plugin can be found in the Protein Analysis folder of the Toolbox.*

1.1 SignalP

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.

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.

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 1.2.
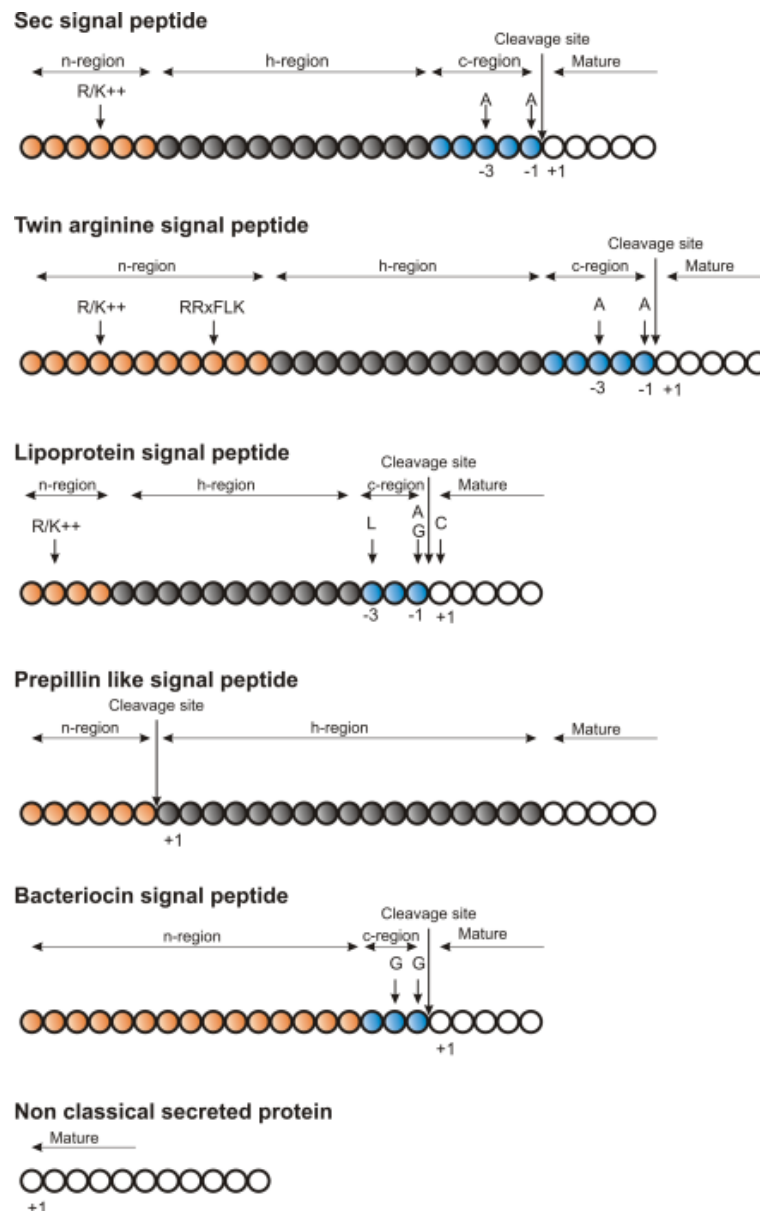
Figure 1.2: *Schematic representation of various signal peptides. Red color indicates n-region, gray color indicates h-region, cyan indicates c-region. All white circles are part of the mature protein. $+1$ indicates the first position of the mature protein. The length of the signal peptides is not drawn to scale.*

1.1.1 The SignalP method

One method for prediction of classical signal peptides is the SignalP method [Nielsen et al., 1997, Bendtsen et al., 2004]. SignalP predicts the actual cleavage site; thus the peptide which is cleaved off during translocation over the membrane. The tool uses SignalP 4.1 which is located at `https://services.healthtech.dtu.dk/service.php?SignalP-4.1`.

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 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.

1.1.2 Signal Peptide Prediction

When the plugin is downloaded and installed, you can use it to predict signal peptides located at https://services.healthtech.dtu.dk/service.php?TMHMM-2.0. In addition, an active internet connection is required to query SignalP [Nielsen et al., 1997, Bendtsen et al., 2004].

To run the signal peptide prediction, go to:

> **Toolbox | Classical Sequence Analysis ( ) | Protein 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.

You should select which organism group the input sequences belong to. the default is eukaryote (see figure 1.3).

- Eukaryote (default)
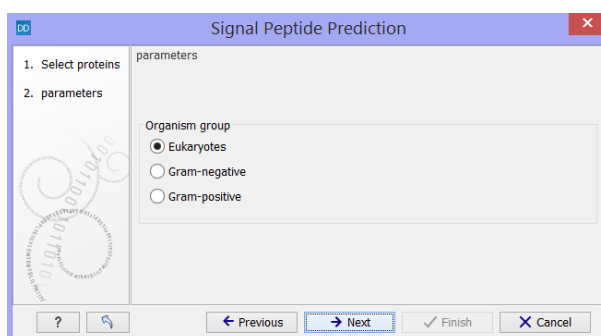
- Gram-negative bacteria

- Gram-positive bacteria

Figure 1.3: *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**.

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 1.1.1). This score has been shown to be the most accurate [Klee and Ellis, 2005] in an evaluation study of signal peptide predictors.

After running the prediction as described above, the protein sequence will show predicted signal peptide as annotations on the original sequence (see figure 1.4). Make sure the Side Panel 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. Undesired annotations can be removed through the **Delete Annotation (⟳)** right-click mouse menu.
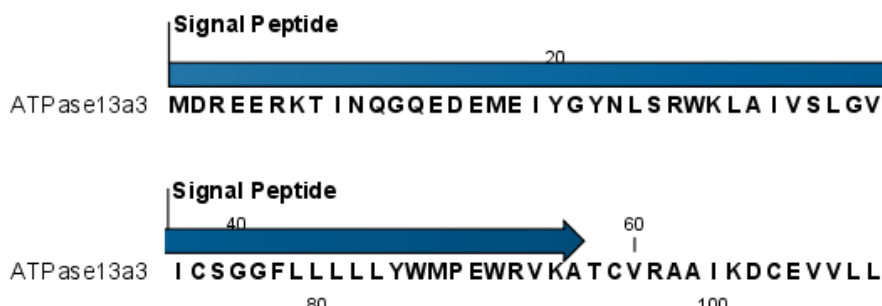
Figure 1.4: *N-terminal signal peptide shown as annotation on the sequence.*

1.2 TMHMM

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, we use TMHMM version 2.0 [Krogh et al., 2001] located at `https://services.healthtech.dtu.dk/service.php?TMHMM-2.0`, 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 `https://www.healthtech.dtu.dk` and in the original research paper [Krogh et al., 2001].

1.2.1 Transmembrane Helix Prediction

When the plugin is downloaded and installed, you can use it to predict transmembrane helices:

> **Toolbox** | **Classical Sequence Analysis (** 🖼 **) | Protein 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, then 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 1.5). Moreover,

annotations showing the topology will be shown, i.e., which part the proteins is located on the inside or on the outside.
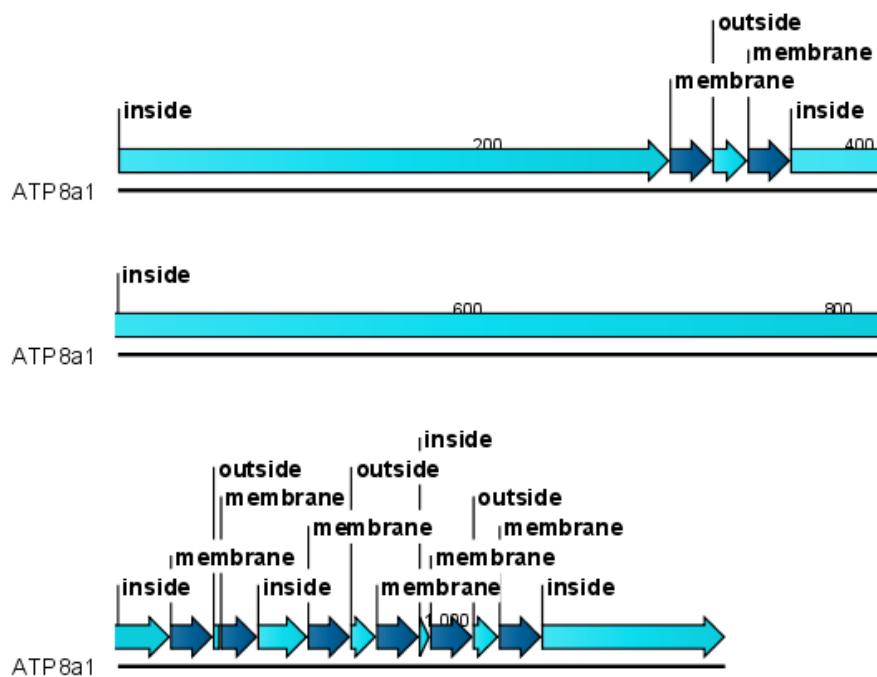

Figure 1.5: *Transmembrane segments shown as annotation on the sequence and the topology.*

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. Undesired annotations can be removed through the **Delete Annotation** (⤳) right-click mouse menu.

Chapter 2

Install and uninstall plugins

SignalP and TMHMM is installed as a plugin.

2.1 Installation of plugins

Note: In order to install plugins and modules, the *CLC Workbench* must be run in administrator mode. On Windows, you can do this by right-clicking the program shortcut and choosing "Run as Administrator". On Linux and Mac, it means you must launch the program such that it is run by an administrative user.

Plugins and modules are installed and uninstalled using the Workbench Plugin Manager. To open the Plugin Manager, click on the **Plugins ( ) button** in the top Toolbar, or go to the menu option:

Utilities | Manage Plugins... ( )

The Plugin Manager has two tabs at the top:

- **Manage Plugins** An overview of your installed plugins and modules is provided under this tab.

- **Download Plugins** Plugins and modules available to download and install are listed in this tab.

To install a plugin, click on the **Download Plugins** tab (figure 2.1). Select a plugin. Information about it will be shown in the right hand panel. Click on the **Download and Install** button to install the plugin.

Accepting the license agreement

The End User License Agreement (EULA) must be read and accepted as part of the installation process. Please read the EULA text carefully, and if you agree to it, check the box next to the text **I accept these terms**. If further information is requested from you, please fill this in before clicking on the **Finish** button.
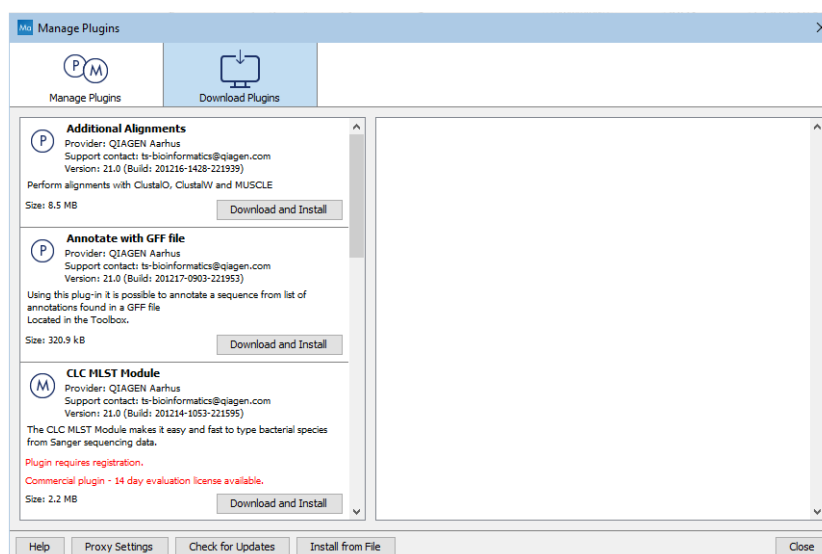
Installing a cpa file

Figure 2.1: *Plugins and modules available for installation are listed in the Plugin Manager under the Download Plugins tab.*

If you have a .cpa installer file for SignalP and TMHMM, you can install it by clicking on the **Install from File** button at the bottom of the Plugin Manager.

If you are working on a system not connected to the internet, plugin and module .cpa files can be downloaded from https://digitalinsights.qiagen.com/products-overview/plugins/using a networked machine, and then transferred to the non-networked machine for installation.

Restart to complete the installation

Newly installed plugins and modules will be available for use after restarting the software. When you close the Plugin Manager, a dialog appears offering the opportunity to restart the *CLC Workbench*.

2.2 Uninstalling plugins

Plugins and modules are uninstalled using the Workbench Plugin Manager.To open the Plugin Manager, click on the **Plugins (⊞) button** in the top Toolbar, or go to the menu option:

Utilities | Manage Plugins... (⊞)

This will open the Plugin Manager (figure 2.2). Installed plugins and modules are shown under the Manage Plugins tab of the Plugins Manager.

To uninstall a plugin or module, click on its entry in the list, and click on the **Uninstall** button.

Plugins and modules are not uninstalled until the Workbench is restarted. When you close the Plugin Manager, a dialog appears offering the opportunity to restart the *CLC Workbench*.

Disabling a plugin without uninstalling it

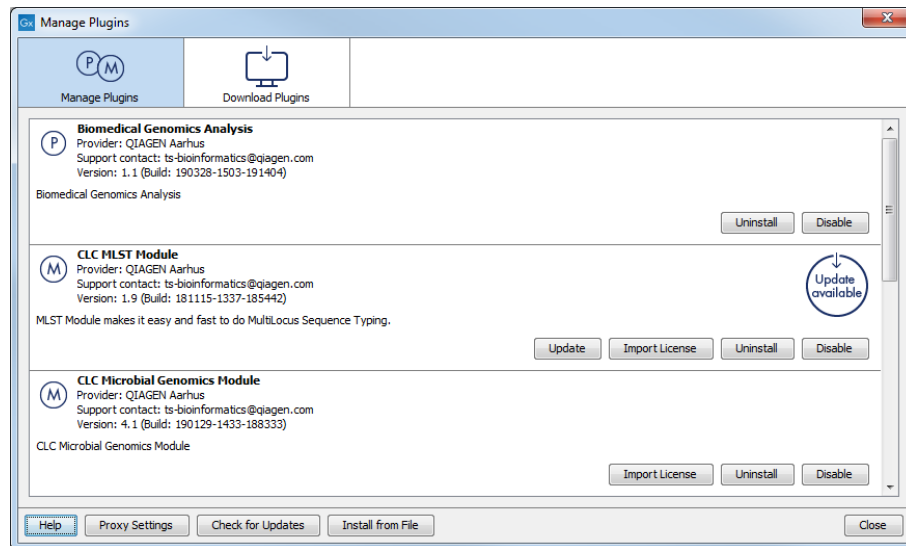If you do not want a plugin to be loaded the next time you start the Workbench, select it in the

Figure 2.2: *Installed plugins and modules are listed in the Plugins Manager under the Manage Plugins tab.*

list under the Manage Plugins tab and click on the **Disable** button.

Bibliography

[Bendtsen et al., 2004] Bendtsen, J. D., Nielsen, H., von Heijne, G., and Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*, 340(4):783–795.

[Blobel, 2000] Blobel, G. (2000). Protein targeting (Nobel lecture). *Chembiochem.*, 1:86–102.

[Klee and Ellis, 2005] Klee, E. W. and Ellis, L. B. M. (2005). Evaluating eukaryotic secreted protein prediction. *BMC Bioinformatics*, 6:256.

[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.

[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.