



CLC **Protein** Workbench

User manual

User manual for
CLC Protein Workbench 2.0
Windows, Mac OS X and Linux

July 6, 2006

CLC bio
Gustav Wieds Vej 10
Dk-8000 Aarhus C
Denmark



Contents

I	Introduction	8
1	Introduction to CLC Protein Workbench	9
1.1	Contact information	11
1.2	Download and installation	11
1.3	System requirements	14
1.4	Licenses	15
1.5	About CLC Workbenches	18
1.6	When the program is installed: Getting started	20
1.7	Network configuration	22
1.8	Adjusting the maximum amount of memory	22
1.9	The format of the user manual	24
2	Tutorials	25
2.1	Tutorial: Starting up the program	26
2.2	Tutorial: View sequence	29
2.3	Tutorial: GenBank search and download	30
2.4	Tutorial: Align protein sequences	32
2.5	Tutorial: Create and modify a phylogenetic tree	34
2.6	Tutorial: Detect restriction sites	35
2.7	Tutorial: Sequence information	36
2.8	Tutorial: BLAST search	38
2.9	Tutorial: Proteolytic cleavage detection	40
2.10	Tips and tricks for the experienced user	41

II Basic Program Functionalities	50
3 User Interface	51
3.1 Navigation Area	52
3.2 View Area	58
3.3 Zoom and selection in View Area	63
3.4 Toolbox and Status Bar	66
3.5 Workspace	67
3.6 List of shortcuts	68
4 User preferences	70
4.1 General preferences	71
4.2 Default View preferences	71
4.3 Advanced preferences	72
4.4 Export/import of preferences	72
4.5 View preference style sheet	72
5 Printing	76
5.1 Selecting which part of the view to print	76
5.2 Page setup	77
5.3 Print preview	77
6 Import/export of data and graphics	79
6.1 Bioinformatic data formats	79
6.2 External files	84
6.3 Export graphics to files	85
6.4 Copy/paste view output	87
7 History	89
7.1 Element history	89
8 Handling of results	91
8.1 How to handle results of analyses	91

III Bioinformatics	94
9 Database search	95
9.1 GenBank search	95
9.2 UniProt (Swiss-Prot/TrEMBL) search	98
9.3 Sequence web info	101
10 BLAST Search	103
10.1 BLAST Against NCBI Database	103
10.2 BLAST Against Local Database	109
10.3 Create Local BLAST Database	110
11 Viewing and editing sequences	113
11.1 View sequence	113
11.2 Sequence information	123
11.3 View as text	124
11.4 Creating a new sequence	125
11.5 Sequence Lists	126
11.6 Circular DNA	128
12 3D molecule viewing	131
12.1 Importing structure files	131
12.2 Viewing structure files	132
12.3 The structure table	133
12.4 Options through the preference panel	134
12.5 3D Output	136
13 General sequence analyses	138
13.1 Dot plots	138
13.2 Shuffle sequence	148
13.3 Local complexity plot	149
13.4 Sequence statistics	151
13.5 Join sequences	158
13.6 Motif Search	159

13.7 Pattern Discovery	162
14 Nucleotide analyses	165
14.1 Convert DNA to RNA	165
14.2 Convert RNA to DNA	166
14.3 Reverse complements of sequences	167
14.4 Translation of DNA or RNA to protein	168
14.5 Find open reading frames	169
15 Protein analyses	172
15.1 Signal peptide prediction	173
15.2 Protein charge	179
15.3 Transmembrane helix prediction	181
15.4 Antigenicity	183
15.5 Hydrophobicity	185
15.6 Pfam domain search	190
15.7 Secondary structure prediction	193
15.8 Protein report	194
15.9 Reverse translation from protein into DNA	197
15.10 Proteolytic cleavage detection	201
16 Restriction site analyses	206
16.1 Restriction sites and enzyme lists	206
16.2 Restriction site analysis	206
16.3 Restriction enzyme lists	210
16.4 Gel electrophoresis	212
17 Sequence alignment	216
17.1 Create an alignment	217
17.2 View alignments	222
17.3 Edit alignments	225
17.4 Join alignments	227
17.5 Bioinformatics explained: Multiple alignments	229

18 Phylogenetic trees	232
18.1 Inferring phylogenetic trees	232
18.2 Bioinformatics explained: phylogenetics	235
 IV Appendix	 240
A Comparison of workbenches	241
 B BLAST databases	 244
B.1 Peptide sequence databases	244
B.2 Nucleotide sequence databases	244
 C Proteolytic cleavage enzymes	 246
 D Formats for import and export	 248
D.1 List of bioinformatic data formats	248
D.2 List of graphics data formats	249
 Bibliography	 250
 V Index	 254

Part I

Introduction

Chapter 1

Introduction to *CLC Protein Workbench*

Contents

1.1	Contact information	11
1.2	Download and installation	11
1.2.1	Program download	11
1.2.2	Installation on Microsoft Windows	12
1.2.3	Installation on Mac OS X	13
1.2.4	Installation on Linux with an installer	13
1.2.5	Installation on Linux with an RPM-package	14
1.3	System requirements	14
1.4	Licenses	15
1.4.1	Demo license description	15
1.4.2	Getting and activating the demo license	15
1.4.3	Commercial license	17
1.4.4	Upgrading from a demo license to a commercial license	18
1.5	About CLC Workbenches	18
1.5.1	New program feature request	19
1.5.2	Report program errors	19
1.5.3	Free vs. commercial workbenches	19
1.6	When the program is installed: Getting started	20
1.6.1	Basic concepts of using CLC Workbenches	20
1.6.2	Quick start	21
1.6.3	Import of example data	22
1.7	Network configuration	22
1.8	Adjusting the maximum amount of memory	22
1.8.1	Microsoft Windows	23
1.8.2	Mac OS X	23
1.8.3	Linux	23
1.9	The format of the user manual	24
1.9.1	Text formats	24

Welcome to *CLC Protein Workbench 2.0* — a software package supporting your daily bioinformatics work.

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

1.1 Contact information

The *CLC Protein Workbench 2.0* is developed by:

CLC bio A/S
Science Park Aarhus
Gustav Wieds Vej 10
8000 Aarhus C
Denmark

<http://www.clcbio.com>

VAT no.: DK 28 30 50 87

Telephone: +45 70 22 32 44

Fax: +45 86 20 12 22

E-mail: info@clcbio.com

If you have questions or comments regarding the program, you are welcome to contact our support function:

E-mail: support@clcbio.com

1.2 Download and installation

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

Furthermore the program can be sent on a CD-Rom by regular mail. To receive the program by regular mail, please write an e-mail to support@clcbio.com, including your postal address.

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 want to include Java or not
(this is necessary if you haven't already installed Java)
- 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.

Download

CLC Protein Workbench 1.5.2

Email

Name

Department

Company / Institution

Mac OS X 10.3 or later (including Intel-based Macs)
35MB disc-image (.dmg)

Windows 2000 or Windows XP
38MB installer (.exe)

☐ Include Java (approximately 15MB extra)

Linux (RedHat/SuSE) installer
32MB installer (.sh)

Linux (RedHat/SuSE) RPM
32MB rpm-package (.rpm)

☐ Include Java (approximately 15MB extra)

Email notifications
☒ Mark this field if you would like to know about new software releases and other relevant bioinformatics information.

Download

Figure 1.1: Download dialog.

1.2.2 Installation on Microsoft Windows

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

If you have downloaded an installer:

Locate the downloaded installer and double-click the icon.

The default location for downloaded files is your desktop.

If you are installing from a CD:

Insert the CD into your CD-ROM drive.

Choose the "Install CLC Protein Workbench" from the menu displayed.

If you already have Java installed on your computer you can choose "Install CLC Protein Workbench without Java".

Installing the program is done in the following steps:

(you must be connected to the Internet throughout the installation process.)

- 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 Protein Workbench and click Next.

- Choose where you would like to create shortcuts for launching *CLC Protein Workbench* and click Next.
- Wait for the installation process to complete, choose whether you would like to launch *CLC Protein 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 choose to create.

1.2.3 Installation on Mac OS X

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

If you have downloaded an installer:

Locate the downloaded installer and double-click the icon.
The default location for downloaded files is your desktop.

If you are installing from a CD:

Insert the CD into your CD-ROM drive and open it by double-clicking on the CD icon on your desktop.

Launch the installer by double-clicking on the "*CLC Protein 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 whether you would like to create desktop icon for launching *CLC Protein Workbench* and click Next.
- Wait for the installation process to complete, choose whether you would like to launch *CLC Protein 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 choose 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 CLCProteinWorkbench_1_5_2_JRE.sh.sh
```

If you are installing from a CD the installers are located in the "linux" directory.

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:

```
# clcproteinwb
```

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

```
# ./clcproteinwb
```

1.2.5 Installation on Linux with an RPM-package

Navigate to the directory containing the rpm-package and install it using the rpm-tool by running a command similar to:

```
# rpm -ivh CLCProteinWorkbench_1_5_2_JRE.sh.rpm
```

If you are installing from a CD the rpm-packages are located in the "RPMS" directory. Installation of RPM-packages usually requires root-privileges.

When the installation process is finished the program can be executed by running the command:

```
# clcproteinwb
```

1.3 System requirements

The system requirements of CLC Protein Workbench 2.0 are these:

- Windows 2000 or Windows XP
- Mac OS X 10.3 or newer
- Linux: Redhat or SuSE

- 256 MB RAM required
- 512 MB RAM recommended
- 1024 x 768 display recommended

1.4 Licenses

The license system of *CLC Protein Workbench 2.0* is based on a license key which is unique for the computer rather than for the user of the workbench.

1.4.1 Demo license description

We offer a fully functional demo version of *CLC Protein Workbench 2.0* to all users, free of charge.

Each user is entitled to four weeks demo of *CLC Protein Workbench 2.0*. In order to make your demo time as valuable as possible, the four weeks can be separated. You can e.g. try two weeks of the demo in January, and the next two weeks in March.

To prevent unauthorized use of the program, you must be connected to the Internet while starting up a demo version of *CLC Protein Workbench*. An additional online check will be conducted 24 hours after the launch of the workbench. After running *CLC Protein Workbench 2.0* for 24 hours, if you are not connected to the Internet, you will be met with the dialog shown in figure 1.2.

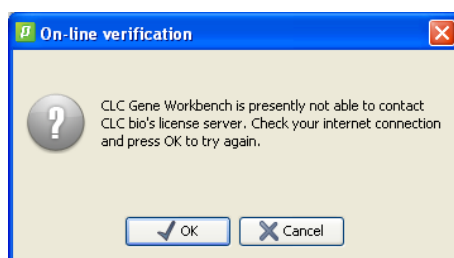


Figure 1.2: This dialog appears when an online license check is conducted by **CLC Protein Workbench**, and the computer is off line. Either at start-up or after 24 hours.

You can then connect to the Internet and retry or you can save your work and close the program. You can run the workbench again later, as long as you are connected to the Internet at start-up.

We use the concept of "*quid quo pro*". The last two weeks of free demo time given to you is therefore accompanied by a short-form questionnaire where you have the opportunity to give us feedback about the program.

The four weeks demo is offered for each major release of *CLC Protein Workbench*. You will therefore have the opportunity to try the next version (*CLC Protein Workbench 2.0.1*) is released. (If you purchase *CLC Protein Workbench* the first year of updates is included)

1.4.2 Getting and activating the demo license

When you start the program for the first time, you will be presented with the dialog shown in figure 1.3.

If you connect to the internet via a proxy server, click the proxy settings button. Otherwise, just click the "Request evaluation license" button in order to get a license key for a demo of *CLC*

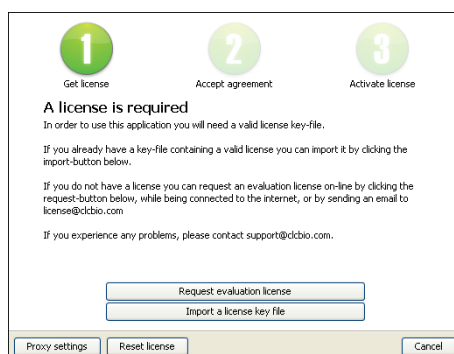


Figure 1.3: Selecting "Request evaluation license".

Protein Workbench 2.0.

Now, our server will issue an evaluation license. This process might take a while depending on your internet connection. When the license key is received, you will be asked to accept the **License agreement** shown in figure 1.4.

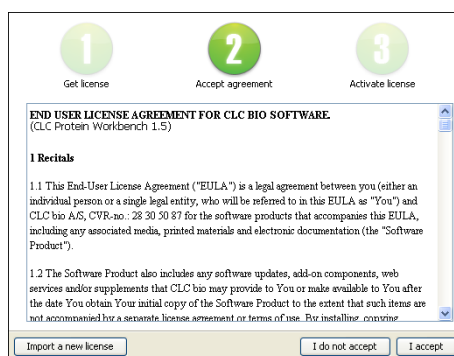


Figure 1.4: License Agreement.

Please read the License agreement carefully before clicking **I accept**. In the next step shown in figure 1.5, select "Activate license on-line". Again, you might have to wait for a short while, because the license key is being activated on our server. A license is related to a specific computer, and therefore it can be used by anyone using that computer.

Like in figure 1.3 you can specify a proxy server if needed.

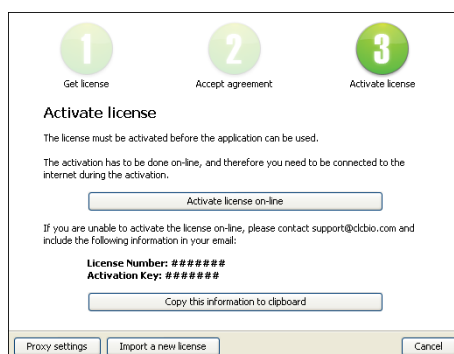


Figure 1.5: Activate the license key online.

Now the license key is activated on your computer, and CLC Protein Workbench 2.0 starts.

Problems with online activation

If you have problems activating the license online, *CLC Protein Workbench* also offers you an opportunity to manually activating your license key.

Step 3 of the license activation dialog provide a **License number** and an **Activation Key**. By clicking **Copy this information to the clipboard** you can open an email editor and paste these two numbers into the mail. If you email this content and a short explanation to support@clcbio.com we will send back a pre-activated license key.

Also, in all steps of the license dialog you have an option of resetting the license. This will allow you to start over, importing another license. However, information about which licenses were used on the computer is stored externally to prevent unauthorized use of demo licenses.

1.4.3 Commercial license

Unlike the demo version, the commercial version is fully functional offline. When you buy a license for *CLC Protein Workbench*, we will provide you with a license key which is activated as described here.

Start the program, and the dialog shown in figure 1.6 will appear:

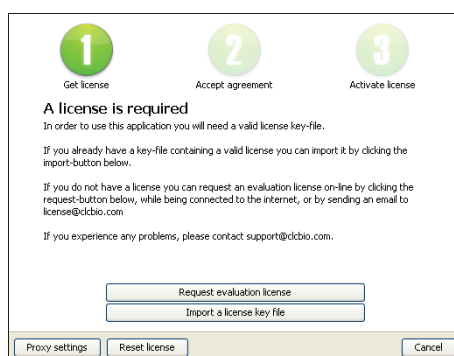


Figure 1.6: Select "Import a license key file".

Choose the option "Import a license key file" in order to specify where your license key is located. Select the license key file provided by CLC bio. When you have selected this file, the **License Agreement** is shown (see figure 1.7). If you want to use another license key instead, click the "Import a license key file" button.

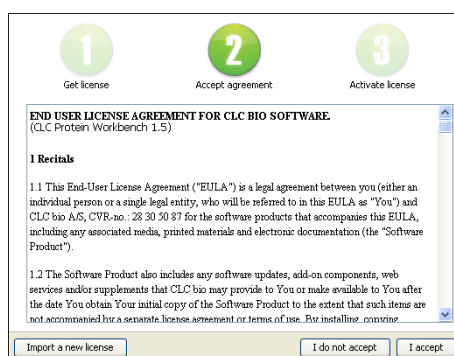


Figure 1.7: Read the License Agreement carefully.

Read the **License Agreement** carefully before clicking the "I accept" button. In the next step shown in figure 1.8, click the "Activate license on-line" button. Your computer must be connected to the internet in order to activate the license. Once the license is activated, you may work

off-line. It will take a little time to activate the license key. When the license key is activated, *CLC Protein Workbench 2.0* will start.

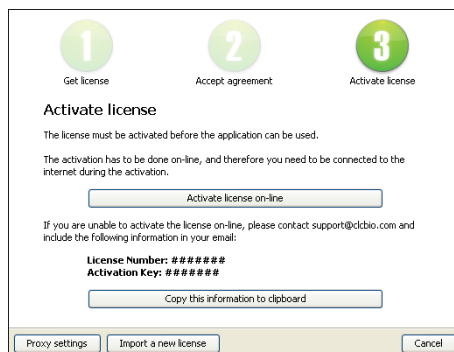


Figure 1.8: Activate the license key online.

A license is related to a specific computer, and therefore it can be used by anyone using that computer. If at some time you want to transfer the license to another computer, please contact license@clcbio.com.

Problems with online activation

If you have problems activating the license online, *CLC Protein Workbench* also offers you an opportunity to manually activating your license key.

Step 3 of the license activation dialog provide a **License number** and an **Activation Key**. By clicking **Copy this information to the clipboard** you can open an email editor and paste these two numbers into the mail. If you email this content and a short explanation to support@clcbio.com we will send back a pre-activated license key.

Also, in all steps of the license dialog you have an option of resetting the license. This will allow you to start over, importing another license. However, information about which licenses were used on the computer is stored externally to prevent unauthorized use of demo licenses.

1.4.4 Upgrading from a demo license to a commercial license

If you are trying a demo of *CLC Protein Workbench* and want to upgrade to a license that you have bought, choose **Upgrade license** in the Help menu. Then follow the description in section 1.4.3.

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 Gene Workbench* and *CLC Combined Workbench* were added to the product portfolio of CLC bio. Like *CLC Protein Workbench*, *CLC Gene 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 Combined Workbench* holds all basic and advanced features of the *CLC Workbenches*.

For an overview of which features the four workbenches include, see <http://www.clcbio.com/features>.

All workbenches will be improved continuously. If you have a CLC Free Workbench or a commercial workbench, and you are interested in receiving news about updates, you should register your e-mail and contact data on <http://www.clcbio.com>, if you haven't already registered when you downloaded the program.

1.5.1 New program feature request

The CLC team is continuously improving the program with our users' interest in mind. Therefore, we welcome all requests from users, and they can be submitted from our homepage <http://www.clcbio.com>. Likewise, you are more than welcome to suggest new features or more general improvements to the program on support@clcbio.com.

1.5.2 Report program errors

CLC bio is doing everything possible to eliminate program errors. Nevertheless, some errors might have escaped our attention. If you discover an error in the program, you can use the **Report a Program Error** function in the **Help** menu of the program to report it. In the **Report a Program Error** dialog you are asked to write your e-mail address. This is because we would like to be able to contact you for further information about the error or for helping you with the problem.

Notice that no personal information is sent via the error report. Only the information which can be seen in the **Program Error Submission Dialog** is submitted.

You can also write an e-mail to support@clcbio.com. Remember to specify how the program error can be reproduced.

All errors will be treated seriously and with gratitude.

We appreciate your help.

Start in safe mode

If the program becomes unstable on start-up, you can start it in **Safe mode**. This is done by pressing 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.

1.5.3 Free vs. commercial workbenches

The advanced analyses of the commercial workbenches, *CLC Protein Workbench* and *CLC Gene Workbench* are not present in *CLC Free Workbench*. Likewise, some advanced analyses are available in *CLC Gene Workbench* but not in *CLC Protein Workbench*, and visa versa. All types of basic and advanced analyses are available in *CLC Combined Workbench*.

However, the output of the commercial workbenches can be viewed in all other workbenches. This allows you to share the result of your advanced analyses from e.g. *CLC Combined Workbench*, with people working with e.g. *CLC Free Workbench*. They will be able to view the results of your analyses, but not redo the analyses.

The CLC Workbenches are developed for Windows, Mac and Linux platforms. Data can be

exported/imported between the different platforms in the same easy way as when exporting/importing between two computers with e.g. Windows.

This is illustrated in figure 1.9.

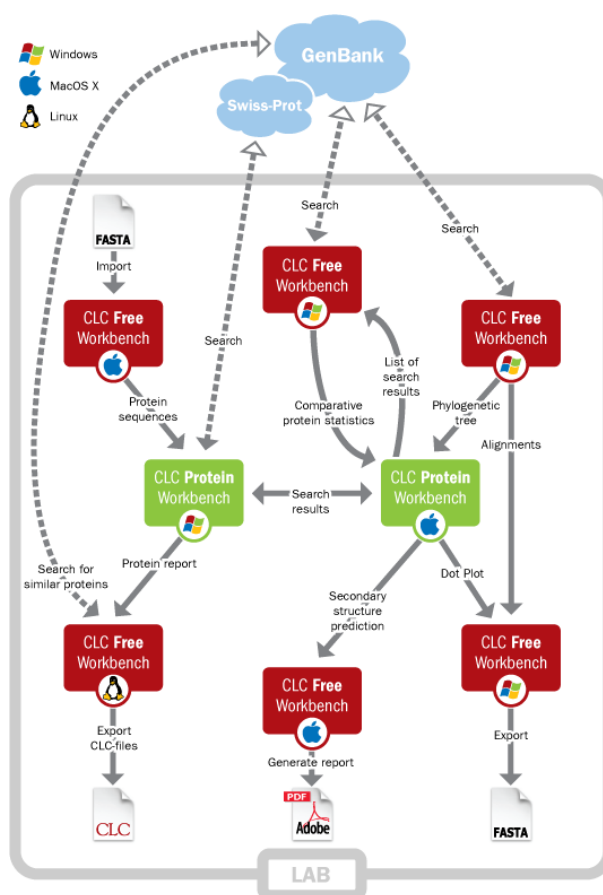


Figure 1.9: An example of how research can be organized and how data can flow between users of different workbenches, working on different platforms.

1.6 When the program is installed: Getting started

CLC Protein Workbench 2.0 includes an extensive **Help** function, which can be found in the **Help** menu of the program's **Menu bar**. The **Help** function can also be launched by pressing F1. The help topics are sorted in a table of contents and the topics can be searched.

1.6.1 Basic concepts of using CLC Workbenches

Here is a short list of basic concepts of how to use CLC Protein Workbench:

- All data for use in the CLC Protein Workbench should be stored inside the program in the **Navigation Area**. This means that you have to either import some of your own data or use e.g. the GenBank search function (🔍).
- The data can be viewed in a number of ways. First, click the element (e.g. a sequence) in the **Navigation Area** and then click **Show** (📄) to find a proper way to view the data (see figure 1.10 for an example).

- When a view is opened, there are three basic ways of interacting:
 1. Using the **Side Panel** to the right to specify how the data should be displayed (these settings are not associated with your data but they can be saved by clicking the icon (☰) in the upper right corner of the **Side Panel**).
 2. Using right-click menus e.g. to edit a sequence (in this case you have to make a selection first using the selection mode(☞)).
 3. Using the Zoom (☞) / (☞) tools.
- In the Toolbox, you find all the tools for analyzing and working on your data. In order to use these tools, your data must be stored in a project in the **Navigation Area**

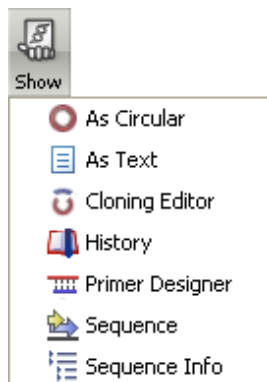


Figure 1.10: The different ways of viewing DNA sequences.

1.6.2 Quick start

When the program opens for the first time, the background of the workspace is visible. In the background are three quick start shortcuts, which will help you getting started. These can be seen in figure 1.11.

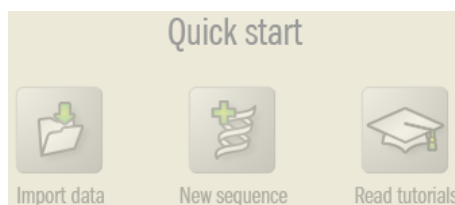


Figure 1.11: Three available Quick start short cuts, available in the background of the workspace.


The function of the three 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.
- **New sequence.** Opens a dialog which allows you to enter your own sequence.
- **Read tutorials.** Opens the tutorials a menu with a number of tutorials. These are also available from the **Help** menu in the **Menu bar**.

It might be easier to understand the logic of the program by trying to do simple operations on existing data. Therefore *CLC Protein Workbench 2.0* includes an example data set, which can be found on our web page, or downloaded from the program (Also found in the **Help** menu).

1.6.3 Import of example data

When downloading *CLC Protein Workbench 2.0* you are asked if you would like to import an 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  **Install example data** in the **Help** menu of the program. This installs the data automatically. You can also go to our website at <http://www.clcbio.com>, Software/CLC Free Workbench/Example data, 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 6.1 for more about importing data.

1.7 Network configuration

If you use a proxy server to access the Internet you must configure *CLC Protein Workbench 2.0* to use this. Otherwise you will not be able to perform any on-line activities (e.g. searching GenBank). *CLC Protein Workbench 2.0* supports the use of a HTTP-proxy and an anonymous SOCKS-proxy.

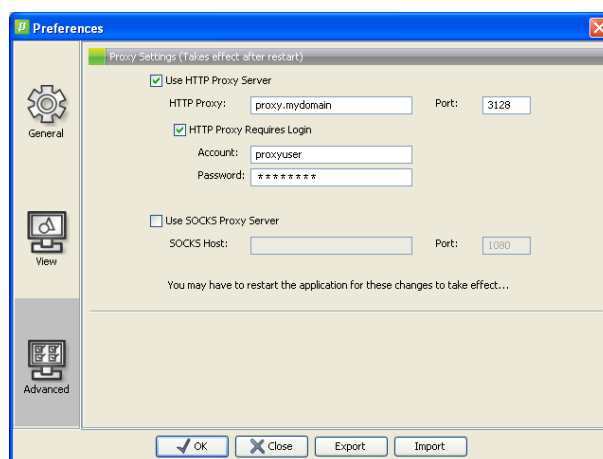


Figure 1.12: Adjusting proxy preferences.

To configure your proxy settings, open *CLC Protein Workbench 2.0*, and go to the **Advanced**-tab of the **Preferences** dialog (figure 1.12) and enter the appropriate information.

You have the choice between a HTTP-proxy and a SOCKS-proxy. *CLC Protein Workbench 2.0* only supports the use of a SOCKS-proxy that does not require authorization.

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

1.8 Adjusting the maximum amount of memory

If you have a large amount of memory (RAM) available in your system and need to work with very large data objects, you can manually change the maximum amount of memory available to the program. Doing so is a somewhat complicated, unsupported procedure and may cause the program to fail if done incorrectly.

Depending on your operating system you may have to repeat these changes if you update *CLC Protein Workbench 2.0* to a newer version.

1.8.1 Microsoft Windows

- Locate the *CLC Protein Workbench 2.0* directory inside your Program Files directory and open it
- Create a new, empty text-file called `clcwb.vmoptions` (make sure the filename does not end with ".txt")
- Add a single line to the file with a syntax similar to:

```
-Xmx512m
```

It is very important that the line looks exactly like the one in the example above, and that you only change the value of the number (512 in the example). For the best performance you should not choose a number greater than the amount (in megabytes) of physical memory available on your system.

1.8.2 Mac OS X

- Locate the *CLC Free Workbench* program file in your Applications folder
- Right-click / control-click the file and choose "Show Package Contents" from the pop-up menu
- Open the file called "Info.plist" located inside the "Contents" folder using the "Property List Editor" application or a text editor like "TextEdit"
- Edit the Root/Java/VMOptions property, and set the maximum amount of memory to a desired value. The property has a specific syntax similar to:

```
-Xmx512m
```

It is very important that you only change the value of the number, 512 in the example above, to the amount of megabytes you want. For the best performance you should not choose a number greater than the amount of physical memory available on your system.

1.8.3 Linux

- Locate the directory where you installed *CLC Protein Workbench 2.0* and open it.
- Create a new, empty text-file called "`clcwb.vmoptions`".
- Add a single line to the file with a syntax similar to:

```
-Xmx512m
```

It is very important that the line looks exactly like the one in the example above, and that you only change the value of the number (512 in the example). For the best performance you should not choose a number greater than the amount (in megabytes) of physical memory available on your system.

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 and some tutorials showing how to apply the most significant functionalities of *CLC Protein Workbench 2.0*.
- The **second part** describes in detail how to operate all the program's basic functionalities.
- The **third part** digs deeper into some of the 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 Protein Workbench 2.0* and provide more general knowledge of 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**)
- Icons, such as "", are included in order to ease the navigation in the **Toolbox**.
- The format of the program name is bold and italic: ***CLC Protein Workbench 2.0***
- The captions of displayed screenshots are in *italic*.

Chapter 2

Tutorials

Contents

2.1 Tutorial: Starting up the program	26
2.1.1 Creating a project and a folder	27
2.1.2 Import data	27
2.1.3 Supported data formats	27
2.2 Tutorial: View sequence	29
2.3 Tutorial: GenBank search and download	30
2.3.1 Saving the search	31
2.3.2 Searching for matching objects	31
2.3.3 Saving the sequence	32
2.4 Tutorial: Align protein sequences	32
2.4.1 Alignment dialog	32
2.5 Tutorial: Create and modify a phylogenetic tree	34
2.5.1 Tree layout	34
2.6 Tutorial: Detect restriction sites	35
2.6.1 View restriction site	35
2.7 Tutorial: Sequence information	36
2.8 Tutorial: BLAST search	38
2.9 Tutorial: Proteolytic cleavage detection	40
2.10 Tips and tricks for the experienced user	41
2.10.1 Open and arrange views using drag and drop	42
2.10.2 Find element in the Navigation Area	42
2.10.3 Find specific annotations on a sequence	43
2.10.4 Split sequences into several lines	44
2.10.5 Make a new sequence of a coding region	44
2.10.6 Translate a coding region	44
2.10.7 Copy annotations from one sequence to another	45
2.10.8 Get overview and detail of a sequence at the same time	45
2.10.9 Smart selecting in sequences and alignments	46
2.10.10 Check for updates and additional information about sequences	46
2.10.11 Quickly import sequences using copy-paste	47

2.10.12 Perform analyses on many elements	47
2.10.13 Drag elements to the Toolbox	48
2.10.14 Export elements while preserving history	48
2.10.15 Avoid the mouse trap - use keyboard shortcuts	49

This chapter contains tutorials representing some of the features of *CLC Protein Workbench 2.0*. The first tutorials are meant as a short introduction to operating the program. The last tutorials give examples of how to use some of the main features of *CLC Protein Workbench 2.0*.

The tutorials are also available as interactive Flash tutorials on <http://www.clcbio.com/tutorials>.

2.1 Tutorial: Starting up the program

This brief tutorial will take you through the most basic steps of working with *CLC Protein Workbench*. The tutorial introduces the user interface, demonstrates how to create a project, and demonstrates how to import your own existing data into the program.

When you open *CLC Protein Workbench* for the first time, the user interface looks like figure 2.1.

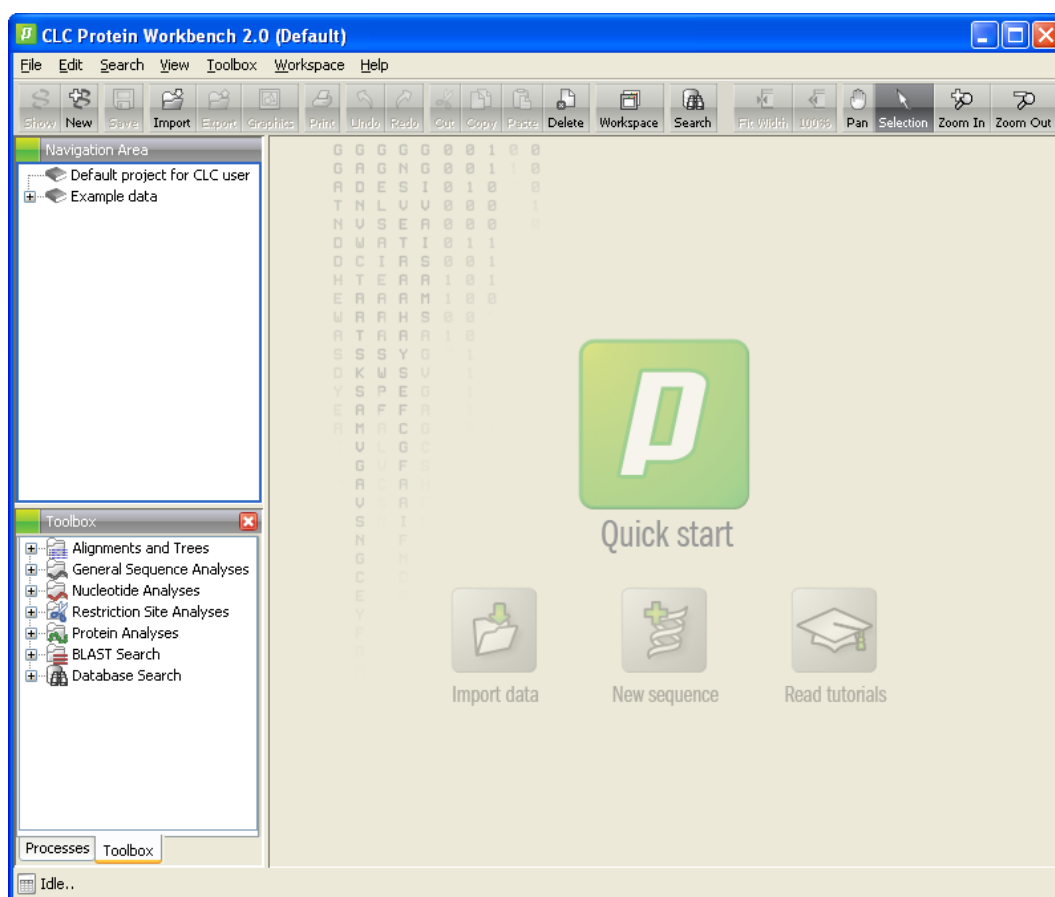


Figure 2.1: The user interface as it looks when you start the program for the first time. (Windows version of **CLC Protein Workbench**. The interface is similar for Mac and Linux.)

At this stage, the important issues are the **Navigation Area** and the **View Area**.

The **Navigation Area** to the left is where you keep all your data for use in the program. Most analyses of *CLC Protein Workbench* require that the data is saved in the **Navigation Area**. There are several ways to get data into the **Navigation Area**, and this tutorial describes how to import existing data.

The **View Area** is the main area to the right. This is where the data can be 'viewed'. In general, a **View** is a display of a piece of data, and the **View Area** can include several **Views**. The **Views** are represented by tabs, and can be organized e.g. by using 'drag and drop'.

2.1.1 Creating a project and a folder

When *CLC Protein Workbench* is started there is one default project in the **Navigation Area**. Create an additional project by:

File in the Menu Bar | New | Project ()
or **Ctrl + R** ($\text{⌘} + R$ on Mac)

Name the project 'Test' and press **Enter**.

The data in the project can be further organized into folders. Create a folder in the 'Test' project by:

Right-click the 'Test'-project in the Navigation Area | New | Folder ()
or **Ctrl + F** ($\text{⌘} + F$ on Mac)

Name the folder 'Subfolder' and press **Enter**.

2.1.2 Import data

Next, we want to import a sequence called HUMDINUC.fsa (FASTA format) from our own Desktop into the new 'Subfolder'. (This file is chosen for demonstration purposes only - you may have another file on your desktop, which you can use to follow this tutorial. You can import all kinds of files.)

In order to import the HUMDINUC.fsa file:

Import () **in the Toolbar | select FASTA (.fsa/.fasta) in the (Files of type) drop down menu | navigate to HUMDINUC.fsa on the desktop | Select**

For files of FASTA or PIR format, you are asked to state which type of sequence you are importing. (This will ensure that *CLC Protein Workbench* treats the sequence in the correct way.)

Click DNA/RNA | OK

The sequence is imported into the project or folder that was selected in the **Navigation Area**, before you clicked **Import**. Double-click the sequence in the **Navigation Area** to view it. The final result looks like figure 2.2.

2.1.3 Supported data formats

CLC Protein Workbench can import and export the following formats:

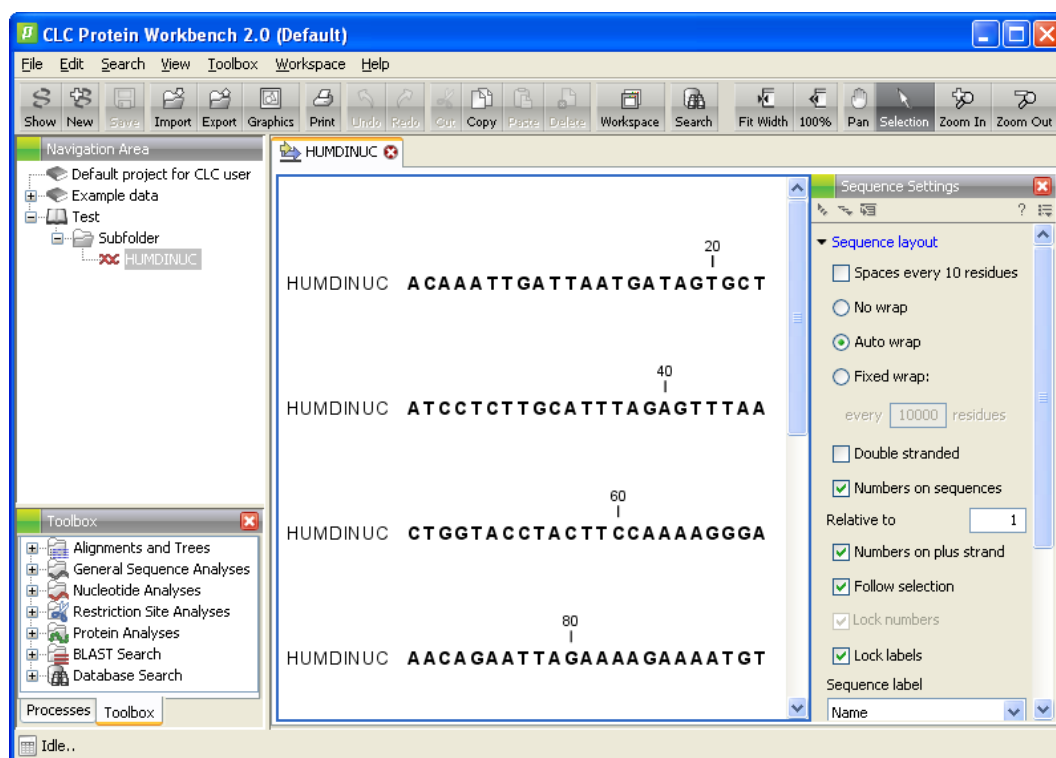


Figure 2.2: The HUMDINUC file is imported and opened.

File type	Suffix	File format used for
Phylip Alignment	.phy	alignments
GCG Alignment	.msf	alignments
Clustal Alignment	.aln	alignments
Newick	.nwk	trees
FASTA	.fsa/.fasta	sequences
GenBank	.gbk/.gb/.gp	sequences
GCG sequence	.gcg	sequences (only import)
PIR (NBRF)	.pir	sequences (only import)
Staden	.sdn	sequences (only import)
VectorNTI		sequences (only import)
DNAstrider	.str/.strider	sequences
Swiss-Prot	.swp	protein sequences
Lasergene sequence	.pro	protein sequence (only import)
Lasergene sequence	.seq	nucleotide sequence (only import)
Embl	.embl	nucleotide sequences
Nexus	.nxs/.nexus	sequences, trees, alignments, and sequence lists
CLC	.clc	sequences, trees, alignments, reports, etc.
Text	.txt	all data in a textual format
ABI		Trace files (only import)
AB1		Trace files (only import)
SCF2		Trace files (only import)
SCF3		Trace files (only import)
Phred		Trace files (only import)
mmCIF	.cif	structure (only import)
PDB	.pdb	structure (only import)
Preferences	.cpf	CLC workbench preferences

Notice that *CLC Protein Workbench* can import 'external' files, too. This means that *CLC Protein Workbench* can import all files and display them in the **Navigation Area**, while the above mentioned formats are the types which can be read by *CLC Protein Workbench*.

2.2 Tutorial: View sequence

This brief tutorial will take you through some different ways to display a sequence in the program. The tutorial introduces zooming on a sequence, dragging tabs, and opening selection in new view.

We will be working with DNA sequence 'AY738615'. Double-click the sequence in the **Navigation Area** to open it. The sequence is displayed with annotations above it. (To provide a better view of the sequence, hide the **Side Panel**. This is done by clicking the red X (X) at the top right corner of the **Side Panel** (in the right side of the **View Area**). (See figure 2.3).

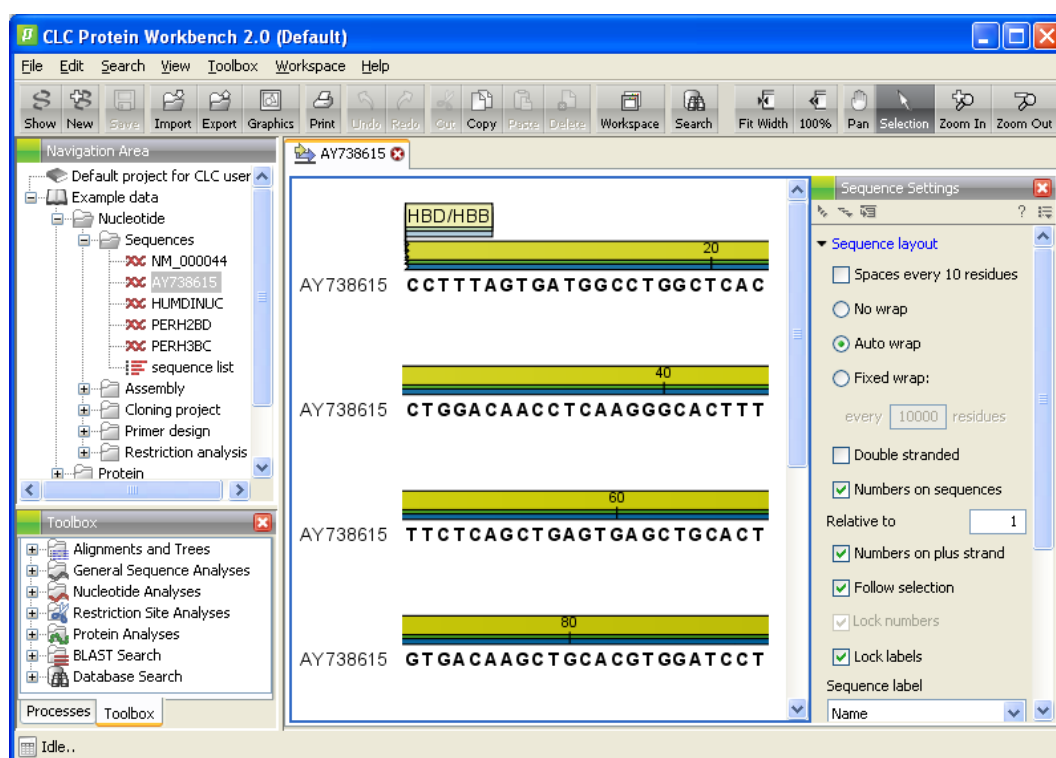


Figure 2.3: DNA sequence 'AY738615' opened in a view. The view preferences has been hidden to provide more space for the view.

As default, *CLC Protein Workbench* displays a sequence with annotations (colored arrows on the sequence) and zoomed to see the residues.

In this tutorial we want to have an overview of the whole sequence. Hence;

click Zoom Out (Z) in the Toolbar | click the sequence until you can see the whole sequence

In the following we will show how the same sequence can be displayed in two different views:

double-click sequence 'AY738615' in the Navigation Area

This opens an additional tab. Drag this tab to the bottom of the view. (See figure 2.4).

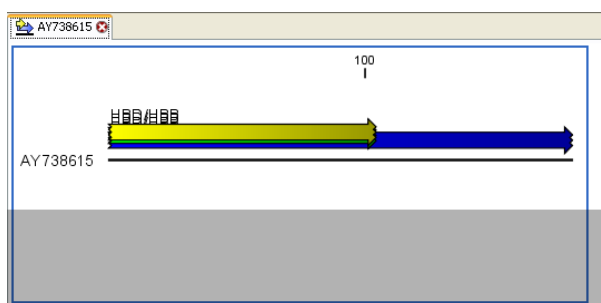


Figure 2.4: Dragging the tab down to the bottom of the view will display a gray area indicating that the tab can be "dropped" here and split the view.

The result is two views of the same sequence in the **View Area**, as can be seen in figure 2.5.

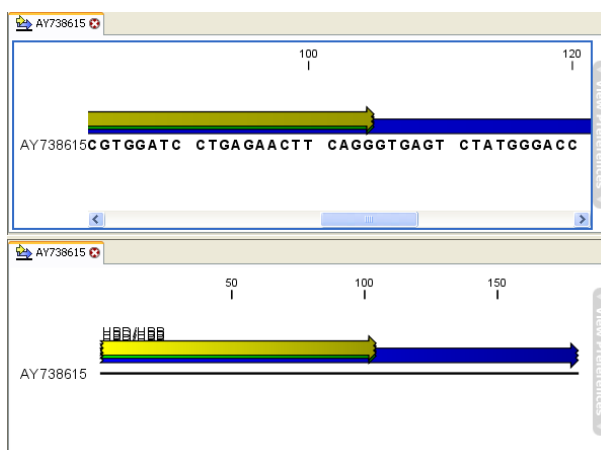


Figure 2.5: The resulting two views which are split horizontally.

If you want to display a part of the sequence, it is possible to select it, and open it in another view:

click Selection () in Toolbar | select a part of the sequence | right-click the selected part of the sequence in the top view | Open Selection in New View

This opens a third display of sequence 'AY738615'. However, only the part which was selected. In order to make room for displaying the selection of the sequence (the most recent view), drag the tab of the view down, next to the tab of the bottom view.

2.3 Tutorial: GenBank search and download

The *CLC Protein Workbench* allows you to search the NCBI GenBank database directly from the program, giving you the opportunity to both open, view, analyze and save the search results without using any other applications. To conduct a search in NCBI GenBank from *CLC Protein Workbench* you must be connected to the Internet.

This tutorial shows how to find a complete human hemoglobin DNA sequence in a situation where you do not know the accession number of the sequence.

To start the search:

Search | Search NCBI Entrez ()

This opens the search view. We are searching for a DNA sequence, hence:

Nucleotide

Now we are going to **Adjust Parameters** for the search. By clicking **More Choices** you activate an additional set of fields where you can enter search criteria. Each search criterion consists of a drop down menu and a text field. In the drop down menu you choose which part of the NCBI database to search, and in the text field you enter what to search for:

Click More Choices until three search criteria are available | choose Organism in the first drop down menu | write 'human' in the adjoining text field | choose All Fields in the second drop down menu | write 'hemoglobin' in the adjoining text field | choose All Fields in the third drop down menu | write 'complete' in the adjoining text field

The screenshot shows the NCBI search interface. At the top, there's a 'Choose database' section with 'Nucleotide' selected. Below this, there are three search criteria, each with a dropdown menu set to 'All Fields' and a text input field containing 'human', 'hemoglobin', and 'complete' respectively. There are buttons for 'Add search parameters', 'Start search', and 'Append wildcard (*) to search words'. At the bottom, there's a table of search results with columns for Accession, Definition, and Modification Date. The table shows several hits related to hemoglobin. At the very bottom, there are buttons for 'Download and Open', 'Download and Save', and a 'more...' link.

Accession	Definition	Modification Date
BC010230	Homo sapiens chromosome 10 open reading frame 83, mRNA (cDNA clone)	2004/03/25
BC015537	Homo sapiens hemoglobin, epsilon 1, mRNA (cDNA clone MGC:9582 IM...	2004/06/29
BC032122	Homo sapiens hemoglobin, alpha 2, mRNA (cDNA clone MGC:29691 IM...	2003/12/19
BC032264	Mus musculus hemoglobin, beta adult minor chain, mRNA (cDNA clone M...	2006/04/13
BC043020	Mus musculus hemoglobin alpha, adult chain 1, mRNA (cDNA clone MGC...	2004/06/30
BC050661	Homo sapiens hemoglobin, alpha 2, mRNA (cDNA clone MGC:60177 IM...	2003/10/07
BC051988	Mus musculus hemoglobin X, alpha-like embryonic chain in Hba complex...	2004/06/30
BC052008	Mus musculus hemoglobin Z, beta-like embryonic chain, mRNA (cDNA cl...	2006/04/27
BC056686	Homo sapiens hemoglobin, theta 1, mRNA (cDNA clone MGC:61857 IM...	2004/06/30
BC057014	Mus musculus hemoglobin Y, beta-like embryonic chain, transcript varia...	2005/12/09
BC069307	Homo sapiens hemoglobin, delta, mRNA (cDNA clone MGC:96894 IMAG...	2004/06/30

Figure 2.6: NCBI search view.

Now you have two choices: Either to click **Start search** (🔍) to commence the search in NCBI, or to click **Save search parameters** (💾) to choose where to save the search.

2.3.1 Saving the search

If you click 'Save search parameters', the program does not save the search results, but rather the search criteria. This allows you to perform exactly the same search later on.

In this tutorial, we are not certain of the quality of our search criteria, and therefore we choose not to save them. Consequently, click **Start search** (🔍) to perform the search.

2.3.2 Searching for matching objects

When the search is complete, the list of hits is shown. If the desired complete human hemoglobin DNA sequence is found, the sequence can be viewed by double-clicking it in the list of hits from the search. If the desired sequence is not shown, you can click the 'More' button below the list to see more hits.

2.3.3 Saving the sequence

The sequences which are found during the search can be displayed by double-clicking in the list of hits. However, this does not save the sequence. It is necessary to save the sequences before any analysis can be conducted. A sequence is saved like this:

click the tab with the name of the sequence | Save in the toolbar (💾)

or **click the tab with the name of the sequence | Ctrl + S (⌘ + S on Mac)**

When you close the view of the sequence, you are asked if you want to save the file.

If you do not want to view the sequence first, the sequence can be saved by dragging it from the list of hits into the **Navigation Area**.

2.4 Tutorial: Align protein sequences

It is possible to create multiple alignments of nucleotide and protein sequences. *CLC Protein Workbench* offers several opportunities to view alignments. The alignments can be used for building phylogenetic trees.

The sequences must be saved in the **Navigation Area** in order to be included in an alignment. To save a sequence which is displayed in the **View Area**, click the tab of the sequence and press Ctrl + S (or ⌘ + S on Mac). In this tutorial eight protein sequences from the Example data will be aligned. (See figure 2.7).

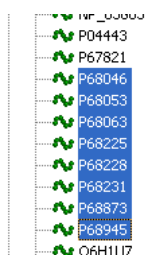


Figure 2.7: Eight protein sequences in a Protein project in the Navigation Area.

To begin aligning the protein sequences:

select the sequences | right-click either of the sequences | Toolbox | Alignments and Trees (🗑️) | Create Alignment (📄)

2.4.1 Alignment dialog

This opens the dialog shown in fig. 2.8.

It is possible to add and remove sequences from **Selected Elements** list. When the relevant proteins are selected there are two options: Click **Next** to adjust parameters for the alignment.

Clicking **Next** opens the dialog shown in fig. 2.9.

Leave the parameters at their default settings. An explanation of the parameters can be found in the program's **Help** function (❓) or in the user manual on <http://www.clcbio.com/download>.

Click **Finish** to start the alignment process which is shown in the **Toolbox** under the **Processes**

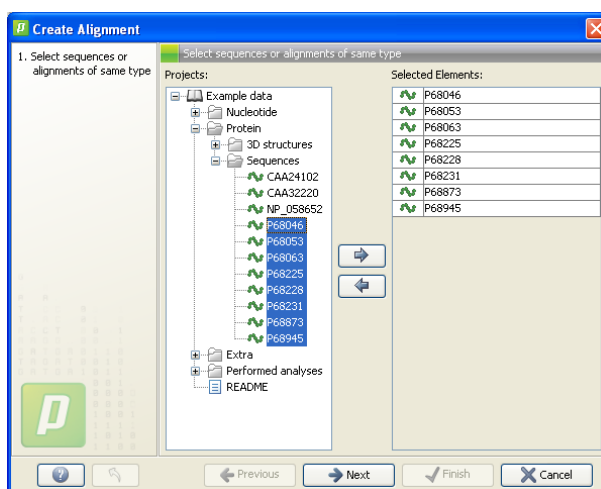


Figure 2.8: The alignment dialog displaying the 8 chosen protein sequences.

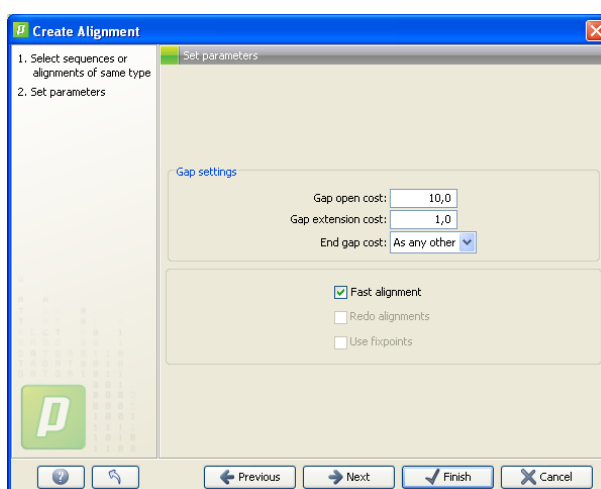


Figure 2.9: The alignment dialog displaying the available parameters which can be adjusted.

tab. When the program is finished calculating it displays the alignment (see fig. 2.10):

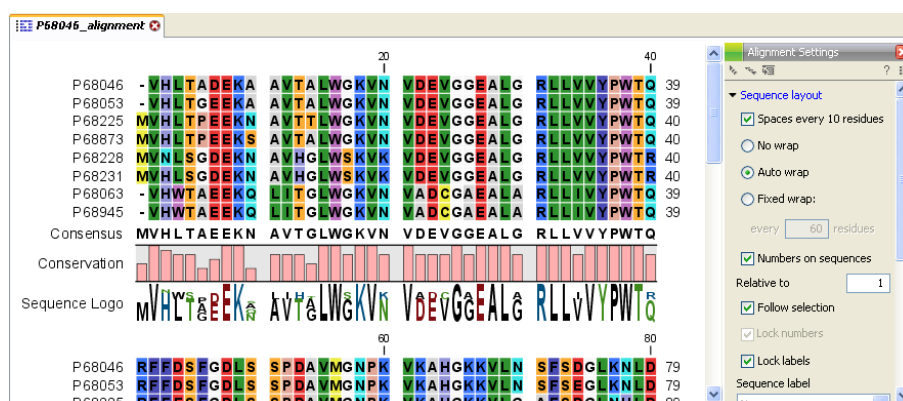


Figure 2.10: The resulting alignment.

Notice! The new alignment is not saved automatically. (The text on the tab is bold and italic to illustrate this.)

To save the alignment, drag the tab of the alignment view into the **Navigation Area**.

2.5 Tutorial: Create and modify a phylogenetic tree

You can make a phylogenetic tree from an existing alignment. (See how to create an alignment in 'Tutorial: Align protein sequence').

We use the 'P04443_alignment' located in Performed Analyses: Protein Workbench in the Example data. To create a phylogenetic tree:

right-click the 'P04443_alignment' in the Navigation Area | Toolbox | Alignments and Trees(🗄️) | Create Tree (🌳)

A dialog opens where you can confirm your selection of the alignment. Moving to the next step in the dialog you can choose between the neighbor joining and the UPGMA algorithms for making trees. You also have the option of including a bootstrap analysis of the result.

Click **Finish** to start the calculation, which can be seen in the **Toolbox** under the **Processes** tab, and after a short while a tree appears in the **View Area** (figure 2.11).

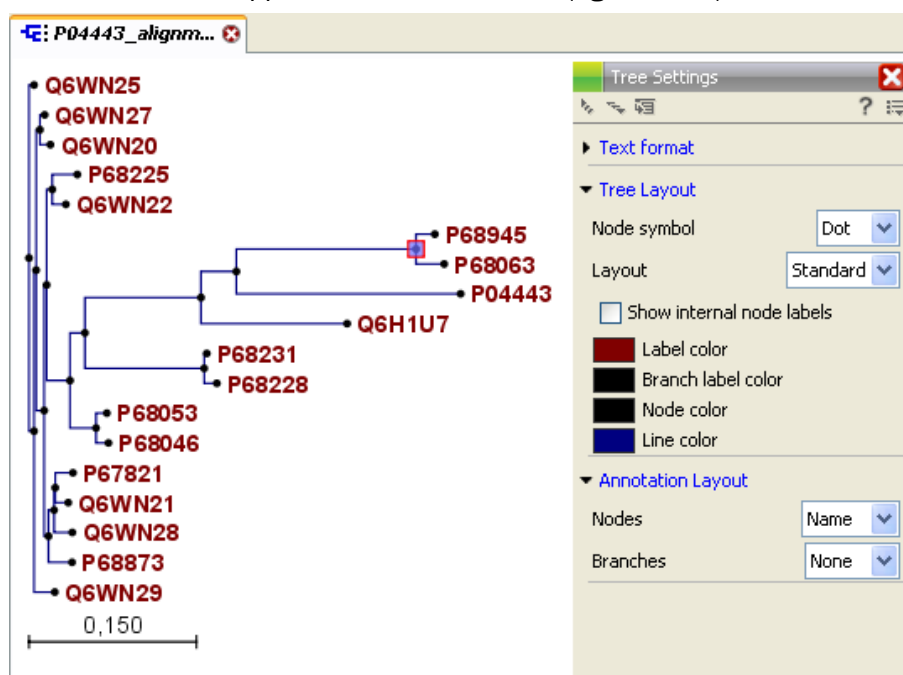


Figure 2.11: After choosing which algorithm should be used, the tree appears in the View Area. The Side panel in the right side of the view allows you to adjust the way the tree is displayed.

2.5.1 Tree layout

Using the **View preferences** (in the right side of the interface) of the tree view, you can edit the way the tree is displayed. Click **Tree Layout** and open the **Layout** drop down menu. Here you can choose between standard and topology layout. The topology layout can help to give an overview of the tree if some of the branches are very short.

When the sequences include the appropriate annotation, it is possible to choose between the accession number and the species names at the leaves of the tree. Sequences downloaded from GenBank, for example, have this information. The **Annotation Layout** preferences allows these different node annotations as well as different annotation on the branches.

The branch annotation includes the bootstrap value, if this was selected when the tree was

calculated. It is also possible to annotate the branches with their lengths.

2.6 Tutorial: Detect restriction sites

This tutorial will show you how to find restriction sites and annotate them on a sequence.

Suppose you are working with sequence PERH3BC from the example data, (can be downloaded from <http://www.clcbio.com/download>) and you wish to know which restriction enzymes will cut this sequence exactly once and create a 3' overhang. Do the following:

select the PERH3BC sequence from the Primer design folder | Toolbox in the Menu Bar | Restriction Site Analyses () | Restriction sites ()

The dialog shown in (fig. 2.12) opens, and you can confirm or change your selection of input sequence.

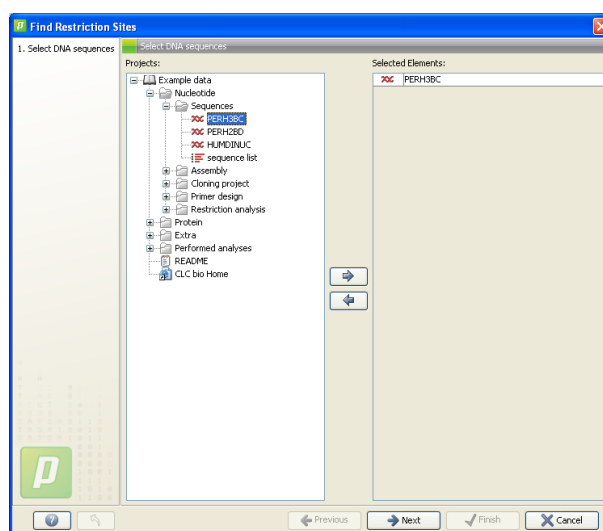


Figure 2.12: Choosing sequence PERH3BC.

In the next step you uncheck "Blunt ends" and "5' overhang" since we only wish to use enzymes with a "3' overhang". Then click **Select all** (see figure 2.13).

Click **Next** and choose both textual and graphical output. (See figure 2.14).

Click **Finish** to start the restriction site analysis.

2.6.1 View restriction site

The restriction sites are shown in two views: one view is in a textual format and the other view displays the sites as annotations on the sequence. To see both views at once:

View in the menu bar | Split Horizontally ()

The result is shown in figure 2.15.

Notice! The results are not automatically saved.

To save the result:

Right-click the tab | File | Save ()

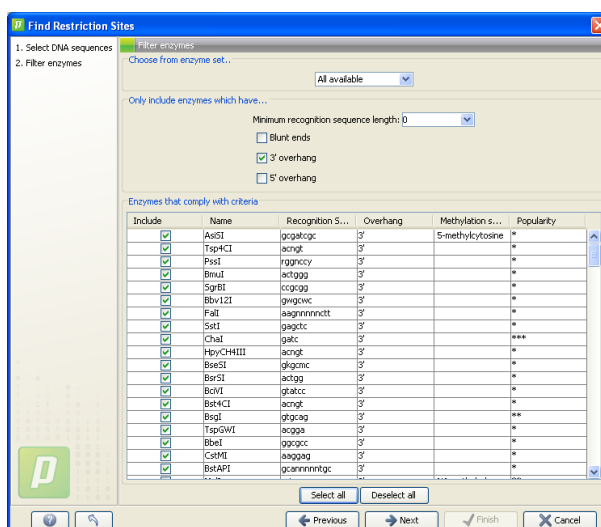


Figure 2.13: Setting parameters for restriction site detection.

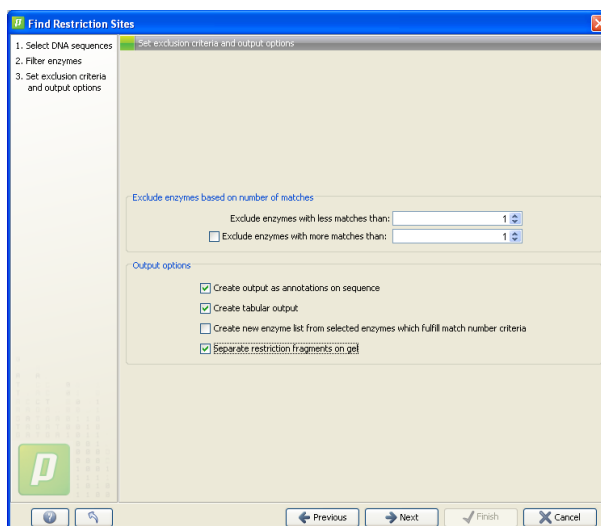


Figure 2.14: Selecting enzymes.

2.7 Tutorial: Sequence information

This tutorial shows you how to see background information about a sequence, including an overview of its annotations.

Suppose you are working with the HUMHBB sequence from the example data, (The Example data can be installed in the program by: clicking **Install Example Data** from the **Help** menu in the **Menu Bar**. The Example data can also be downloaded from <http://www.clcbio.com/download>.) and you wish to see more background information about this sequence. This can be done using the **Sequence Info** functionality of CLC Protein Workbench:

Select HUMHBB in the Navigation Area | Show (📄) in Menu Bar | Sequence Info (📄)

This opens a new view shown in figure 2.16.

The sequence is originally downloaded from GenBank, and it is the information from the GenBank file which is shown as a list of headings. Click the heading **Modification Date** to see when the

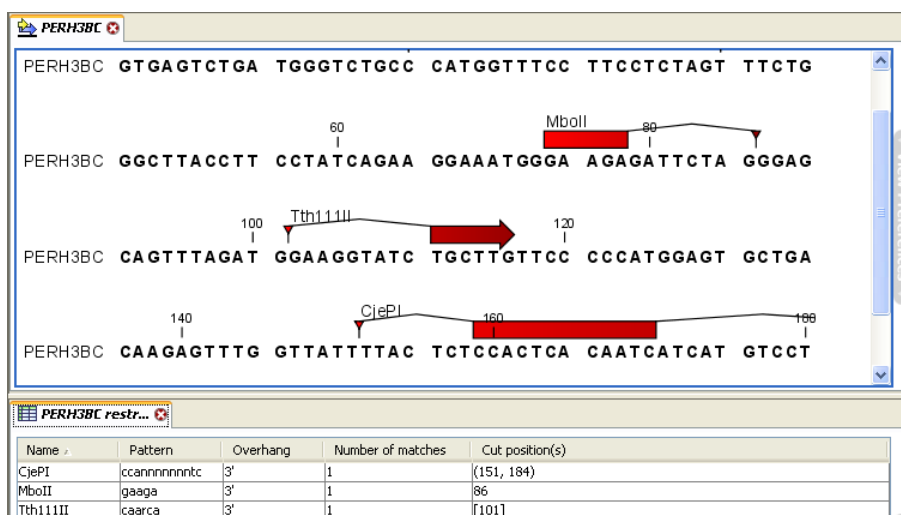


Figure 2.15: The result of the restriction site detection is displayed as text, and in this tutorial the View shares the View Area with a View of the PERH3BC sequence displaying the restriction sites (split-screen-view).

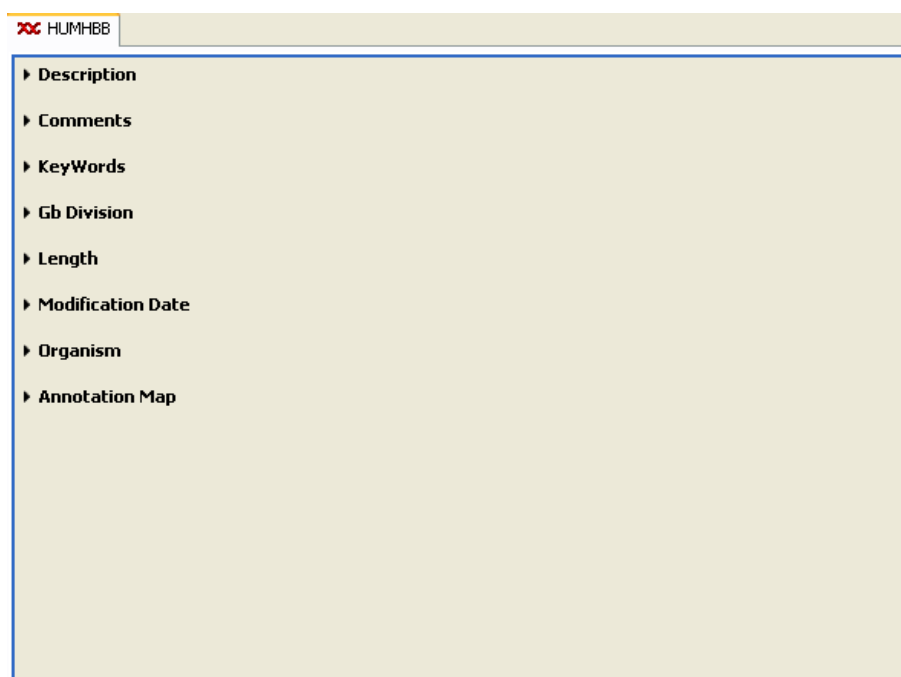


Figure 2.16: The initial view of sequence info of HUMHBB.

sequence was modified in GenBank.

At the bottom there is an **Annotation Map** providing an overview of the annotations on the sequence. The annotations are divided into types. We are interested in the coding sequences of HUMHBB:

Click Annotation Map | Click CDS

The seven coding sequences are displayed with the corresponding positions in GenBank syntax. In order to make full use of the **Annotation Map**, open a normal view of the HUMHBB sequence below the **Sequence Info**:

Select the HUMHBB in the Navigation Area | Drag it to the bottom of the View Area until a gray shadow appears

Now, clicking a coding sequences in the **Annotation Map** will make a selection representing the coding sequence in the view below . You can see that the selection matches the CDS annotation the yellow boxes in figure 2.17).

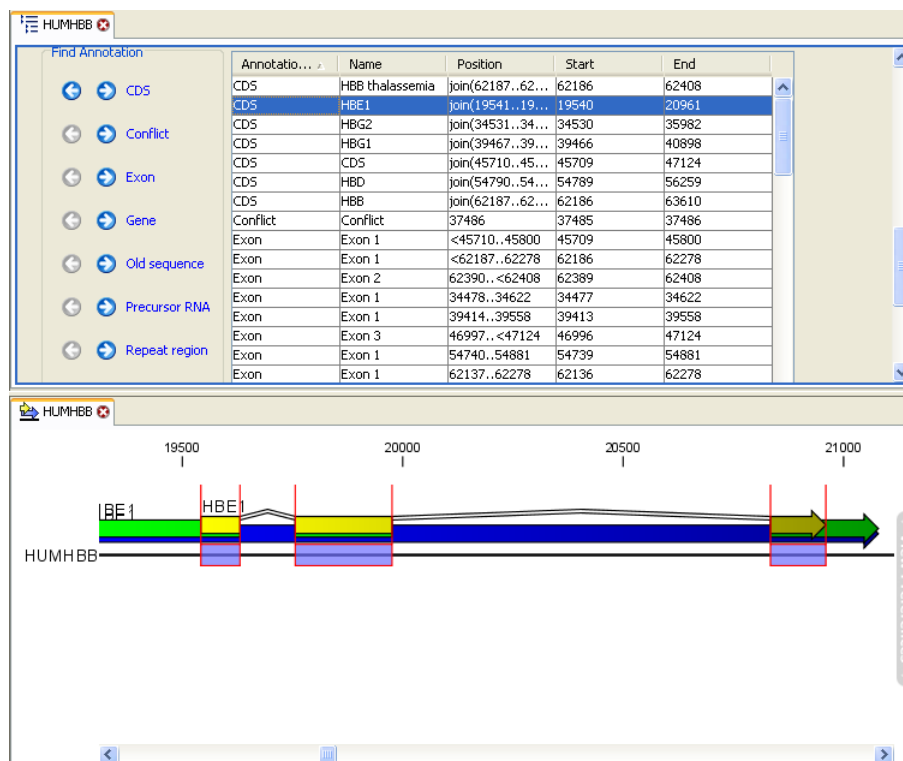


Figure 2.17: Two views of the HUMHBB sequence. The upper view shows the coding sequences (CDS), and the bottom view shows a selection corresponding to the CDS chosen in the upper view.

2.8 Tutorial: BLAST search

This tutorial shows you how to perform a BLAST search using *CLC Protein Workbench*.

Suppose you are working with the NP_058652 protein which constitutes the beta part of the hemoglobin molecule that is expressed in the adult house mouse, *Mus musculus*. To obtain more information about this molecule you wish to query the Swiss-Prot database to find homologous proteins in humans *Homo sapiens*, using the **Basic Local Alignment Search Tool** (BLAST) algorithm.

Please note that your computer must be connected to the Internet to complete this tutorial.

Start out by:

select protein NP_058652 in the Navigation Area | Toolbox | BLAST Search(📄) | BLAST Against NCBI Databases

In **Step 1** you can choose which sequence to use as query sequence. Since you have already chosen the sequence it is displayed in the **Selected Elements** list.

Click **Next**.

In **Step 2** (figure 2.18), choose the default BLAST program: **BLASTp: Protein sequence against Protein database** and select the **Swiss-Prot** database in the **Database** drop down menu.

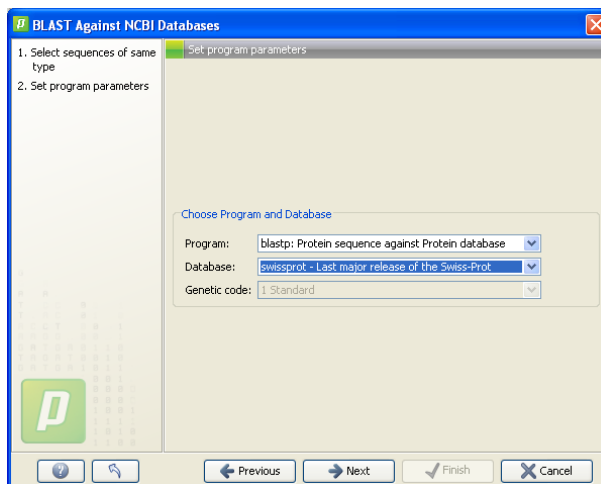


Figure 2.18: Choosing BLAST program and database.

Click **Next**.

In the **Limit by Entrez query** in **Step 3**, choose **Homo sapiens[ORGN]** from the drop down menu to arrive at the search configuration seen in figure 2.19. Including this term limits the query to proteins of human origin.

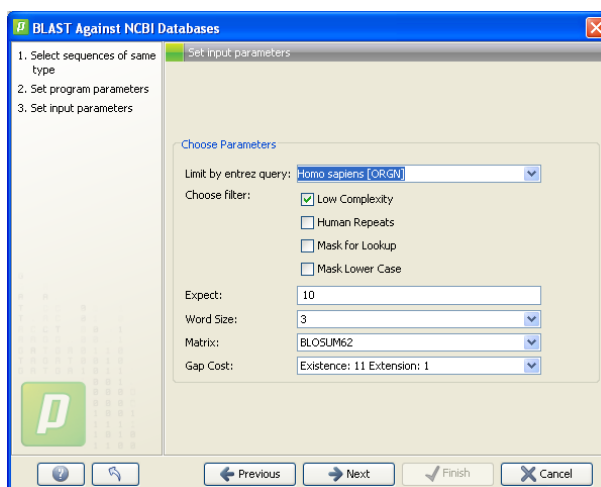


Figure 2.19: The BLAST search is limited to homo sapiens[ORGN]. The remaining parameters are left as default.

Click **Finish** to accept the default 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 is received and opened in a new view.

The output is shown in figure 2.20 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.

Try placing your mouse pointer 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.

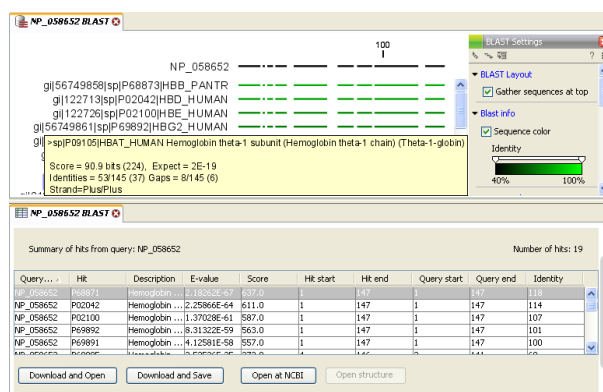


Figure 2.20: Output of a BLAST search. By holding the mouse pointer over the lines you can get information about the sequence.

For now, we will focus our attention on sequence P02042 - the BLAST hit that is second from the top of the list. To open sequence P02042:

right-click the line representing sequence P02042 | Open Sequence in New View

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 part of your project 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 sequences.

2.9 Tutorial: Proteolytic cleavage detection

This tutorial shows you how to find cut sites and see an overview of fragments when cleaving proteins with proteolytic cleavage enzymes.

Suppose you are working with protein CAA32220 from the example data, and you wish to see where enzyme **trypsin** will cleave the protein. Furthermore, you want to see details for the resulting fragments which are between 10 and 15 amino acids long.

right-click protein CAA32220 in the Navigation Area | Toolbox | Protein Analyses
 **| Proteolytic Cleavage**

This opens **Step 1** of the Proteolytic Cleavage dialog. In this step you can choose which sequences to include in the analysis. Since you have already chosen protein CAA32220, click **Next**.

In this step you should select **Trypsin**. This is illustrated in figure 2.21.

Click **Next** to go to **Step 3** of the dialog.

In **Step 3** you can adjust the parameters for which fragments of the cleavage you want to include in the table output of the analysis.

Type '10' in the Exclude fragments shorter than | Check the box: Exclude fragments longer than | enter '15' in the corresponding text field

These parameter adjustments are shown in figure 2.22:

Click **Finish** to make the analysis. The result of the analysis can be seen in figure 2.23

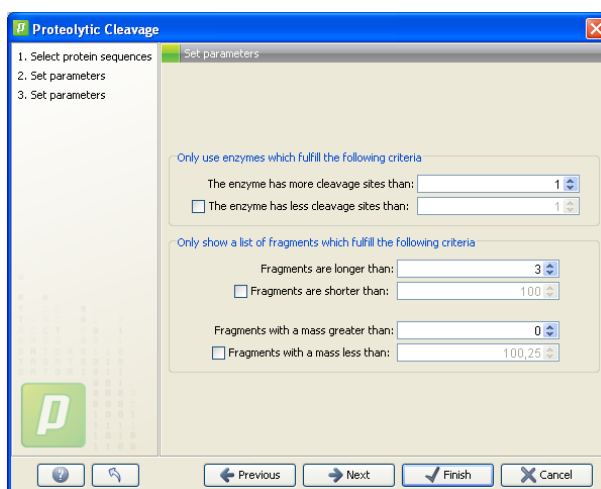


Figure 2.21: Selecting trypsin as the cleaving enzyme.

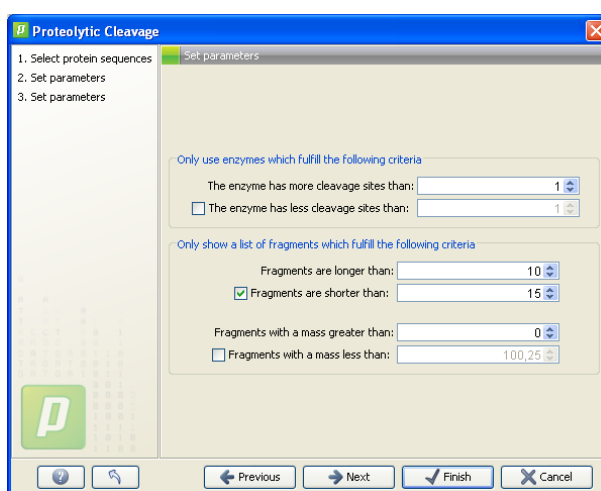


Figure 2.22: Adjusting the output from the cleavage to include fragments which are between 10 and 15 amino acids long.

Notice! The output of proteolytic cleavage is two related views. The sequence view displays annotations where the sequence is cleaved. The table view shows information about the fragments satisfying the parameters set in the dialog. Subsequently, if you have restricted the fragment parameters, you might have more annotations on the sequence than fragments in the table.

If you conduct another proteolytic cleavage on the same sequence, the output consists of: (possibly) new annotations on the original sequence and an additional table view, listing all fragments.

2.10 Tips and tricks for the experienced user

In this tutorial you will get to know a number of ways to cut corners when using *CLC Protein Workbench*. The following sections will show you how to get your tasks done quickly and easily. When you are using the program it is hard to discover these shortcuts yourself which is the reason why this tutorial was written.

The tutorial assumes that you have used the program for a while, since the basic usages are not

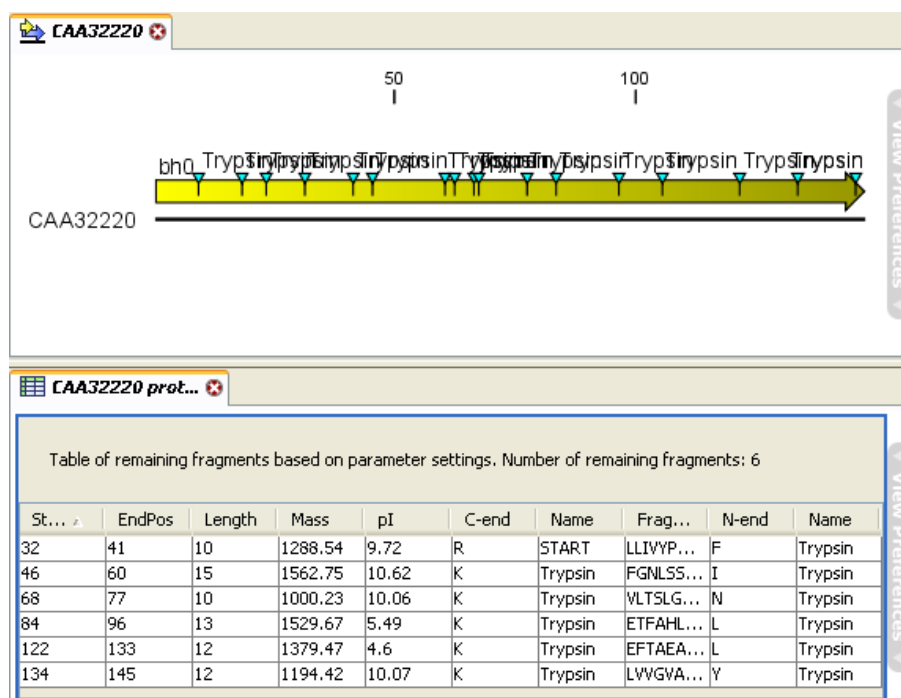


Figure 2.23: The output of the proteolytic cleavage shows the cleavage sites as annotations in the protein sequence. The accompanying table lists all the fragments which are between 10 and 15 amino acids long.

explained.

2.10.1 Open and arrange views using drag and drop

Instead of opening views using double click or **Show**, you can use drag and drop both to open and arrange views. Drag and drop is supported both within the **Navigation Area**, within the **View Area** and between the two areas:

- 1. Drag and drop an element within the Navigation Area:** Moves the element to the drop location.
- 2. Drag an element from the Navigation Area to the View Area:** Opens the element in a new view. The view will be opened in the part of the **View Area** where the element is dropped.
- 3. Drag the tab of a view within the View Area:** If there are other views open, this will split the **View Area** and make it possible to see several views at the time.
- 4. Drag the tab of a view into the Navigation Area:** If the view is new and has not been saved to a project before, this will save the view at the drop location. If the view is already represented in the **Navigation Area**, this will save a copy of the view at the drop location.

2.10.2 Find element in the Navigation Area

If you have a view of e.g. a sequence and you wish to know in which project this sequence is saved, use the **Find in Project** function:

right-click the tab of the view | View | Find in Project()

This will select the sequence in the **Navigation Area** (see figure 2.24).

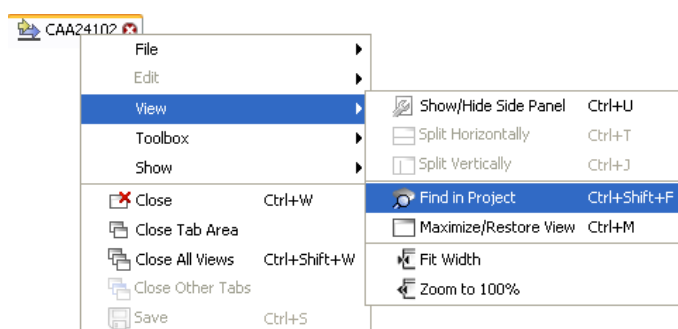


Figure 2.24: This will select the sequence in the Navigation Area.

You can also use the shortcut key: Ctrl + Shift + F on Windows or ⌘ + Shift + F on Mac.

2.10.3 Find specific annotations on a sequence

If you are looking for a specific annotation on a sequence, you may benefit from viewing the **Sequence info** while keeping an ordinary view of the sequence on the screen. In the **Sequence info** you find an Annotation map which displays all the annotations of the sequence. The annotations serve as links, selecting the annotation in the ordinary view of the sequence (see figure 2.25).

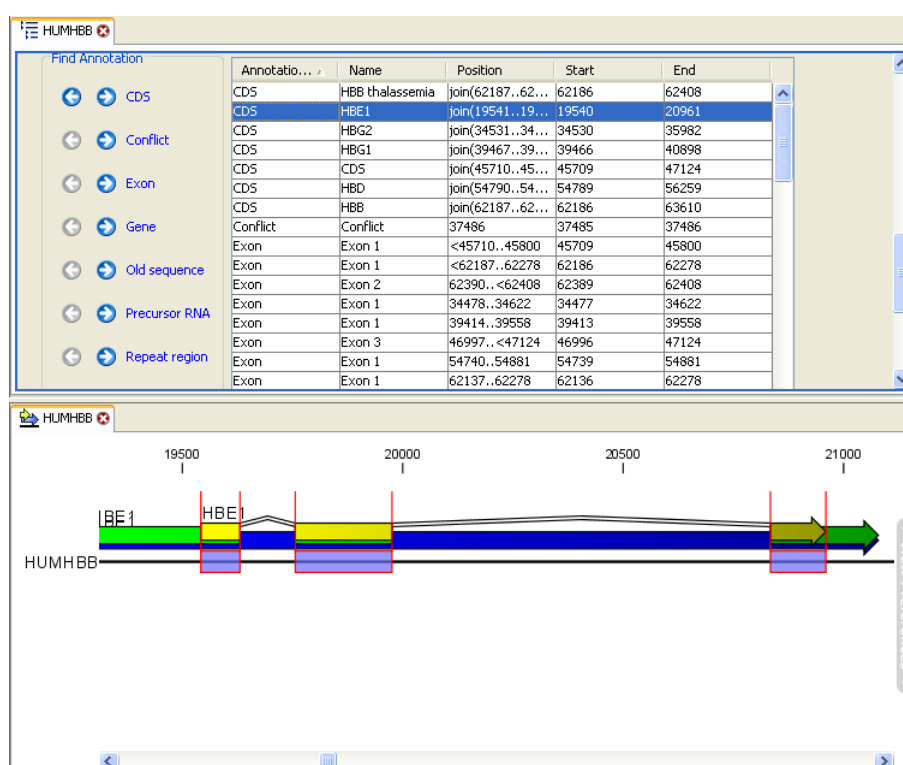


Figure 2.25: Clicking the HBE1 coding region in the top view selects the annotation on the sequence in the bottom view.

For sequences with many annotations, it is easier to navigate using these links compared to of scrolling in the ordinary view of the sequence.

2.10.4 Split sequences into several lines

Producing graphics of long sequences can be a strenuous task, especially if you have not discovered the "Wrap sequence" option. If you just export graphics of a long sequence without wrapping, you will get an extremely wide graphics file which probably has been edited in a graphics program before use. Wrapping the sequence allows you to control the width and height of the graphics file (see figure 2.26).

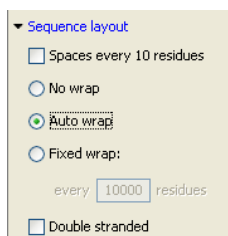


Figure 2.26: *Wrapping the sequence automatically.*

2.10.5 Make a new sequence of a coding region

If you have a genomic sequence containing a coding region, you can easily make a new sequence which only consists of the coding region (see figure 2.27):

right-click the coding region's annotation | Open Annotation in New View

This will open a new sequence which only consists of the residues covered by the annotation.

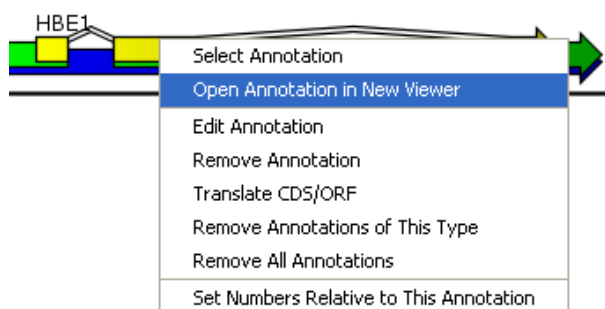


Figure 2.27: *Opening the coding region in a new view.*

2.10.6 Translate a coding region

If you have a genomic sequence containing one or more coding regions, you can translate these regions in a quick and easy way. If you want to translate a single coding region (see figure 2.28):

right-click the coding region's annotation | Translate CDS/ORF

This will open a new view with the translated sequence.

In order to translate all the coding regions of a sequence:

Toolbox | Nucleotide Analyses (📁) | Translate to protein (🧬) | Translate CDS and ORF in Step 2

This will extract all the coding regions of the sequence and for each region it will open a new view with the translation.

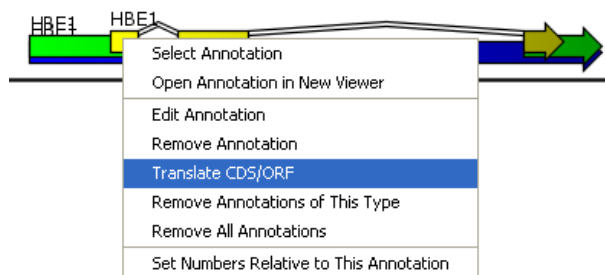


Figure 2.28: Opening a new view with the translation of the coding region.

2.10.7 Copy annotations from one sequence to another

If you have a collection of similar sequences, and you have annotated one of the sequences, you can copy these annotations to the rest of the sequences. First, create an alignment of the sequences. Next, find the annotated sequence and for each of the annotations that you want to copy:

right-click the annotation | Copy Annotation to other Sequences

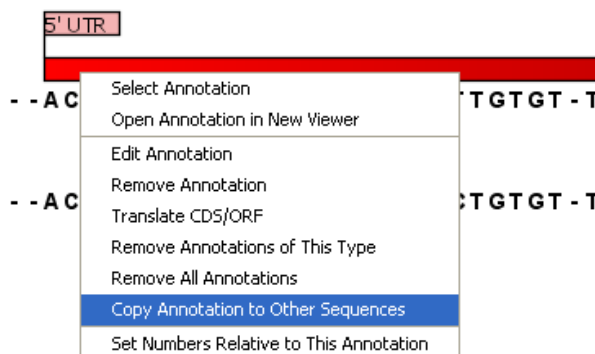


Figure 2.29: Copying annotation to other sequences in the alignment.

A dialog listing all the sequences in the alignment is shown. The annotation will be copied to the sequences that you select in this dialog. If the sequences are not identical, the annotation will still be copied.

2.10.8 Get overview and detail of a sequence at the same time

If you have a large sequence and you want to be able to get an overview of the whole and still keep the details of the residues, you can use the **Split views** functionality. In the example below (figure 2.30), the end of the red annotation is examined in detail in the bottom view, and in the upper view you have the overview of the whole alignment.

In this example, a selection was made in the upper view, and the bottom view automatically scrolls to display this selection (this behavior can be turned off by unchecking the "Follow selection" option in the **Side Panel**).

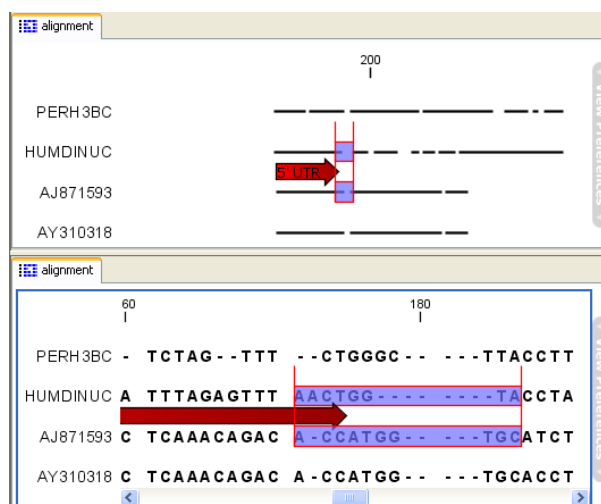


Figure 2.30: Using the split views and follow selection functionalities.

2.10.9 Smart selecting in sequences and alignments

There are a number of ways to select residues in sequences and alignments:

Using the mouse. This is the most basic way of selecting. Place the mouse cursor where you want the selection to start, press and hold the mouse button, move the mouse to the location where the selection should end and release the mouse button.

Using the mouse in combination with the Shift key. If you have made a selection and want to extend or reduce the selection, hold the Shift key while clicking the location where you want the boundary of the selection.

Using the arrow keys in combination with the Shift key. If you have made a selection and want to extend or reduce the selection, hold the Shift key while pressing the left and right arrow keys.

Using the mouse in combination with the Ctrl (for Windows) or ⌘ (for Mac) key. By holding this key, you can make multiple selections that are not contiguous.

Selecting an annotation. Double-click an annotation in order to select the residues that the annotation covers. This is especially helpful if the annotation is not contiguous (as the CDS region in figure 2.27).

Using the Search function. At the bottom of **Side Panel** to the right, there is a search field, which can be used for selections (use Ctrl + F on Windows or ⌘ + F on Mac). You can both search for annotations, residues or positions. The result of the search is a selection (as shown in figure 2.31). Remember to separate the start and end numbers with two punctuation marks (..).

No matter how you make your selection, you can see the start and end positions in right part of the status bar below the **View Area**.

2.10.10 Check for updates and additional information about sequences

If you have downloaded a sequence from NCBI or UniProt, you can easily check if the online information about the sequence has been updated and get additional information about the

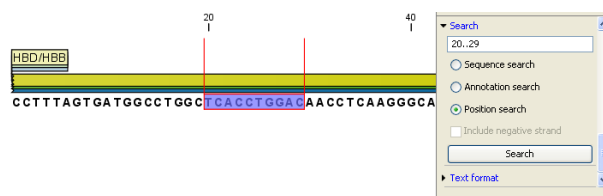


Figure 2.31: Making a selection from position 20 to 29 (both included) using the Search function.

sequence:

right-click the sequence | Web info | NCBI or UniProt

This will open your default web-browser showing the information about the sequence at either NCBI or UniProt. Clicking "PubMed" instead of "NCBI/UniProt" gives you a direct link to the sequence's PubMed references.

2.10.11 Quickly import sequences using copy-paste

Instead of using the **Import** (📁) function to import a sequence, you can use copy-paste. If you have copied the sequence from a source outside the program (e.g. a webpage or text document), you can paste it into the text field in the **Create new sequence** dialog (shown in figure 2.32).

Figure 2.32: Pasting a sequence into the text field at the bottom is a quick way of importing sequence data.

This dialog lets you paste all kinds of characters into the text field, including numbers and spaces. If you have pasted e.g. numbers into the field, just press and hold the space key on your keyboard until the numbers have been deleted. Spaces are not included in the new sequence.

2.10.12 Perform analyses on many elements

If you have a folder with a lot of mixed elements (e.g. both nucleotide and protein sequences, alignments, reports), you can often select the whole folder for an analysis, even if the analysis should only be performed on a special type of element (e.g. nucleotide sequences). In the example below (figure 2.33), the dialog says "Select nucleotide sequences", but the project contains both protein and nucleotide sequences. Instead of carefully pinpointing the nucleotide

sequences, you can just press Ctrl+A (⌘ +A on Mac), selecting all the visible elements. When you add these elements (➡), the protein sequences are filtered out.

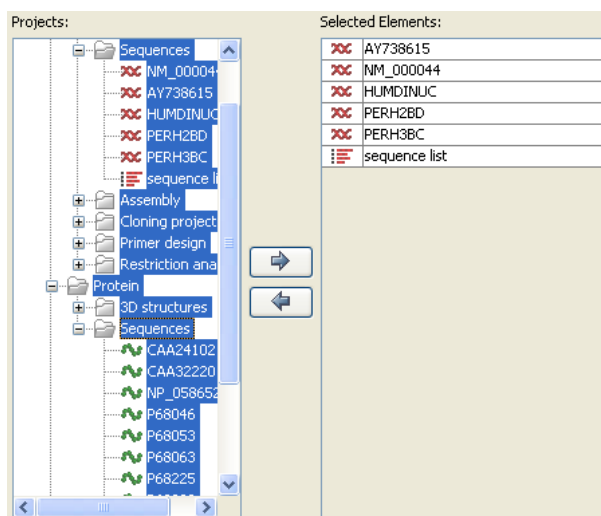


Figure 2.33: Selecting protein and dna sequences, but the dialog automatically filters out the protein sequences.

2.10.13 Drag elements to the Toolbox

If you have selected e.g. some protein sequences in the **Navigation Area** that you wish to use for creating an alignment,

2.10.14 Export elements while preserving history

If you have created e.g. an alignment and wish to export it to a colleague with the detailed history of all the source sequences, you can select the alignment and all the sequences for export. There is, however, a much easier way to do this (see figure 2.34):

Select the alignment | File | Export with dependent elements

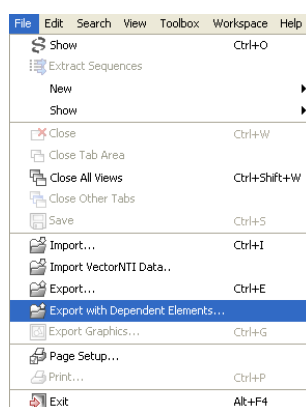


Figure 2.34: Export with dependent elements in order to preserve the detailed history of an element.

This will export the alignment including all the source sequences in one clc-file. When your colleague import the alignment, its detailed history is preserved.

2.10.15 Avoid the mouse trap - use keyboard shortcuts

Many tasks can be performed without using the mouse. When you do the same task again and again, you can save some time by learning its shortcut key. As an example you can navigate and zoom a view of sequence or an alignment using the keyboard:

- **Navigate the view using the four arrow keys.** This is equivalent to scrolling with the mouse using the scroll bars.
- **Use the '+' and '-' keys to zoom in and out.** This is equivalent to using the zoom modes in the toolbar.

Note that you have to click once inside the view with the mouse first in order to use this functionality.

There are many other shortcuts in *CLC Protein Workbench* which may save you a lot of time when performing repetitive tasks. See section 3.6 for a list of available shortcuts.

Part II

Basic Program Functionalities

Chapter 3

User Interface

Contents

3.1 Navigation Area	52
3.1.1 Data structure	52
3.1.2 Create new projects and folders	53
3.1.3 Multiselecting elements	54
3.1.4 Moving and copying elements	54
3.1.5 Change element names	55
3.1.6 Delete elements	56
3.1.7 Show folder elements in View	57
3.1.8 Sequence properties	58
3.2 View Area	58
3.2.1 Open View	58
3.2.2 Close Views	59
3.2.3 Save changes in a View	60
3.2.4 Undo/Redo	60
3.2.5 Arrange Views in View Area	61
3.2.6 Side Panel	62
3.3 Zoom and selection in View Area	63
3.3.1 Zoom In	63
3.3.2 Zoom Out	65
3.3.3 Fit Width	65
3.3.4 Zoom to 100%	65
3.3.5 Move	65
3.3.6 Selection	65
3.4 Toolbox and Status Bar	66
3.4.1 Processes	66
3.4.2 Toolbox	66
3.4.3 Status Bar	67
3.5 Workspace	67
3.5.1 Create Workspace	67
3.5.2 Select Workspace	67

3.5.3 Delete Workspace	67
3.6 List of shortcuts	68

This chapter provides an overview of the different areas in the user interface of *CLC Protein Workbench 2.0*. As can be seen from figure 3.1 this includes a **Navigation Area**, **View Area**, **Menu Bar**, **Toolbar**, **Status Bar** and **Toolbox**.

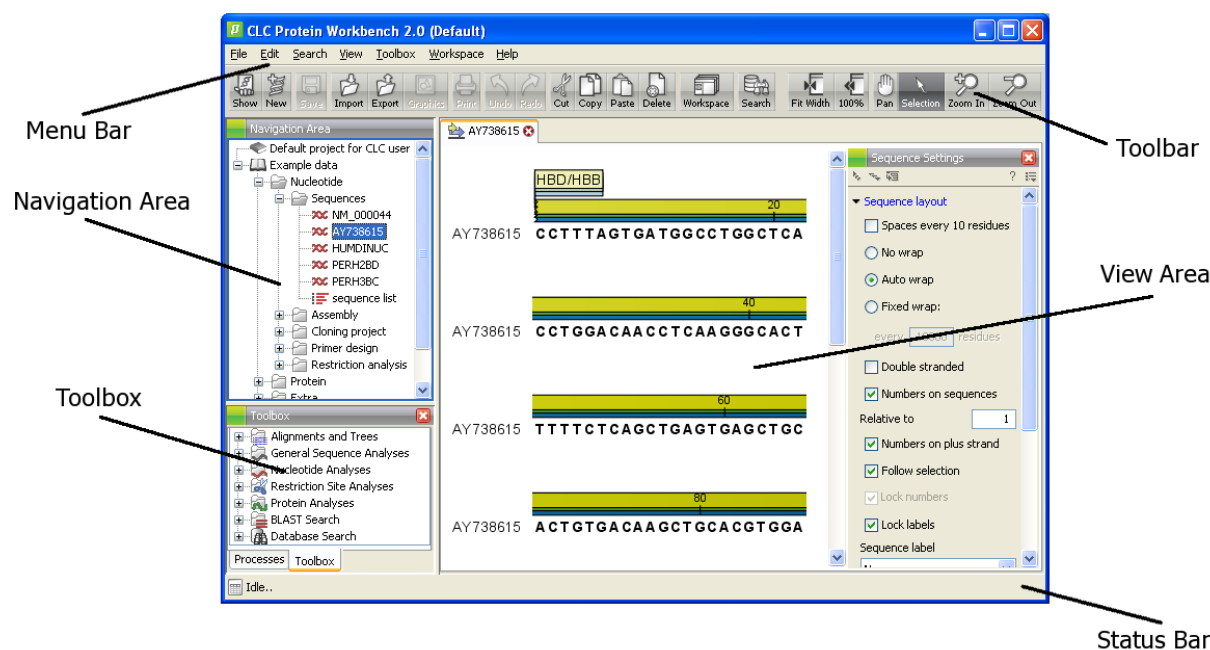


Figure 3.1: The user interface consists of the Menu Bar, Toolbar, Status Bar, Navigation Area, Toolbox, and View Area.

3.1 Navigation Area

The **Navigation Area** is located in the left side of the workbench, under the **Toolbar**. It is used for organizing and navigating data. The **Navigation Area** displays a **Project Tree** (see figure 3.2), which is similar to the way files and folders are usually displayed on your computer. The **Project Tree** contains one or more projects. The elements which are available in the **Navigation Area** remain the same when changing **Workspaces** (see section 3.5).

A project can be a collection of elements which are related, e.g. because the elements are used in the same assignment or research project.

The word 'Element' is used to refer to sequences, saved searches, lists, folders etc. In other words, everything which can be stored in a project in the **Navigation Area**.

3.1.1 Data structure

Elements, or data, in *CLC Protein Workbench 2.0* are stored in a kind of database. Hence, the data cannot be browsed from e.g. Windows Explorer or similar file systems. However, elements are available from the **Navigation Area**. To open an element:

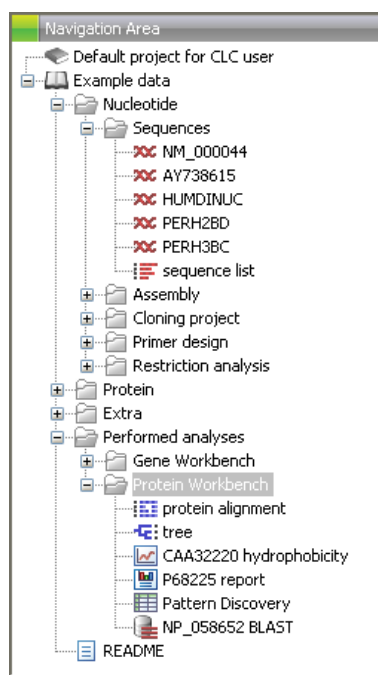


Figure 3.2: The Navigation Area.

Double-click the element

or **Click the element | Show () in the Toolbar | Select the desired way to view the element**

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

Adding data

Data can be added to a project in a number of ways. Files can be imported from the file system, and elements from the **Navigation Area** can also be exported to the file system. (For more about import and export, see chapter 6.)

Furthermore, an element can be added to a project by dragging it into the **Navigation Area**. Elements on lists, e.g. search hits or sequence lists, can also be dragged to the **Navigation Area**.

When dragging from the **View Area** to the **Navigation Area**, the element, e.g. a sequence, an alignment, or a search report, is selected by clicking on the tab and dragging it into the navigation area. If the element already exists, you are asked whether you want to save a copy.

If a piece of data is dropped on a folder or a project, the data is placed at the bottom of the list of elements in the folder or project in question.

If a piece of data is dropped on an element, which is not a folder or a project, the data is added just after that element.

3.1.2 Create new projects and folders

In the **Navigation Area** all files and folders are stored in one or more projects. Creating a new project can be done in two ways:

right-click an element in the Navigation Area | New | New Project (📁)

or **File | New | New Project (📁)**

Regardless of which element is selected when you create a new project, the new project is placed at the bottom of the **Project Tree**.

You can move the project manually by selecting it and dragging it to the desired location. Projects are always placed at the upper-most level in the **Project Tree**.

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 | New Folder (📁)

or **File | New | New Folder (📁)**

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

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

3.1.3 Multiselecting elements

Multiselecting elements in the **Navigation Area** can be done in the following ways:

- Holding down the <Ctrl> key 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 cursor with the arrow-keys while holding down the <Shift> key, enables you to increase the number of elements selected.

3.1.4 Moving and copying elements

Elements can be moved and copied in two ways: using the copy, cut and paste functions, or using drag and drop.

Copy, cut and paste elements

Copies of elements, folders, and projects 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 disappears until you activate the paste function.

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 projects and folders in the **Project Tree**
- 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. Further description of the function is found in connection with the relevant functions.

3.1.5 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).
- Species.

- Species (accession).
- Common Species.
- Common Species (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 project, folder, piece of data etc. can be done in three different ways:

right-click the element | Rename

or **select the element | Edit in the Menu Bar | Rename**

or **select the element | F2**

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.

3.1.6 Delete elements

Deleting a project, folder, piece of data, etc. 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 a **Recycle Bin** where it is kept as a precaution.

Restore Deleted Elements

The elements in the **Recycle Bin** can be restored and saved in the **Navigation Area** again. This is done by:

Edit in the Menu Bar | Restore Deleted Elements ()

This opens the dialog shown in fig. 3.3.

The dialog shows a list of all the deleted elements. Select the elements you want to restore and click next. This opens the dialog shown in fig. 3.4.

Choose where to restore the deleted elements. Click **Finish**

Notice! Only files which were saved in the **Navigation Area**, and then deleted, can be restored.

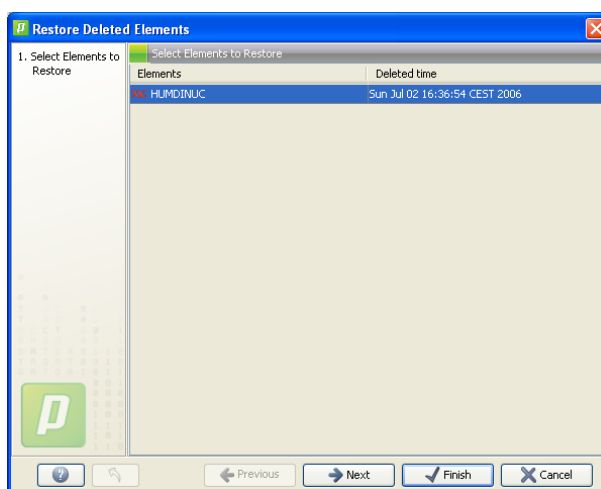


Figure 3.3: The Restore Deleted Elements dialog.

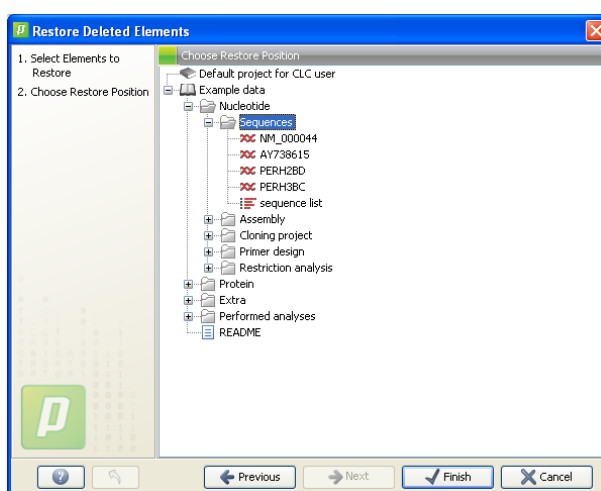


Figure 3.4: The Restore Deleted Elements dialog.

The deleted elements remain in the **Recycle Bin** until the **Recycle Bin** is emptied. To empty the bin:

Edit in the Menu Bar | **Empty recycle bin** (🗑️)

3.1.7 Show folder elements in View

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

select a project | **Show** (📁) in the **Toolbar** | **Folder Contents** (📁)

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 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**.

Notice! The **View** only displays one layer of the **Project Tree** at a time.

3.1.8 Sequence properties

Sequences downloaded from databases have a number of properties, which can be displayed using the **Sequence Properties** function:

Right-click a sequence in the Navigation Area | Properties

This will show a dialog as shown in figure 3.5.



Figure 3.5: Sequence properties for the HUMDINUC sequence.

For a more comprehensive view of sequence information, see section 11.2.

3.2 View Area

The **View Area** is the right-hand part of the workbench interface, 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 3.6.

Notice I.e., the tab concept is central to working with *CLC Protein Workbench 2.0*, 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 3.3 deals with the zooming and selecting functions.

3.2.1 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.

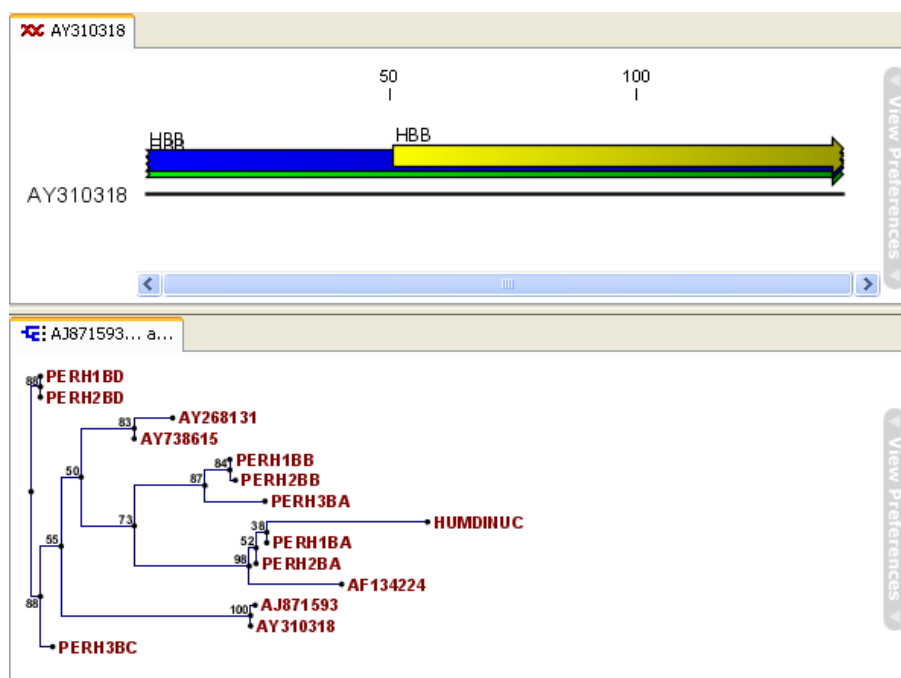


Figure 3.6: A View Area can enclose several Views, each View is indicated with a tab (see top left View, which shows protein P12675). Furthermore, several Views can be shown at the same time (in this example, three views are displayed).

Notice! 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 3.1.4 for instructions on how to open a **View** using drag and drop.

3.2.2 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. See figure 3.7

- **Close.** See above.
- **Close Tab Area.** Closes all tabs in the tab area.
- **Close All Views.** Closes all tabs, in all tab areas. Leaves an empty workspace.
- **Close Other Tabs.** Closes all other tabs in the particular tab area.

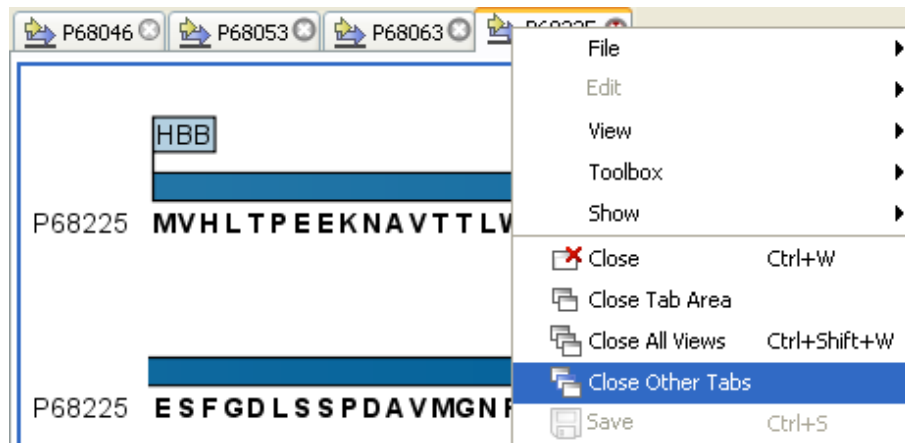


Figure 3.7: By right-clicking a tab, several close options are available.

3.2.3 Save changes in a View

When changes are made in a view, the text on the tab appears *bold and italic*. 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 **View** containing an element that has been changed since you opened it, you are asked if you want to save.

When saving 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 3.8).

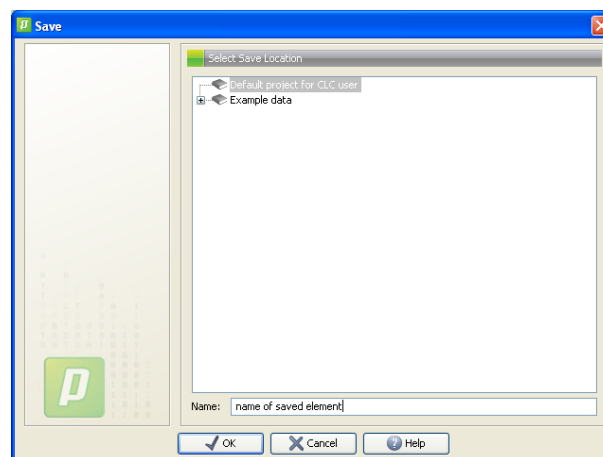


Figure 3.8: Save dialog.

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

After naming the element, press **OK**

3.2.4 Undo/Redo

If you make a change 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**

Notice! Actions in the **Navigation Area**, e.g. renaming and moving elements, cannot be undone. However, you can restore deleted elements (see section 3.1.6).

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

3.2.5 Arrange Views in View 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. 3.9) illustrating that the view will be placed in this part of the **View Area**.

The results of this action is illustrated in figure 3.10.

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 3.11). 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.

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 size of View (□)**

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

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

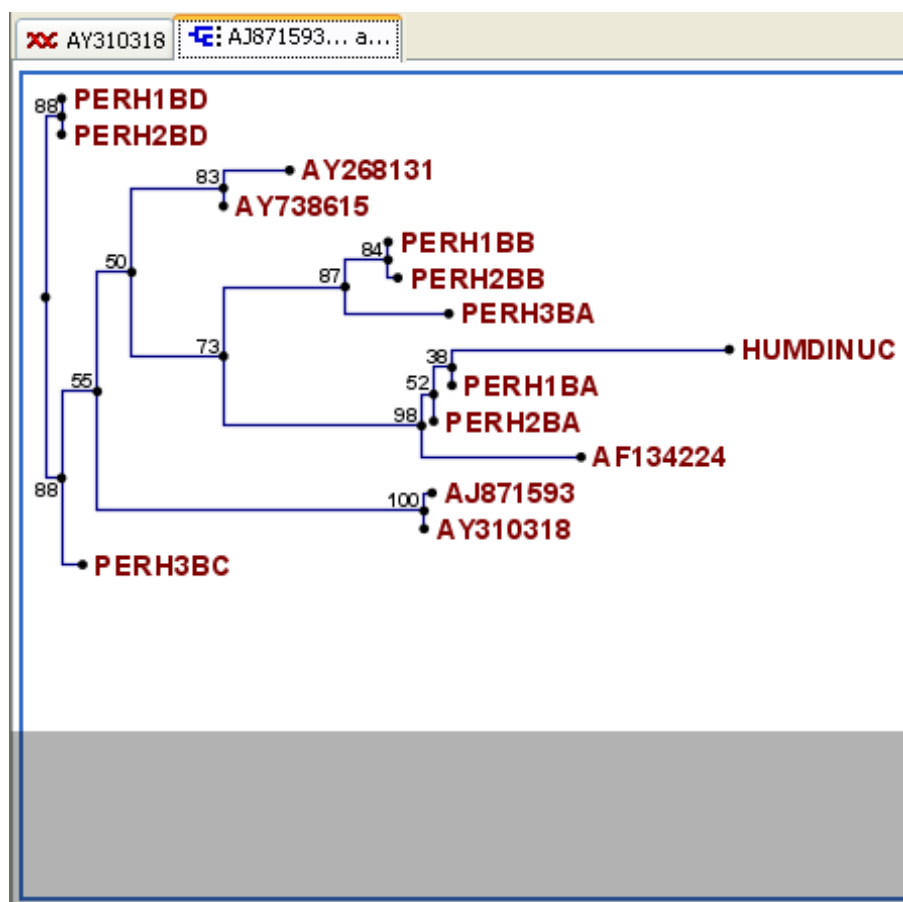


Figure 3.9: When dragging a View, a gray area indicates where the View will be shown.

The following restores the size of the **View**:

Ctrl + M

or **View | Maximize/restore size of View** ()

or **click close-button** () **in the corner of the View Area**

or **double-click title of View**

3.2.6 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.

Side Panel are activated in this way:

select the View | Ctrl + U ( + U on Mac)

or **right-click the tab of the View | View | Show/Hide Side Panel** ()

Notice! Changes made to the **Side Panel** will not be saved when you save the **View**. See how to save the changes in the **Side Panel** in chapter 4 .

The **Side Panel** consists of a number of groups of preferences (depending on the kind of data

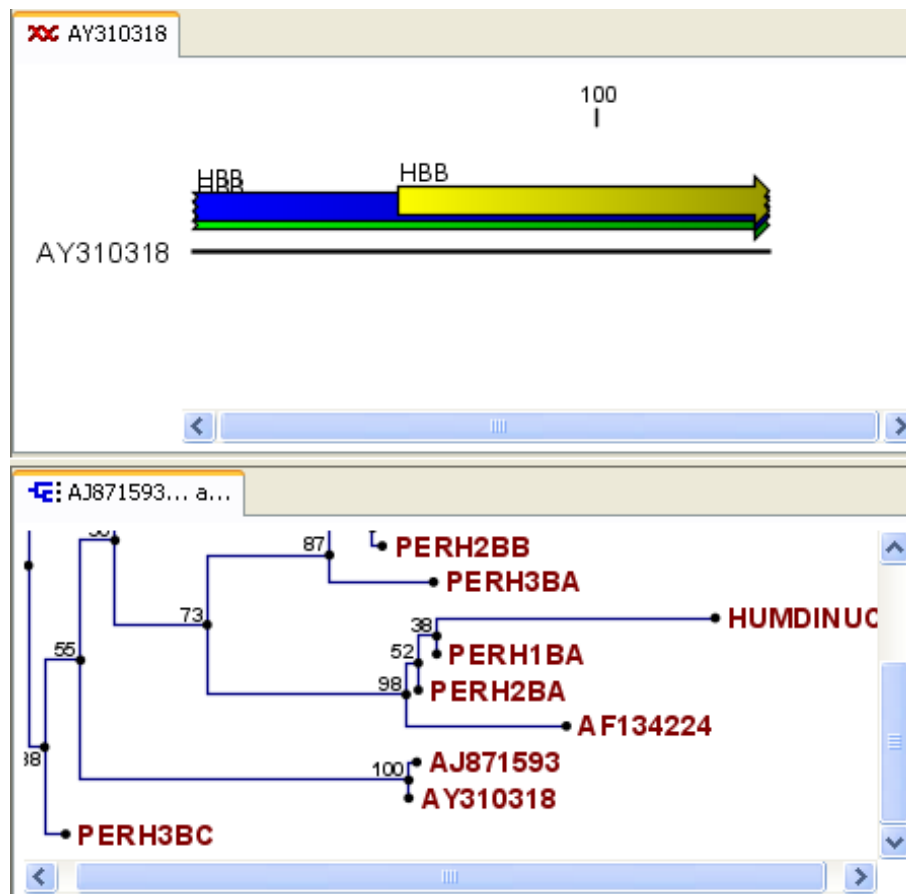




Figure 3.10: A horizontal split-screen. The two Views split the View Area.

being viewed), which can be expanded and collapsed by clicking the header of the group. You can also expand or collapse all the groups by clicking the icons ()/() at the top.

3.3 Zoom and selection in View Area


The mode toolbar items in the right side of the **Toolbar** apply to the function of the mouse pointer. When e.g. **Zoom Out** is selected, the **Zoom Out**-function is applied each time you click in a **View** where zooming is relevant (texts, tables and lists cannot be zoomed). The chosen mode is active until another mode toolbar item is selected. (**Fit Width** and **Zoom to 100%** do not apply to the mouse pointer.)

3.3.1 Zoom In

There are two ways to **Zoom In**:

The first way enables you to zoom in, step by step, on a sequence:

Click Zoom In () in the toolbar | click the location in the view that you want to zoom in on

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

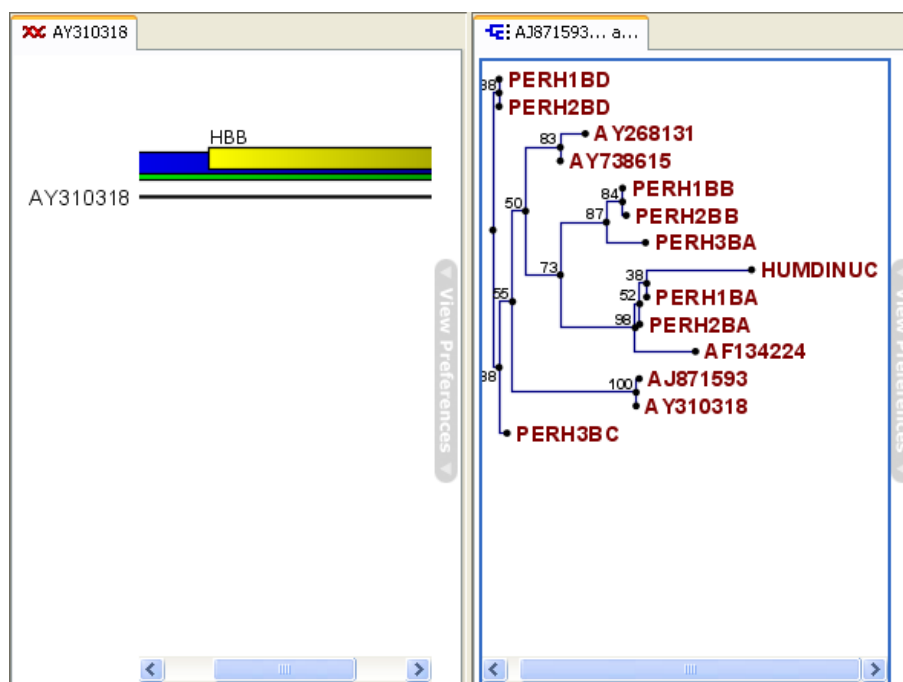


Figure 3.11: A vertical split-screen.

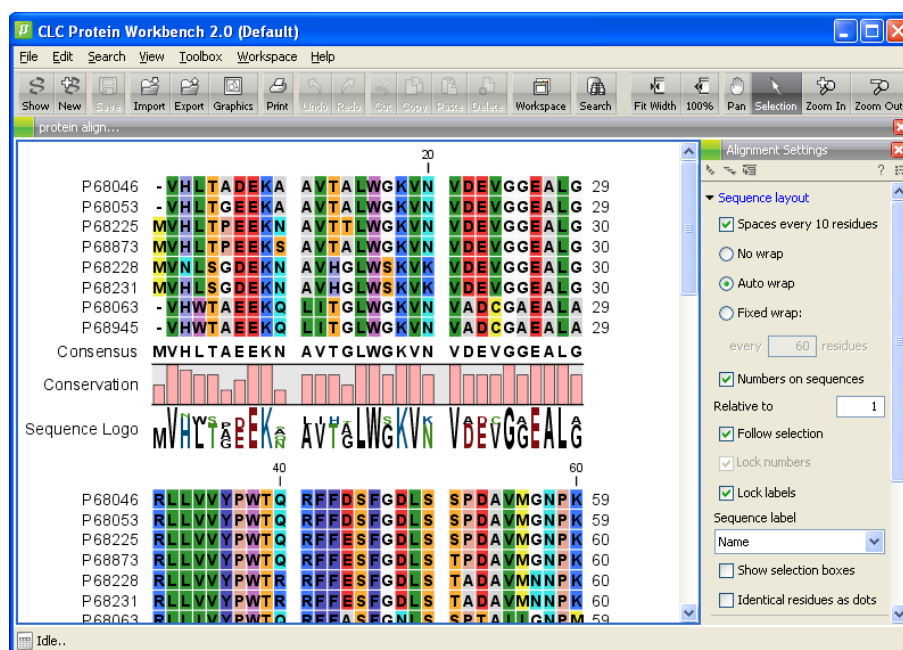


Figure 3.12: A maximized View. The function hides the Navigation Area and the Toolbox.

When you choose the Zoom In mode, the mouse pointer changes to a magnifying glass to reflect the mouse mode.

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



Figure 3.13: The mode toolbar items.

3.3.2 Zoom Out

It is possible to zoom out, step by step, on a sequence:

Click Zoom Out (🔍) in the toolbar | click in the view until you reach a satisfying zoomlevel

When you choose the Zoom In mode, the mouse pointer changes to a magnifying glass to reflect the mouse mode.

If you want to get a quick overview of a sequence or a tree, use the **Fit Width** function instead of the **Zoom Out** function.

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.

3.3.3 Fit Width

The **Fit Width** (📏) function adjusts the content of the **View** so that both ends of the sequence, alignment, or tree is visible in the **View** in question. (This function does not change the mode of the mouse pointer.)

3.3.4 Zoom to 100%

The **Zoom to 100%** (📏) function zooms the content of the **View** so that it is displayed with the highest degree of detail. (This function does not change the mode of the mouse pointer.)

3.3.5 Move

The Move mode allows you to drag the content of a **View**. E.g. if you are studying a sequence, you can click anywhere in the sequence and hold the mouse button. By moving the mouse you move the sequence in the **View**.

3.3.6 Selection

The Selection mode (🖱️) is used for selecting in a **View** (selecting a part of a sequence, selecting nodes in a tree etc.). It is also used for moving e.g. branches in a tree or sequences in an alignment.

When you make a selection on a sequence or in an alignment, the location is shown in the bottom right corner of your workbench. E.g. '23^24' means that the selection is between two residues. '23' means that the residue at position 23 is selected, and finally '23..25' means that 23, 24 and 25 are selected. By holding ctrl / ⌘ you can make multiple selections.

3.4 Toolbox and Status Bar

The **Toolbox** is placed in the left side of the user interface of *CLC Protein Workbench 2.0* below the **Navigation Area**.

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

3.4.1 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.

Active buttons are blue.

If a process is terminated, the stop, pause, and play buttons of the process in question are made gray.

The terminated processes can be removed by:

View | Remove Terminated Processes (X)

Running and paused processes are not deleted.

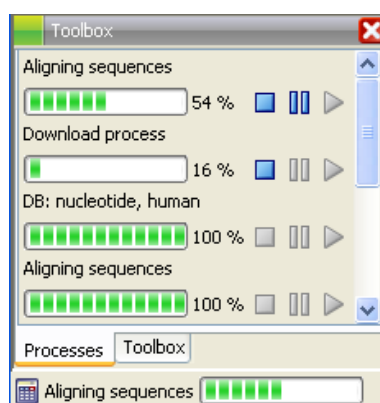


Figure 3.14: Two running, and a number of terminated processes in the Toolbox.

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.

3.4.2 Toolbox

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

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

View | Show/Hide Toolbox

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**.

3.4.3 Status Bar

As can be seen from figure 3.1, 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 3.3.6 for more about the Selection mode button.)

3.5 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 Protein Workbench 2.0*. The next time you run the program, the **Workspaces** are reopened exactly as you left them.

Notice! It is not possible to run more than one version of *CLC Protein Workbench 2.0* at a time. Use two or more **Workspaces** instead.

3.5.1 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 Protein Workbench* opens one **Workspace**, (the largest window in the right side of the workbench, see 3.1). 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 **Project Tree** in the **Navigation Area** is collapsed and the **View Area** is empty and ready to work with. (See figure 3.15).

3.5.2 Select Workspace

When there is more than one **Workspace** in the 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 Protein Workbench 2.0*" at the top left corner of the main window, in this case: (default).

3.5.3 Delete Workspace

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

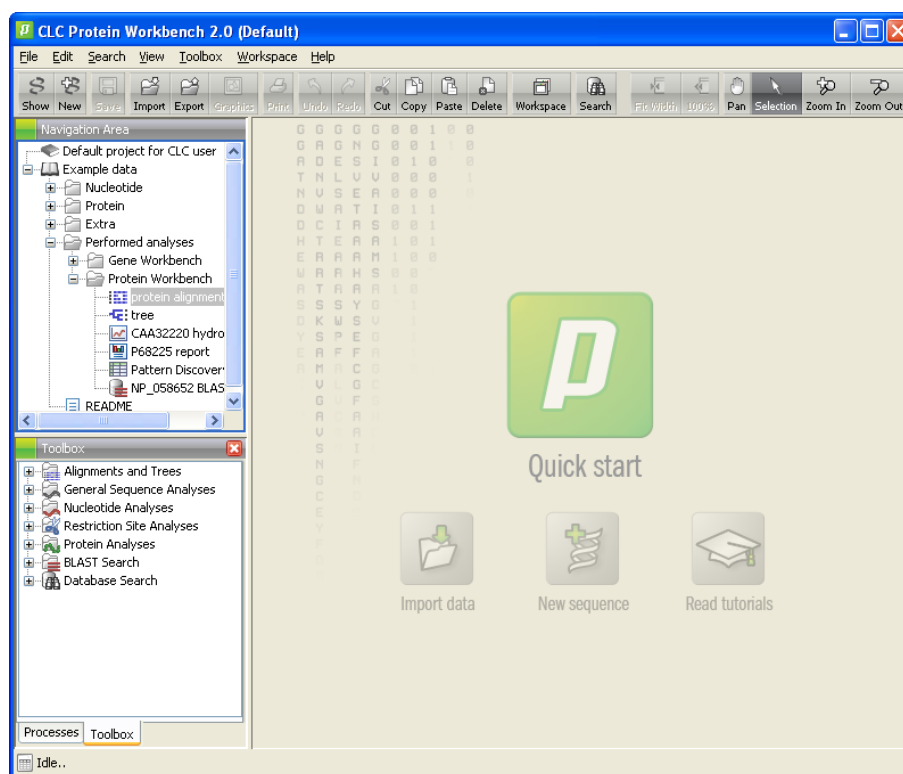


Figure 3.15: An empty Workspace.

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

Notice! 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.

3.6 List of shortcuts

The keyboard shortcuts in *CLC Protein Workbench 2.0* are listed below.

Action	Windows/Linux	Mac OS X
Adjust selection	Shift + arrow keys	Shift + arrow keys
Change between tabs	Ctrl + tab	⌘ + tab
Close	Ctrl + W	⌘ + W
Close all views	Ctrl + Shift + W	⌘ + Shift + W
Copy	Ctrl + C	⌘ + C
Cut	Ctrl + X	⌘ + X
Delete	Delete	Delete
Exit	Alt + F4	⌘ + Q
Export	Ctrl + E	⌘ + E
Export graphics	Ctrl + G	⌘ + G
Find Inconsistency	Space	Space
Find Previous Inconsistency	,	,
Help	F1	F1
Import	Ctrl + I	⌘ + I
Maximize/restore size of View	Ctrl + M	⌘ + M
Move gaps in alignment	Ctrl + arrow keys	⌘ + arrow keys
Navigate sequence views	left/right arrow keys	left/right arrow keys
New Folder	Ctrl + Shift + N	⌘ + Shift + N
New Project	Ctrl + R	⌘ + R
New Sequence	Ctrl + N	⌘ + N
View	Ctrl + O	⌘ + O
Paste	Ctrl + V	⌘ + V
Print	Ctrl + P	⌘ + P
Redo	Ctrl + Y	⌘ + Y
Rename	F2	F2
Save	Ctrl + S	⌘ + S
Search in an open sequence	Ctrl + F	⌘ + F
Search NCBI	Ctrl + B	⌘ + B
Search UniProt	Ctrl + Shift + U	⌘ + Shift + U
Select All	Ctrl + A	⌘ + A
Selection Mode	Ctrl + 2	⌘ + 2
User Preferences	Ctrl + K	⌘ + ;
Split Horizontally	Ctrl + T	⌘ + T
Split Vertically	Ctrl + J	⌘ + J
Show/hide Preferences	Ctrl + U	⌘ + U
Undo	Ctrl + Z	⌘ + Z
Zoom In Mode	Ctrl + + (plus)	⌘ + + (plus)
Zoom In (without clicking)	+ (plus)	+ (plus)
Zoom Out Mode	Ctrl + - (minus)	⌘ + - (minus)
Zoom Out (without clicking)	- (minus)	- (minus)

Combinations of keys and mouse movements are listed below.

Action	Windows/Linux	Mac OS X	Mouse movement
Maximize View			Double-click the tab of the View
Restore View			Double-click the View title
Reverse zoom function	Shift	Shift	Click in view
Select multiple elements	Ctrl	⌘	Click elements
Select multiple elements	Shift	Shift	Click elements

Chapter 4

User preferences

Contents

4.1 General preferences	71
4.2 Default View preferences	71
4.3 Advanced preferences	72
4.4 Export/import of preferences	72
4.5 View preference style sheet	72
4.5.1 Floating Side Panel	73

The **Preferences** dialog offers opportunities for changing the default settings for different features of the program. For example, if you adjust **Number of hits** under **General Preferences** to 40 (instead of 50), you see the first 40 hits each time you conduct a search (e.g. NCBI search).

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)

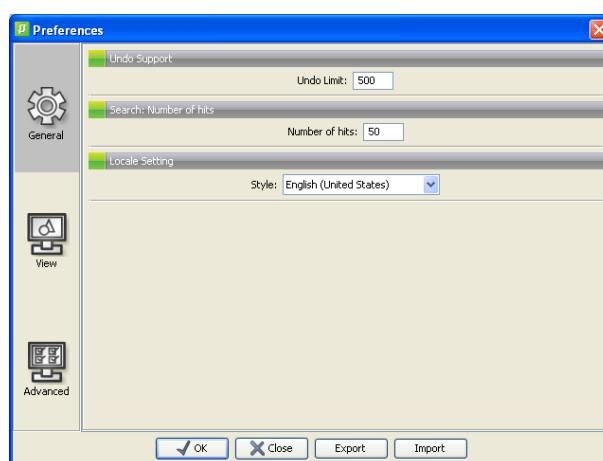


Figure 4.1: Preferences include General preferences, View preferences, Colors preferences, and Advanced settings.

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 sequences, alignments or trees. See section 3.2.4 for more on this topic.
- **Number of hits.** The number of hits shown in *CLC Protein Workbench 2.0*, when e.g. searching NCBI. (The sequences shown in the program are not downloaded, until dragged/saved into the Navigation Area.
- **Locale Setting.** i.e. in which country you are located. This determines the punctuation to be used.

4.2 Default View preferences

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

1. **Toolbar**
2. **Side Panel Location**
3. **New View**
4. **View Format**
5. **Default view settings sheet.**

In general, these are default settings for the user interface.

The **fToolbar** preferences let you choose the size of the toolbar icons, and you can choose whether to display names below the icons.

The **Side Panel Location** setting lets you choose between **Dock in views** and **Float in window**. When docked in view, view preferences will be located in the right side of the view of e.g. an alignment. When floating in window, the side panel can be placed everywhere in your screen, also outside the workspace, e.g. on a different screen. See section 4.5 for more about floating side panels.

The **New view** setting allows you to choose whether the **View preferences** are to be shown automatically when opening a new view. If this option is not chosen, you can press (Ctrl + U (⌘ + U on Mac)) to see the preferences panels of an open view.

The **View Format** 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).
- Species.

- Species (accession).
- Common Species.
- Common Species (accession).

The **User Defined View Settings** gives you an overview of different style sheets for your **View preferences**. See section 4.5 for more about how to create and save style sheets.

The first time you use the program, only the **CLC Standard Settings** is available. However, the tab allowing you to choose the style sheet for a viewer (e.g. a sequence viewer) only appears after you have launched the viewer for the first time.

4.3 Advanced preferences

The **Advanced** settings include the possibility to set up a proxy server. This is described in section 1.7 .

4.4 Export/import of preferences

The user preferences of the *CLC Protein Workbench 2.0* 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

Notice! 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.

Notice! Before exporting, you are asked about which of the different settings you want to include in the exported file. "Default View Settings Sheet", which is one of the preferences which can be selected for export, does not include the Style sheets themselves, but only information about which of the Style sheets is default style sheets.

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

4.5 View preference style sheet

Depending on which view you have opened in the Workbench, you have different options of adjusting the **View preferences**.

Figure 4.2 shows the preference groups which are available for a sequence.

By clicking the black triangles, the different preference groups can be opened. An example is shown in figure 4.3.



Figure 4.2: View preferences for a view of a sequence include several preference groups. In this case the groups are: Sequence layout, Annotation types, Annotation layout, etc. Several of these preference groups are present in more views. E.g. Sequence layout is also present when an alignment is viewed.

The content of the different preference groups, are described in connection to those chapters where the functionality is explained. E.g. **Sequence Layout** View preferences are described in chapter 11.1.1 which is about editing options of a sequence view.

When you have adjusted a view of e.g. a sequence, your settings can be saved in a so called style sheet. When you open other sequences, which you want to display in a similar way, the saved style sheet can be applied. These options are available in the top of the View preferences. (See figure 4.4).

To manage style sheets click (☰) seen in figure 4.4. This opens a menu, where the following options are available:

- Save Settings
- Delete Settings
- Apply Saved Settings

Style sheets for the View preference differ between views. Hence, you can have e.g. three style sheets for sequences, two for alignments, and four for graphs. To adjust which of the style sheets is default for e.g. an alignment, go to the general **Preferences** (Ctrl + K (⌘ + ; on Mac).

CLC Standard Settings represents the way the program was set up, when you first launched the program.

The remaining icons of figure 4.4 are used to; **Expand all preferences**, **Collapse all preferences**, and **Dock/Undock Preferences**. **Dock/Undock Preferences** is used when making the View preferences "floating". See next section

4.5.1 Floating Side Panel

The Side Panel of the views can be placed in the right side of a view, or they can be floating. (See figure 4.5).

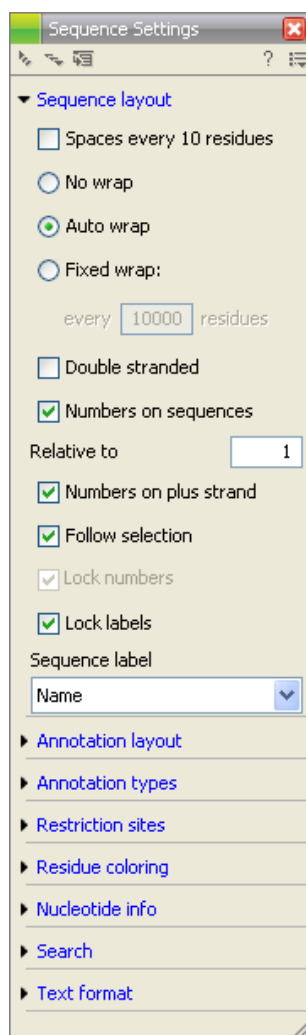


Figure 4.3: The many preferences for each view are stored in preference groups which can be opened and closed.

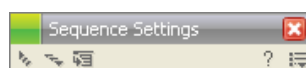



Figure 4.4: The top of the View preferences contain Expand all preferences , Collapse all preferences, Dock/Undock preferences, Help, and Save/Restore preferences.

By clicking the Dock icon () the floating Side Panel reappear in the right side of the view. The size of the floating Side Panel can be adjusted by dragging the hatched area in the bottom right.

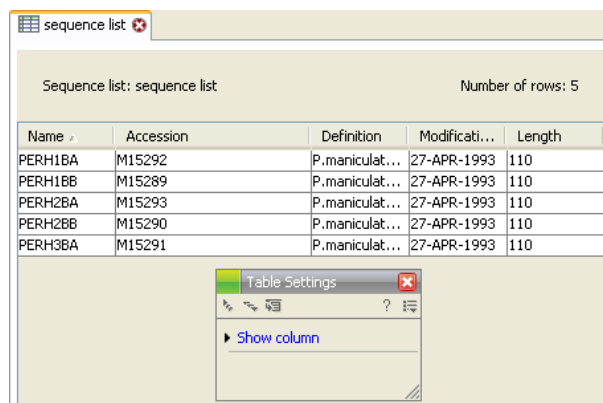


Figure 4.5: The floating Side Panel can be moved out of the way, e.g. to allow for a wider view of a table.

Chapter 5

Printing

Contents

5.1 Selecting which part of the view to print	76
5.2 Page setup	77
5.3 Print preview	77

CLC Protein Workbench 2.0 offers different choices of printing the result of your work.

This chapter deals with printing directly from the workbench. Another option for using the graphical output of your work, is to export graphics (see chapter 6.3) in a graphic format, and then import it into a document or into a presentation.

All the kinds of data that you can view in the **View Area** can be printed. 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:

select relevant view | Print () in the toolbar

If you are printing e.g. alignments, sequences and graphs, you will be faced with three different dialogs, allowing you to adjust the way your view is printed.

- A dialog to let you select which part of the view you want to print.
- A dialog to adjust page setup.
- A **Print preview** window.

These three kinds of dialogs are described in the two following sections.

5.1 Selecting which part of the view to print

Views that are printed exactly like they look on the screen, have an option for selecting which part of the view to print (see figure 5.1).

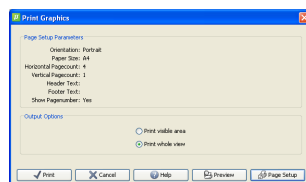


Figure 5.1: When printing graphics you get the options of printing the visible area or printing the whole view.

Printing the whole view is useful if you have zoomed in on an area of the view, and you want to print the whole view (also the part of e.g. a sequence, which is not visible). On the other hand, if you want to print some details of an area of the view, you can use the zoom and navigate functions first, and then print the visible area. This will result in a print of only some 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 5.2

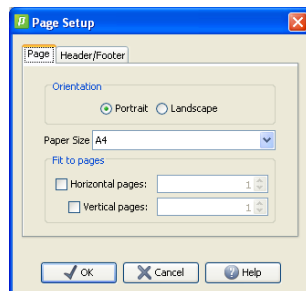


Figure 5.2: In this dialog the default settings Portrait and A4 apply to print of an alignment. By checking Fit to pages it is possible to adjust Horizontal pages to 2. This is done allow a long sequence to stretch the width of two A4 pages. This is illustrated in the Page Layout field.

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** to see the print preview with the settings you have made.

5.3 Print preview

The preview is shown in figure 5.3).

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.

Notice that if you wish to change e.g. the colors of the residues in the alignment, this must be changed in the **View preferences** of the specific dot plot.

Figure 5.3: *Print preview.*

Chapter 6

Import/export of data and graphics

Contents

6.1 Bioinformatic data formats	79
6.1.1 Import of bioinformatic data	80
6.1.2 Export of bioinformatic data	82
6.2 External files	84
6.2.1 Import external files	84
6.2.2 Export external files	84
6.2.3 Technical details	85
6.3 Export graphics to files	85
6.3.1 Exporting protein reports	87
6.4 Copy/paste view output	87

CLC Protein Workbench 2.0 handles a large number of different data formats. All data stored in the Workbench is available in the **Navigation Area** of the program. The data of the **Navigation Area** can be divided into two groups. The data is either one of the different bioinformatic data formats, or it can be an 'external file'. Bioinformatic data formats are those formats which the program can work with, e.g. sequences, alignments and phylogenetic trees. External files are files or links which are stored in *CLC Protein Workbench 2.0*, but are opened by other applications, e.g. pdf-files, Microsoft Word files, Open Office spreadsheet files, or it could be links to programs and webpages etc.

Furthermore, this chapter deals with the export of graphics.

6.1 Bioinformatic data formats

The different bioinformatic data formats are imported in the same way, therefore, the following description of data import is an example which illustrates the general steps to be followed, regardless of which format you are handling.

6.1.1 Import of bioinformatic data

Here follows a short list of the formats which *CLC Protein Workbench 2.0* handles, and a description of which type of data the different formats support.

File type	Suffix	File format used for
Phylip Alignment	.phy	alignments
GCG Alignment	.msf	alignments
Clustal Alignment	.aln	alignments
Newick	.nwk	trees
FASTA	.fsa/.fasta	sequences
GenBank	.gbk/.gb/.gp	sequences
GCG sequence	.gcg	sequences (only import)
PIR (NBRF)	.pir	sequences (only import)
Staden	.sdn	sequences (only import)
VectorNTI		sequences (only import)
DNAstrider	.str/.strider	sequences
Swiss-Prot	.swp	protein sequences
Lasergene sequence	.pro	protein sequence (only import)
Lasergene sequence	.seq	nucleotide sequence (only import)
Embl	.embl	nucleotide sequences
Nexus	.nxs/.nexus	sequences, trees, alignments, and sequence lists
CLC	.clc	sequences, trees, alignments, reports, etc.
Text	.txt	all data in a textual format
ABI		Trace files (only import)
AB1		Trace files (only import)
SCF2		Trace files (only import)
SCF3		Trace files (only import)
Phred		Trace files (only import)
mmCIF	.cif	structure (only import)
PDB	.pdb	structure (only import)
Preferences	.cpf	CLC workbench preferences

Notice that *CLC Protein Workbench* can import 'external' files, too. This means that *CLC Protein Workbench* can import all files and display them in the **Navigation Area**, while the above mentioned formats are the types which can be read by *CLC Protein Workbench*.

The *CLC Protein Workbench 2.0* offers a lot of possibilities to handle bioinformatic data. Read the next sections to get information on how to import different file formats or to import data from a Vector NTI database.

Import of common bioinformatic data

Before importing a file, you must decide where you want to import it, i.e. which project or folder. The imported file ends up in the project or folder you selected in the **Navigation Area**.

select project or folder | click Import (📁) in the Toolbar | browse to the relevant file | Select

The imported file is placed at the location which was selected when the import was initiated. E.g. if you right-click on a file in the **Navigation Area** and choose import, the imported file is placed

immediately below the selected file. If you right-click a folder, the imported file is placed as the last file in that folder. If you right-click a project, the imported file is placed as the last file in that project (and after existing folders).

It is also possible to drag a file from e.g. the desktop into the **Navigation Area** of *CLC Protein Workbench*. If *CLC Protein Workbench* recognizes the file format, the file is automatically parsed (changed) into CLC format and stored in the **Navigation Area**. If the format is not recognized, the following dialog is displayed (see figure 6.1):

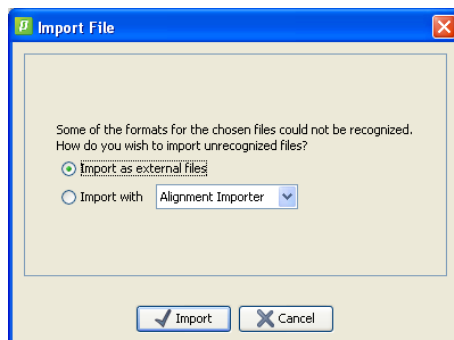


Figure 6.1: If the dragged file is not recognized by **CLC Protein Workbench** the dialog allows you to "force" the import in a certain format.

Notice! When browsing for files to import, the dialog only displays files of the format chosen in the **File of type** drop down menu at the bottom of the import dialog. If the format .clc is chosen, only .clc-files are shown in the **Import** dialog. Choose **All Files** to ensure the file you are looking for is displayed.

When you import a file containing several sequences, you will be asked whether you want to save the sequences as individual elements or as a sequence list (see section 11.5 for more about sequence lists).

Import of data in clc-format from older versions

If you want to import data in clc-format generated in an older version of either of the workbenches, it has to be converted first. If you try to import it without conversion, you will see a warning dialog.

Import of Vector NTI data

CLC Protein Workbench 2.0 can import DNA, RNA, and protein sequences from a Vector NTI Database. The import can be done for Vector NTI Advance™10 for Windows machines and Vector NTI Suite 7.1 for Mac OS X for Panther and former versions. A new Project will be placed in the **Navigation Area** and you can find all sequences in different folders ready to work with. In order to import all DNA/RNA and protein sequences:

**select File in the Menu Bar | Import VectorNTI Data.. | select a database directory
| Import | confirm the information**

Notice! The default installation of the VectorNTI program for the database home is

- C:/VNTI Database/
for Windows machines and

- /Library/Application Support/VNTI Database/
for Mac OS X for Panther.

Therefore the *CLC Protein Workbench 2.0* will check if there is a default installation and will ask whether you want to use the default database directory or another directory.

Notice! Make sure that the Vector NTI database directory (default or backup) contains folders like ProData and MolData. These folders are necessary when we import the data into *CLC Protein Workbench 2.0*.

In order to import all DNA/RNA and protein sequences if a default database directory is installed:

select File in the Menu Bar | Import VectorNTI Data | select Yes if you want to import the default database | confirm the information

or **select File in the Menu Bar | Import VectorNTI Data | select No to choose a database | select a database directory | Import | confirm the information**

After the import there is a new Project called **Vector NTI Data** in the **Navigation Area**. In **Vector NTI Data** you can see two folders: **DNA/RNA** containing the DNA and RNA sequences, and **Protein** containing all protein sequences. (See figure 6.2).

The project, folders and all sequences are automatically saved.

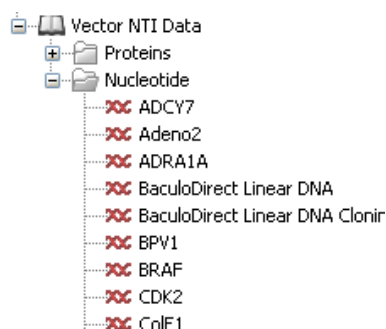


Figure 6.2: Project Vector NTI Data containing all imported sequences of the Vector NTI Database.

6.1.2 Export of bioinformatic data

CLC Protein Workbench 2.0 can export bioinformatic data in most of the formats that can be imported. There are a few exceptions. See section 6.1.1.

To export a file:

select the element to export | Export () | choose where to export to | select 'File of type' | enter name of file | Save

Notice! The **Export** dialog decides which types of files you are allowed to export into, depending on what type of data you want to export. E.g. protein sequences can be exported into GenBank, Fasta, Swiss-Prot and CLC-formats.

Export of projects, folders and multiple files

The .clc file type can be used to export all kinds of files and is therefore especially useful in these situations:

- Export of one or more file folders including all underlying files and folders.
- Export of one or more project folders including all underlying files and folders.
- If you want to export two or more files into one .clc-file, you have to copy them into a folder or project, which can be exported as described below:

Export of projects and folders is similar to export of single files. Exporting multiple files (of different formats) is done in .clc-format. This is how you export a project:

select the project to export | Export (📁) | choose where to export to | enter name of project | Save

You can export multiple files of the same type into formats other than CLC (.clc). E.g. two DNA sequences can be exported in GenBank format:

select the elements to export by <Ctrl>-click or <Shift>-click | Export (📁) | choose where to export to | choose GenBank (.gbk) format | enter name of project | Save

Export of dependent objects

When exporting e.g. an alignment, *CLC Protein Workbench 2.0* can export all dependent objects. I.e. the sequences which the alignment is calculated from. This way, when sending your alignment (with the dependent objects), your colleagues can reproduce your findings with adjusted parameters, if desired.

To export with dependent files:

select the element in Navigation Area | File in Menu Bar | Export with dependent objects | enter name of project | choose where to export to | Save

The result is a folder containing the exported file with dependent objects, stored automatically in a folder on the desired location of your desk.

Export history

To export an element's history:

select the element in Navigation Area Export(📁) | select History PDF(.pdf) | choose where to export to | Save

The entire history of the element is then exported in pdf format.

The CLC format

CLC Protein Workbench keeps all bioinformatic data in the CLC format. Compared to other formats, the CLC format contains more information about the object, like its history and comments. The CLC format is also able to hold several objects of different types (e.g. an alignment, a graph and a phylogenetic tree). This means that if you are exporting your data to another CLC Workbench, you can use the CLC format to export several objects in one file, and all the objects' information is preserved.

Notice! CLC files can be exported from and imported into all the different CLC Workbenches.

Back up

The CLC format is practical for making manual back up of your files. All files are stored in Projects and these can easily be exported out of *CLC Protein Workbench*, :

select the project to export | Export (📁) | choose where to export to | enter name of project | Save

Other than that, the files of the **Navigation Area** are stored in a persistence folder on your computer. Hence, your regular back up system should be set up to include this folder.

On Mac the folder can be found: Library/Application Support/CLC bio/Workbench/<version number>/persistence

On Windows: Documents and Settings/<username>/CLC bio/Workbench/<version number>/persistence

On Linux: home/<username>/.clcbio/workbench/<version number>/persistence

6.2 External files

In order to help you organize your projects, *CLC Protein Workbench 2.0* 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 a project in *CLC Protein Workbench 2.0*. Importing an external file creates a copy of the file which is saved in a project in *CLC Protein Workbench 2.0*. The file can now be opened by double-clicking the file name 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).

CLC Protein Workbench can also show web links (URLs) in the **Navigation Area**. This can be done by using the **Import** function of the program or by dragging the file e.g. from the desktop to the **Navigation Area**.

6.2.1 Import external files

To import an external file:

click a project or folder to import into | Import (📁) in the toolbar | Choose All files in Files of type | browse to the relevant file | Select

or **drag the file from the file system into a project in the Navigation Area (only possible under Windows)**

Notice! When you import an external file, a copy of the original file is created. This means that you should always make sure that you open the file from within *CLC Protein Workbench 2.0*.

6.2.2 Export external files

If you export an entire project or folder from *CLC Protein Workbench 2.0*, the exported CLC file will include all external files stored in it. This means that you can export the project as a CLC file, and send it to a colleague who can import it and access all the files in the project.

You can also export individual files in their original format. To export a file from *CLC Protein Workbench 2.0*:

click a file in the Navigation Area | Export (📁) in the toolbar | browse to the desired folder | Save

If the file already exists, you are asked if you want to replace it.

6.2.3 Technical details

This section explains the more technical aspects of how *CLC Protein Workbench 2.0* stores the external files. When you import the file, a copy of the file is created in a database. When you open the file from the **Navigation Area**, it's checked out to a repository (a folder called "CLCWorkbenchRepository" located in your operating system's user folder) where it stays until you close the application that has the file open. When you exit *CLC Protein Workbench 2.0*, it checks all the files in the repository into the database, unless they are still open in another application. If the latter is the case, the file stays in the repository even after the file is closed, and it will not be checked in until the next time *CLC Protein Workbench 2.0* is closed.

If you have made changes to a file after the *CLC Protein Workbench 2.0* was closed, a dialog is shown asking which version to use. The date and time of the latest change of the file is displayed in the dialog helping you to decide which one to keep (see figure 6.3).

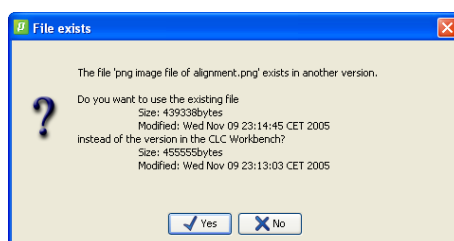


Figure 6.3: A dialog asking which version of the file you want to keep.

6.3 Export graphics to files

CLC Protein Workbench 2.0 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 Protein Workbench 2.0 exports graphics exactly the way it is shown in the **View Area**. Thus, all settings made in the **Side Panel** will be reflected in the exported file.

To show you how to export graphics, we choose to export the phylogenetic tree of the example data set in .png-format. See 6.4.

When the relevant file is opened and shown in the **View Area** do the following:

select tab of View | Graphics (🖨️) on Toolbar | select location on disc | name file and select type | Save

After clicking **Save**, you are prompted for whether to **Export visible area** or **Export whole view**. The first parameter exports 'what you see' and the latter parameter also exports the part of the view that is not visible. Hence, choosing **Export whole view** will generate a larger file.

Furthermore, when saving in .png, .jpg, and .tif-formats you are prompted for which quality to save the graphics in.

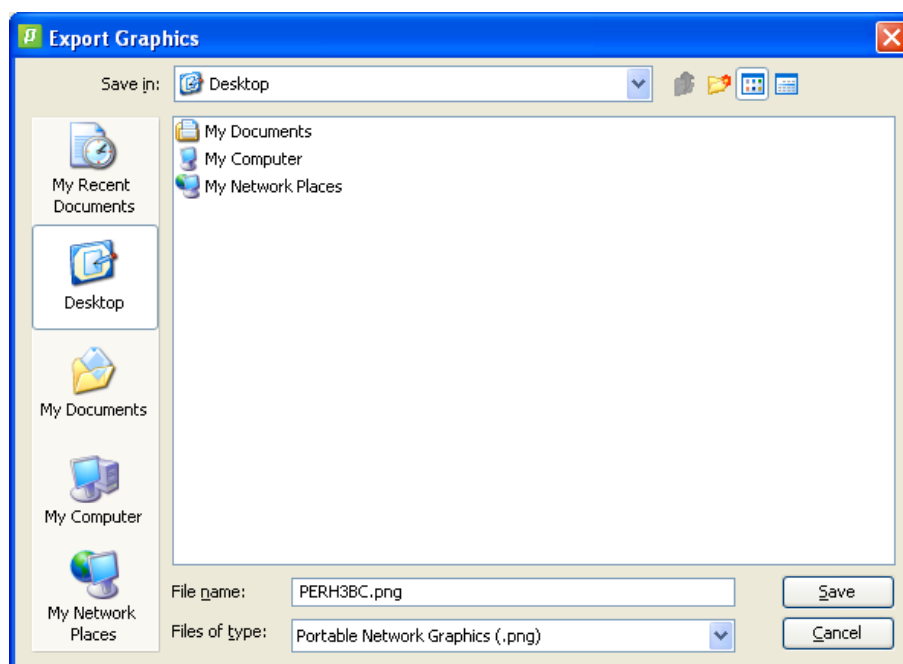


Figure 6.4: Exporting a phylogenetic tree from **CLC Protein Workbench 2.0**.

To see the exported file browse to the file on your computer and open it. In our case the .png-file is opened in a browser, the result can be seen in figure 6.5.

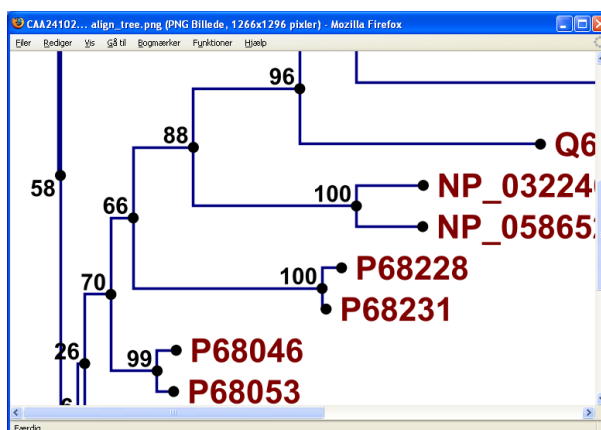


Figure 6.5: The exported .png-file opened in a browser. (Due to high resolution of the exported graphics, it is not possible to see the entire file in the browser window.)

The following file types are available for exporting graphics in *CLC Protein Workbench 2.0*:

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. This format is a good choice for storing images without large shapes (e.g. dot plots).

Vector graphics

Vector graphics 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 zoomfactor is, thereby always giving a correct

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

image. This format is good for e.g. graphs and reports, but less usable for e.g. dotplots.

Graphics files can also be imported into the **Navigation Area**. However, no kinds of graphics files can be displayed in *CLC Protein Workbench 2.0*. See section 6.2.1 for more about importing external files into *CLC Protein Workbench 2.0*.

6.3.1 Exporting protein reports

Protein reports cannot be exported in the same way as other data. Instead, they can be exported from the **Navigation Area**:

Click the report in the Navigation Area | Export (📎) in the Toolbar | select pdf

When the report is exported, the file can be opened with Adobe Reader. Opening and printing in Adobe Reader is also the only way to print the report.

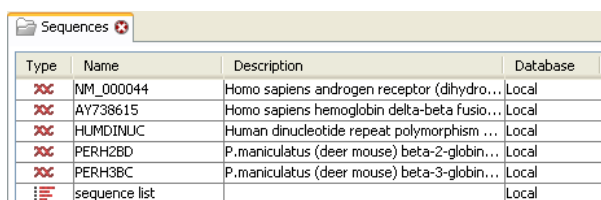
6.4 Copy/paste view output

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 Protein Workbench 2.0* 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 | Push arrow down (or up)

See figure 6.6.



Type	Name	Description	Database
X	NM_000044	Homo sapiens androgen receptor (dihydro...	Local
X	AY738615	Homo sapiens hemoglobin delta-beta fusio...	Local
X	HUMDINUC	Human dinucleotide repeat polymorphism ...	Local
X	PERH2BD	P.maniculatus (deer mouse) beta-2-globin...	Local
X	PERH3BC	P.maniculatus (deer mouse) beta-3-globin...	Local
	sequence list		Local

Figure 6.6: 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 Protein Workbench 2.0* can be produced. (Except the icons which are replaced by file references in Excel.)

Chapter 7

History

Contents

7.1 Element history	89
7.1.1 Sharing data with history	90

CLC Protein Workbench 2.0 keeps a log of all operations you make in the program. If e.g. 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.

This can be useful in several situations: It can be used for documentation purposes, where you can specify exactly how your data has been created and modified. It can also be useful if you return to a project after some time and want to refresh your memory on how the data was created. Also, if you have performed an analysis and you want to reproduce the analysis on another element, you can check the history of the analysis which will give you all parameters you set.

This chapter will describe how to use the **History** functionality of *CLC Protein Workbench 2.0*.

7.1 Element history

You can view the history of all elements in the **Navigation Area** except files that are opened in other programs (e.g. Word and pdf-files). The history starts when the element appears for the first time in *CLC Protein Workbench 2.0*. To view the history of an element:

Right-click the element in the Navigation Area | Show | History ()

or **Select the element in the Navigation Area | Show** () **in the Toolbar | History** ()

This opens a view that looks like the one in figure 7.1.

When opening an element's history is opened, the newest change is submitted in the top of the view. The following information is available:

- **Title.** The action that the user performed.

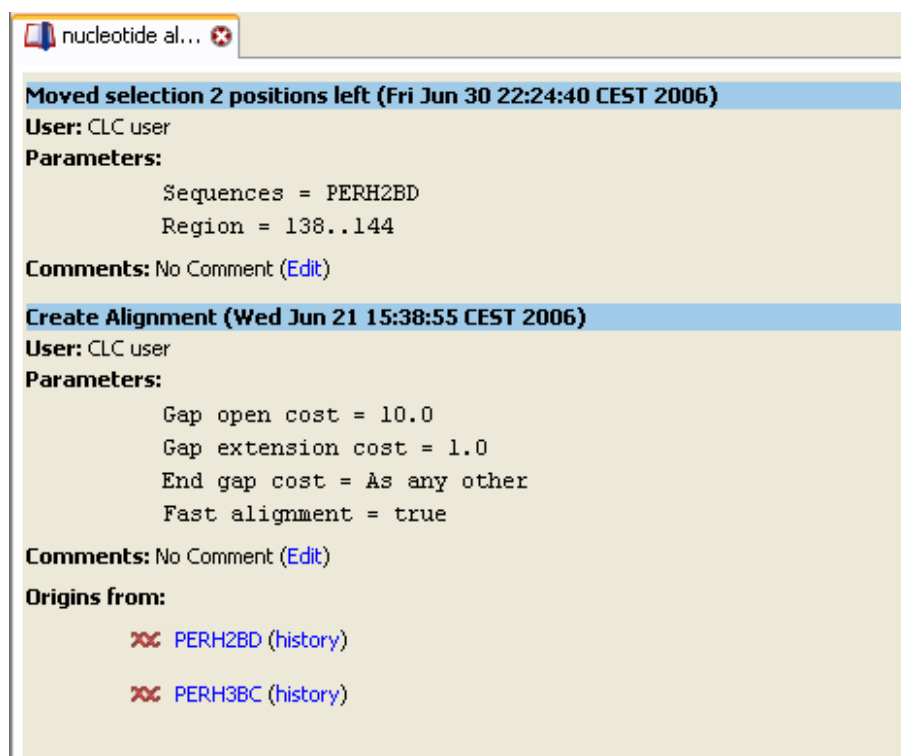



Figure 7.1: An element's history.

- **Date and time.** Date and time for the operation. The date and time are displayed according to your locale settings (see section 4.1).
- **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 was chosen for an analysis.
- **Origins 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. If you have e.g. 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.

7.1.1 Sharing data with history

The history of an element is attached to that element, which means that exporting an element in CLC format (*.clc) will export the history too. In this way, you can share projects and files with others while preserving the history. If an element's history includes source elements (i.e. if there are elements listed in 'Origins from'), they must also be exported in order to see the full history. Otherwise, the history will have entries named "Element deleted". An easy way to export an element with all its source elements is to use the **Export Dependent Objects** function described in section 6.1.2.

The of a history view can be printed. To do so, click the **Print** icon (.

Chapter 8

Handling of results

Contents

8.1 How to handle results of analyses	91
8.1.1 When the analysis does not create new elements	91
8.1.2 Batch log	92

Most of the analyses in the **Toolbox** are able to perform the same analysis on several elements in one batch. This means that analyzing large amounts of data is very easily accomplished. If you e.g. wish to translate a large number of DNA sequence to protein, you can just select the DNA sequences and set the parameters for the translation once. Each DNA sequence will then be treated individually as if you performed the translation on each of them. The process will run in the background and you will be able to work on other projects at the same time.

8.1 How to handle results of analyses

All the analyses in the **Toolbox** are performed in a step-by-step procedure. First, you select elements for analyses, and then there are a number of steps where you can specify parameters (some of the analyses have no parameters, e.g. when translating DNA to RNA). The final step concerns the handling of the results of the analysis, and it is almost identical for all the analyses so we explain it in this section in general.

In this step, shown in figure 8.1, you have two options:

- **Open.** This will open the result of the analysis in a view. This is the default setting.
- **Save.** This means that the result will not be opened but saved to a folder in the **Navigation Area**. If you select this option, click **Next** and you will see one more step where you can specify where to save the results (see figure 8.2). In this step, you *have to select a folder*. You also have the option of creating a new folder in this step.

8.1.1 When the analysis does not create new elements

When an analysis does not create new elements, as e.g. **Find Open Reading Frames** which adds annotations to the sequences, the options for saving are different (see figure 8.3):

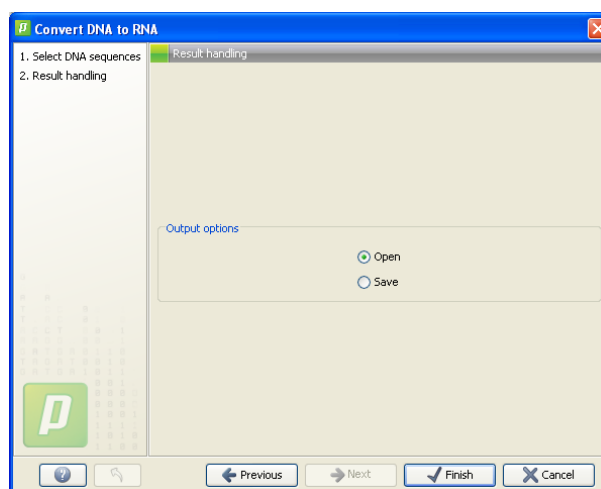


Figure 8.1: The last step of the analyses exemplified by Translate DNA to RNA.

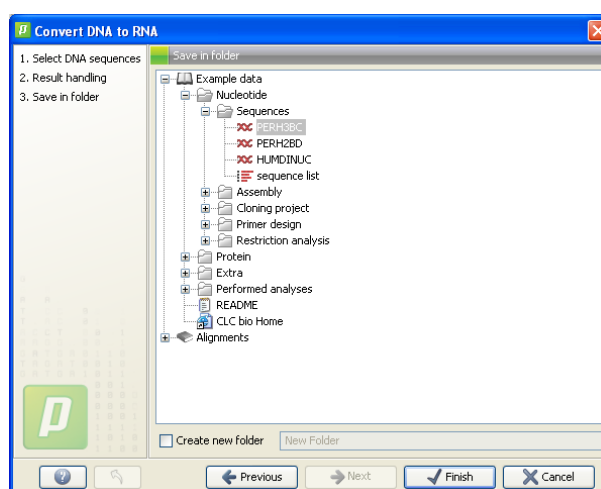


Figure 8.2: Specify a folder for the results of the analysis.

- **Open.** This will open each of the selected sequences in a view.
- **Save.** This will not open the sequences but just add the annotations.
- **Copy and save in new folder.** This option does not add annotations to the existing sequences but saves a copy of the selected sequences. Choosing this option means that there will be an extra step for selecting a folder where the copies of the sequences can be saved.

8.1.2 Batch log

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. See an example of a log in figure 8.4. In this example, the log displays information about how many open reading frames were found.

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.

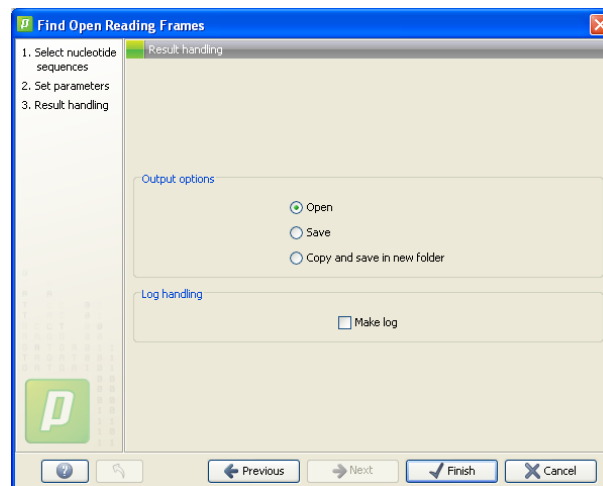


Figure 8.3: The final step when the analysis does not create new elements but add annotations to existing elements.

The screenshot shows a window titled "Log" with a table of results. The table has three columns: "Name", "Description", and "Time". The data is as follows:

Name	Description	Time
HJMIDNUC	Found 5 reading frames	Sun Jun 11 13:06:17 CEST 2006
PERH1BA	Found 5 reading frames	Sun Jun 11 13:06:17 CEST 2006
PERH1BB	Found 5 reading frames	Sun Jun 11 13:06:17 CEST 2006
PERH2BA	Found 4 reading frames	Sun Jun 11 13:06:17 CEST 2006
PERH2BB	Found 4 reading frames	Sun Jun 11 13:06:17 CEST 2006
PERH2BD	Found 7 reading frames	Sun Jun 11 13:06:17 CEST 2006
PERH3BA	Found 3 reading frames	Sun Jun 11 13:06:17 CEST 2006
PERH3BC	Found 7 reading frames	Sun Jun 11 13:06:17 CEST 2006

On the right side of the table, there is a vertical scrollbar and a label "Batch logs".

Figure 8.4: An example of a batch log when finding open reading frames.

Part III

Bioinformatics

Chapter 9

Database search

Contents

9.1 GenBank search	95
9.1.1 GenBank search options	95
9.1.2 Handling of GenBank search results	97
9.2 UniProt (Swiss-Prot/TrEMBL) search	98
9.2.1 UniProt search options	99
9.2.2 Handling of UniProt search results	100
9.3 Sequence web info	101
9.3.1 Google sequence	102
9.3.2 NCBI	102
9.3.3 PubMed References	102
9.3.4 UniProt	102

CLC Protein Workbench 2.0 offers different ways of searching data on the Internet. You must be online when initiating and performing the following searches:

9.1 GenBank search

This section describes searches in GenBank - the **NCBI Entrez** database - and the import of search results. The NCBI search view is opened in this way (figure 9.1):

Search | Search NCBI Entrez 

or **Ctrl + B** (**⌘ + B** on Mac)

This opens the following view:

9.1.1 GenBank search options

Conducting a search in the **NCBI Database** from *CLC Protein Workbench 2.0* corresponds to conducting the search on NCBI's website. When conducting the search from *CLC Protein Workbench 2.0*, the results are available and ready to work with straight away.

You can choose whether you want to search for nucleotide sequences or protein sequences.

NCBI search

Choose database: ☒ Nucleotide ☐ Protein

All Fields

All Fields

All Fields

☐ Append wildcard (*) to search words

Accession	Definition	Modification D...
BC010230	Homo sapiens chromosome 10 open reading frame 83, mRNA (cDNA clo...	2004/03/25
BC015537	Homo sapiens hemoglobin, epsilon 1, mRNA (cDNA clone MGC:9582 IM...	2004/06/29
BC032122	Homo sapiens hemoglobin, alpha 2, mRNA (cDNA clone MGC:29691 IMA...	2003/12/19
BC032264	Mus musculus hemoglobin, beta adult minor chain, mRNA (cDNA clone M...	2006/04/13
BC043020	Mus musculus hemoglobin alpha, adult chain 1, mRNA (cDNA clone MGC...	2004/06/30
BC050661	Homo sapiens hemoglobin, alpha 2, mRNA (cDNA clone MGC:60177 IMA...	2003/10/07
BC051988	Mus musculus hemoglobin X, alpha-like embryonic chain in Hba complex,...	2004/06/30
BC052008	Mus musculus hemoglobin Z, beta-like embryonic chain, mRNA (cDNA cl...	2006/04/27
BC056686	Homo sapiens hemoglobin, theta 1, mRNA (cDNA clone MGC:61857 IMA...	2004/06/30
BC057014	Mus musculus hemoglobin Y, beta-like embryonic chain, transcript varia...	2005/12/09
BC069307	Homo sapiens hemoglobin, delta, mRNA (cDNA clone MGC:96894 IMAG...	2004/06/30

(50 of 236 hits shown)

Figure 9.1: The GenBank search dialog.

As default, *CLC Protein Workbench* offers one text field where the search parameters can be entered. Click **Add search parameters** to add more parameters to your search.

Notice! 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 NCBI database at the same time.
- **Organism.** Text.
- **Description.** Text.
- **Modified Since.** Between 30 days and 10 years.
- **Gene Location.** Genomic DNA/RNA, Mitochondrion, or Chloroplast.
- **Molecule.** Genomic DNA/RNA, mRNA or rRNA.
- **Sequence Length.** Number for maximum or minimum length of the sequence.
- **Gene Name.** Text.

The search parameters are the most recently used. The **All fields** allows searches in all parameters in the NCBI 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 sequences. 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, you can either **Save search parameters** or **Start search**.

When applying the **Save search parameters** option, only the parameters are saved - not the results of the search. The search parameters can also be saved by dragging the tab of the Search view into the **Navigation Area**.

If you don't save the search, the search parameters are saved in **Search NCBI** view until the next time you conduct an NCBI search.

Notice! 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.

The search process runs in the **Toolbox** under the **Processes** tab. It is possible to stop the search process by clicking stop (.

Because the process runs in the **Processes** tab it is possible to perform other tasks while the search is running.

9.1.2 Handling of GenBank 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 sequence hit is represented by text in three columns:

- Accession.
- Definition.
- Modification 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 sequences can be selected, and by clicking the buttons in the bottom of the search view, you can do the following:

- Download and open, doesn't save the sequence.
- Download and save, lets you choose location for saving sequence.
- Open at NCBI, searches the sequence at NCBI'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. Finally, you can also

Drag and drop from GenBank search results

The sequences from the search results can be opened by dragging them into a position in the **View Area**.

Notice! 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 GenBank 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 9.2). Choosing **Save sequence** lets you select a folder or project where the sequences are saved when they are downloaded. Choosing **Open sequence** opens a new view for each of the selected sequences.

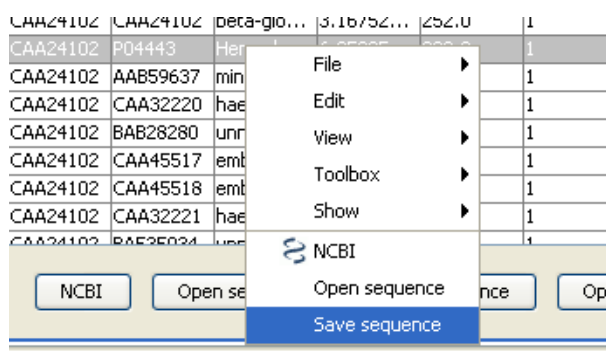


Figure 9.2: By right-clicking a search result, it is possible to choose how to handle the relevant sequence.

Copy/paste from GenBank search results

When using copy/paste to bring the search results into the **Navigation Area**, the actual files are downloaded from GenBank.

To copy/paste files into the **Navigation Area**:

select one or more of the search results | Ctrl + C (⌘ + C on Mac) | select project or folder in the Navigation Area | Ctrl + V

Notice! 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.

9.2 UniProt (Swiss-Prot/TrEMBL) search

The

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 9.3) is opened in this way:

Search | Search UniProt (🔍)

Accession	Name	Description	Organism
Q29397	SV2A_BOVIN	Synaptic vesicle glycoprotein 2A (p87)	Bos taurus (Bovine)
Q4R4N1	CPLX1_MACFA	Complexin-1	Macaca fascicularis (Crab eati...
Q4R4X3	SV2A_MACFA	Synaptic vesicle glycoprotein 2A	Macaca fascicularis (Crab eati...
Q5R4L9	SV2A_PONPY	Synaptic vesicle glycoprotein 2A	Pongo pygmaeus (Orangutan)
Q7L033 Q948...	SV2A_HUMAN	Synaptic vesicle glycoprotein 2A	Homo sapiens (Human)
Q86YN6 Q86...	PRGC2_HUMAN	Peroxisome proliferator-activated receptor g...	Homo sapiens (Human)
Q9GLR1	NEC1_BOVIN	Neuroendocrine convertase 1 precursor (EC ...	Bos taurus (Bovine)
Q9JIS5 Q80T...	SV2A_MOUSE	Synaptic vesicle glycoprotein 2A (Synaptic ve...	Mus musculus (Mouse)

Figure 9.3: The UniProt search dialog.

9.2.1 UniProt search options

Conducting a search in **UniProt** from *CLC Protein Workbench 2.0* corresponds to conducting the search on UniProt's website. When conducting the search from *CLC Protein Workbench 2.0*, 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 curated manually and data are entered according to the original research paper.
- **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 Protein Workbench* offers one text field where the search parameters can be entered. Click **Add search parameters** to add more parameters to your search.

Notice! 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 NCBI 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, you can either **Save search parameters** or **Start search**.

When applying the **Save search parameters** option, only the parameters are saved - not the results of the search. The search parameters can also be saved by dragging the tab of the Search view into the **Navigation Area**.

If you don't save the search, the search parameters are saved in **Search NCBI** view until the next time you conduct an NCBI search.

Notice! 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.

The search process runs in the **Toolbox** under the **Processes** tab. It is possible to stop the search process by clicking stop (.

Because the process runs in the **Processes** tab it is possible to perform other tasks while the search is running.

9.2.2 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

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. Finally, you can also

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**.

Notice! 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 9.2). Choosing **Download and Save** lets you select a folder or project 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 project
or folder in the Navigation Area | Ctrl + V**

Notice! 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.

9.3 Sequence web info

CLC Protein Workbench 2.0 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 11.3). 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 9.4):

Right-click a sequence in the Navigation Area | Sequence Web Info | select the desired search function

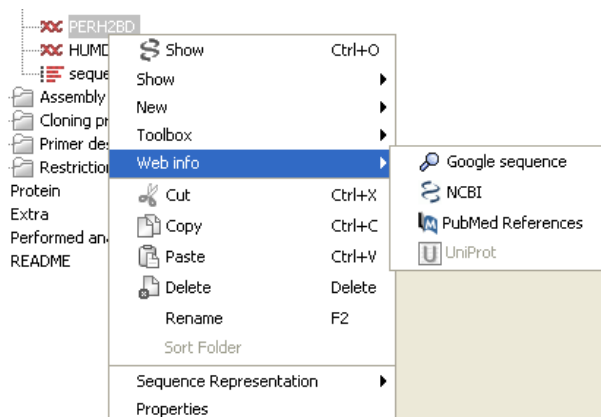


Figure 9.4: By right-clicking a search result, it is possible to choose how to handle the relevant sequence.

This will open your computer's default browser searching for the sequence that you selected.

9.3.1 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>.

9.3.2 NCBI

The NCBI search function searches in GenBank at NCBI (<http://www.ncbi.nlm.nih.gov>) using an identification number (when you view the sequence as text it is the "GI" number). Therefore, the sequence file must contain this number in order to look it up in NCBI. All sequences downloaded from NCBI have this number.

9.3.3 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 see a dialog and the browser will not open.

9.3.4 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.

Chapter 10

BLAST Search

Contents

10.1 BLAST Against NCBI Database	103
10.1.1 Output from BLAST search	106
10.1.2 BLAST table	108
10.2 BLAST Against Local Database	109
10.3 Create Local BLAST Database	110

CLC Protein Workbench offers to conduct BLAST searches on protein and DNA sequences. In short, a BLAST search identifies homologous sequences by searching one or more databases hosted by NCBI (<http://www.ncbi.nlm.nih.gov/>), on your query sequence [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.

From CLC Protein Workbench 2.0 it is also possible to conduct BLAST searches on a database stored locally on your computer. Local BLAST and the creation of a database for local BLAST search is described later in this chapter.

10.1 BLAST Against NCBI Database

To conduct a BLAST search:

right-click the tab of an open sequence | Toolbox | BLAST Search() | BLAST Against NCBI Databases ()

or **click an element in the Navigation Area | Toolbox | BLAST Search() | BLAST Against NCBI Databases ()**

Alternatively, use the keyboard shortcut: Ctrl+Shift+B for Windows and ⌘ +Shift+B on Mac OS. This opens the BLAST dialog. You can not use sequences longer than 8190 for BLAST search.

This opens the dialog seen in figure 10.1

Click **Next**

In **Step 2**, you can choose which type of BLAST search you want to conduct, and you can limit your

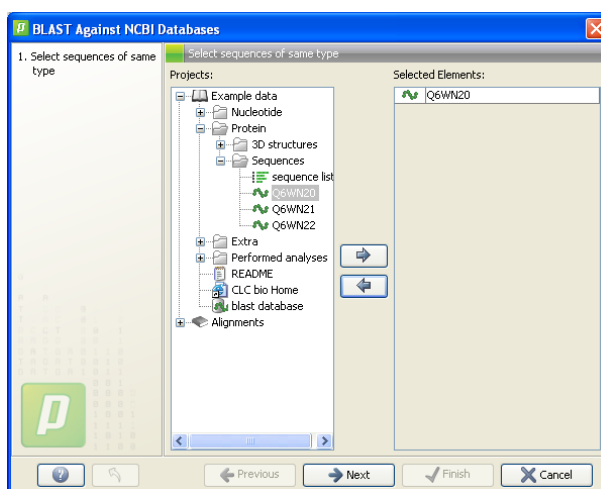


Figure 10.1: Choose one or more sequences to conduct a BLAST search.

search to a particular database (see section B in the appendix for a list of available databases). Step 2 can be seen in figure 10.2:

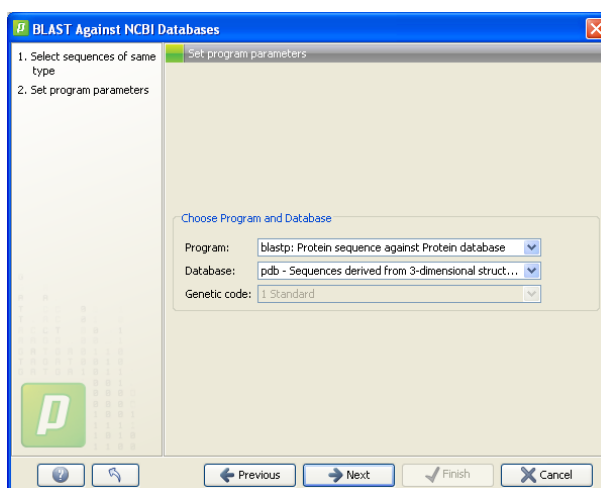


Figure 10.2: Choose a BLAST Program and a database for the search.

BLAST search for DNA sequences:

- **BLASTn: DNA sequence against DNA database** This BLAST method is used to identify homologous DNA sequences to your query sequence.
- **BLASTx: Translated DNA sequence against Protein database** If you want to search in protein databases, this BLAST method allows for automated translation of the DNA input sequence and searching in various protein databases.
- **tBLASTx: Translated DNA sequence against Translated DNA database** Here is both the input DNA sequence and the searched DNA database automatically translated.

BLAST search for protein sequences:

- **BLASTp: Protein sequence against Protein database** This the most common BLAST method used when searching for homologous protein sequences having a protein sequence as search input.

- **tBLASTn: Protein sequence against Translated DNA database** Here is the protein sequence searched against an automatically translated DNA database.

Depending on whether you choose a protein or a DNA sequence, a number of different databases can be searched. A complete list of these databases can be found in Appendix B. When **nr** appears in the **Database** parameter drop down menu, the search will include all relevant databases at NCBI. The **nr** database is the most complete, but also the most redundant database that can be searched. Searches can be limited to less complete databases. As an example, when choosing **pdb** only sequences with a known structure are searched. If homologous sequences are found to the query sequence, these can be downloaded and opened with the 3D viewer of *CLC Protein Workbench*.

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 is particularly useful when working with organisms or organelles which have a genetic code that differs from the standard genetic code.

In **Step 3** you can limit the BLAST search by adjusting the parameters seen in figure 10.3

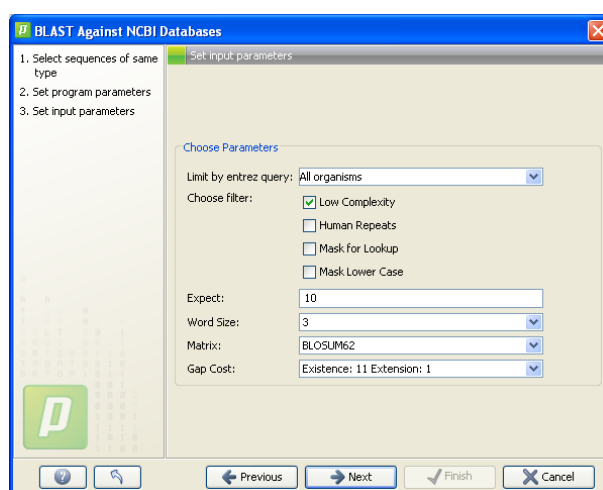


Figure 10.3: Examples of different limitations which can be set before submitting a BLAST search.

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 the BLAST databases. Any terms can be entered that would normally be allowed in an Entrez search session. Some queries are preentered and can be chosen in the drop down menu.
- **Choose filter**
 - 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.

- **Human Repeats.** This option masks Human repeats (LINE's and SINE's) and is especially useful for human sequences that may contain these repeats. Filtering for repeats can increase the speed of a search especially with very long sequences (>100 kb) and against databases which contain large number of repeats (htgs).
 - **Mask for Lookup.** This option masks only for purposes of constructing the lookup table used by BLAST. BLAST searches consist of two phases, finding hits based upon a lookup table and then extending them.
 - **Mask Lower Case.** With this option selected you can cut and paste a FASTA sequence in upper case characters and denote areas you would like filtered with lower case. This allows you to customize what is filtered from the sequence during the comparison to the BLAST databases
- **Expect** The statistical significance threshold for reporting matches against database sequences: the default value is 10, meaning that 10 matches are expected to be found merely by chance, according to the stochastic model of Karlin and Altschul (1990). If the statistical significance 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 shows less stringent matches. Fractional values are acceptable.
 - **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.
 - **Matrix** 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).
 - **Gap Cost** The pull down menu shows the Gap Costs (Penalty to open Gap and penalty to extend Gap). There can only be a limited number of options for these parameters. Increasing the Gap Costs and Lambda ratio will result in alignments which decrease the number of Gaps introduced.

The more limitations are submitted to the search parameters, the faster the search will be conducted. If no limitations are submitted, the BLAST search may take several minutes.

When the **Advanced parameters** of **Step 3** are adjusted, click **Next** to choose whether you want to open the BLAST output in an editor and/or in a table. [10.1.1](#)

Click **Next** if you wish to adjust how to handle the results (see section [8.1](#)). If not, click **Finish**.

10.1.1 Output from BLAST search

The two different outputs from a BLAST search are shown in figure [10.4](#):

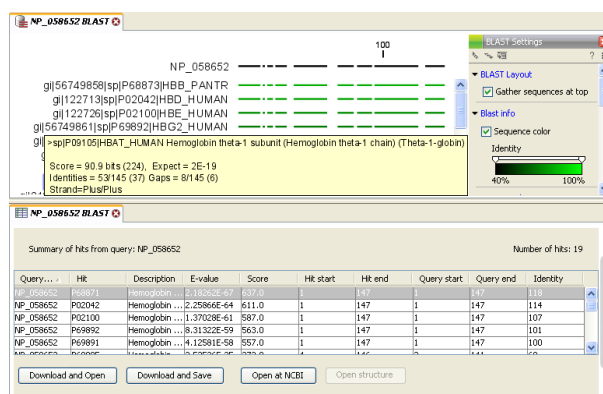


Figure 10.4: Display of the output of a BLAST search. At the top is there a graphical representation of BLAST hits with tool-tips showing additional information on individual hits. Below is shown a tabular form of the BLAST results.

The **BLAST Graphics** and the **BLAST table** are described in the following chapters.

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.

There are several View preferences available for in the **BLAST Graphics** view.

- **BLAST Layout** You can choose whether to **Gather sequences at top**, which means that vertical gaps between sequences are eliminated to assist comparison between the query sequence and the hit sequences.
- **BLAST info** In this View preference group, you can choose whether to color hit sequences and you can adjust the coloring.

The remaining View preferences for BLAST Graphics are the same as those of alignments. See section 17.2.

Some of the information available in the tooltips 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).
- **Length of sequence** This shows the entire length of the found sequence.
- **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.
- **Query** This is the sequence (or part of the sequence) which you have used for the BLAST search.
- **Subjct (subject)** This is the sequence found in the database.

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.

In addition to the latter described output of a BLAST search, 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. In the tabular view it is possible to select multiple sequences and for example download all of these in one single step. Moreover, it is possible to look additional information on each single hit is the BLAST result on the NCBI homepage. These possibilities are either available through a right-click with the mouse or by using the buttons at the end of the table.

10.1.2 BLAST table

If the **BLAST table** view was not selected in **Step 4** of the BLAST search, the table can be generated in the following way:

Right-click the tab of the initial BLAST result view | Show | BLAST Table

Figure 10.5 is an example of a BLAST Table.

Query	Hit	Description	E-value	Score	Hit start	Hit end	Query start	Query end	Identity
CAA26204	1DXT-D	Chain D, H...	7.65273E-67	629.0	1	125	1	121	121
CAA26204	1Y85-D	Chain D, T...	2.90803E-66	624.0	2	125	1	120	120
CAA26204	2DN3-B	Chain B, I...	2.90803E-66	624.0	2	125	1	120	120
CAA26204	1O1N-D	Chain D, D...	6.47842E-66	621.0	2	125	1	120	119
CAA26204	1Y83-D	Chain D, T...	6.47842E-66	621.0	2	125	1	120	119
CAA26204	1Y7F-B	Chain B, T...	8.46108E-66	620.0	2	125	1	120	119
CAA26204	1HD8-D	Chain D, A...	8.46108E-66	620.0	2	125	1	120	119

Figure 10.5: Display of the output of a BLAST search in the tabular view. The hits can be sorted by the different columns, simply by clicking the column heading.

The BLAST 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
- **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 bit score of the local alignment generated through the BLAST search.
- **Hit start** Shows the start position in the hit sequence

- **Hit end** Shows the end position in the hit sequence
- **Query start** Shows the start position in the query sequence
- **Query end** Shows the end position in the query sequence
- **Identity** Shows the number of identical residues in the query and hit sequence

In the **BLAST table** view you can handle the hit sequences. Select one or more sequences from the table, and apply one of the following functions.

- **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 sequence**
Opens the selected sequence(s) in one or more sequence views.
- **Save sequence**
Downloads and saves the sequence without opening it.
- **Open structure** If the hit sequence contain structure information, the sequence is opened in a text view or a 3D view (3D view in CLC Protein Workbench and CLC Combined Workbench).

10.2 BLAST Against Local Database

CLC Protein Workbench will let you conduct a BLAST search in a local database. See section 10.3 for more about how to create a database.

The advantage of conducting a local BLAST search is the speed and that it is possible to BLAST sequences longer than 8900 residues.

To conduct a Local BLAST search:

right-click the tab of an open sequence | Toolbox | BLAST Search() | BLAST Against Local Databases ()

or **click an element in the Navigation Area | Toolbox | BLAST Search() | BLAST Against Local Databases ()**

This opens the dialog seen in figure 10.6:

Click **Next**

This opens the dialog seen in figure 10.7:

In **Step 2**, you can choose between different BLAST methods. See section 10.1 for information about these methods.

In **step 2** you can also choose which of your local BLAST databases you want to conduct the search in. Clicking **Select Database** opens the dialog shown in figure 10.8:

Select a Click **Next**

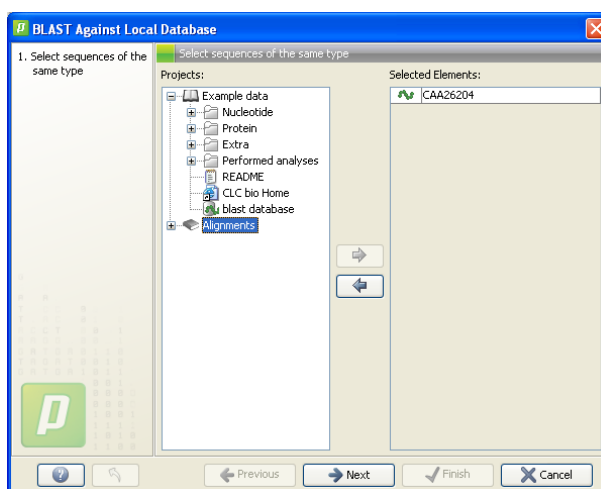


Figure 10.6: Choose one or more sequences to conduct a Local BLAST search.

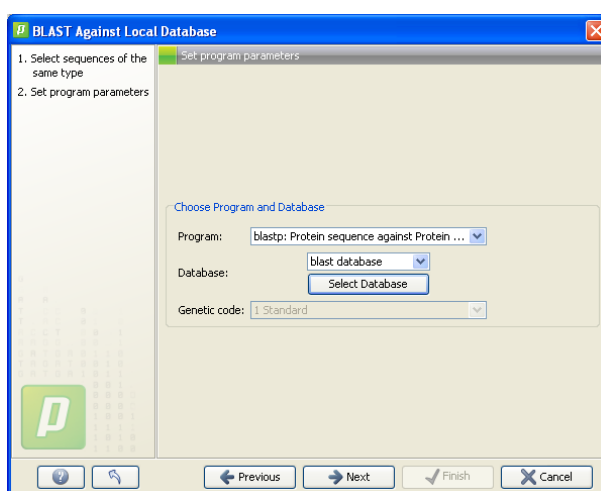


Figure 10.7: Choose a BLAST program and a local database to conduct BLAST search.

This opens the dialog seen in figure 10.9:

See section 10.1 for information about these limitations.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**.

10.3 Create Local BLAST Database

In *CLC Protein Workbench* you can create a local database which you can use for local BLAST. Both DNA, RNA, and protein sequences can be used.

It is not necessary to import the sequences into *CLC Protein Workbench* before creating the database. The local database can be created from sequences which are stored in the **Navigation Area** or the sequences can be browsed from the computer's file system. In the latter case, the files must be in fasta (.fsa/.fa/.fasta) format.

To create a local BLAST data base from the file system or from the **Navigation Area**:

BLAST search in Toolbox(📁) | Create Local BLAST Database(🛠️)

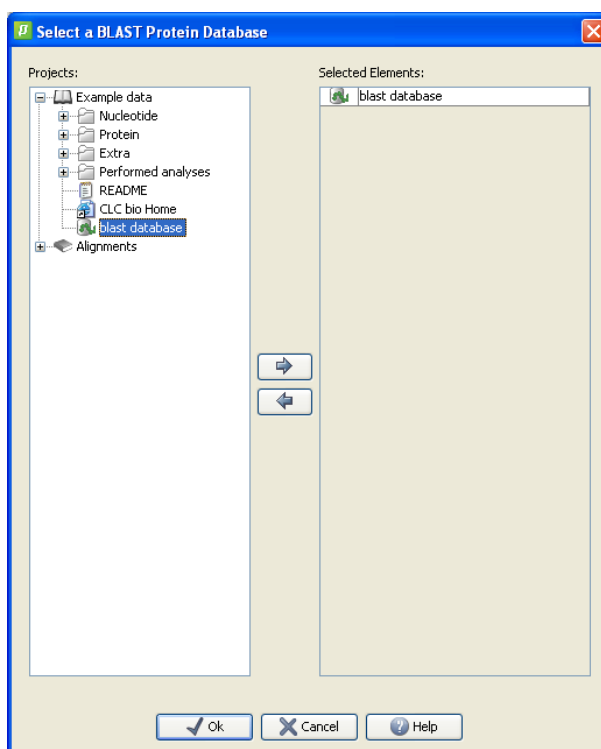


Figure 10.8: Select your local BLAST database.

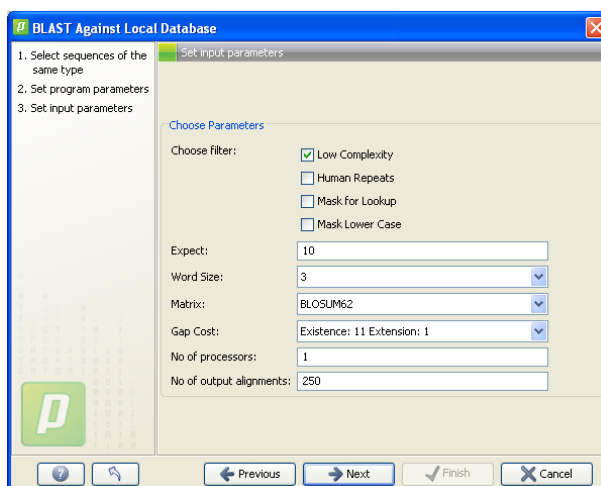


Figure 10.9: Examples of different limitations which can be set before submitting a BLAST search.

This opens the dialog seen in figure 10.10

- **Select Input Source.** Lets you choose whether to include sequences from the **Navigation Area** or from the computer's file system (External FASTA file).
- **Sequence type.** If you choose to import sequences from an external FASTA file into the database, you must choose whether the sequences are nucleotide or protein sequences.
- **FormatDB Option.** Enables or disables parsing of SeqId and creation of indices.
- **Input Sequences.** Depending on the choice of **Select Input Source** above, clicking the button will let you browse the **Navigation Area** or the external file system for the sequences

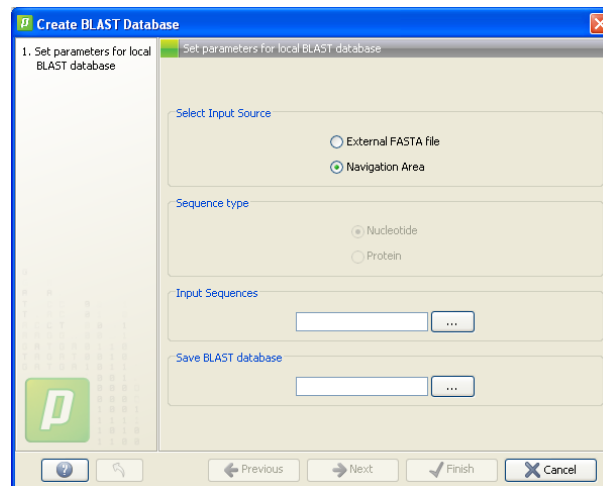


Figure 10.10: Setting parameters for the local BLAST database.

which you want to include in the database.

- **Save BLAST database.** Lets you browse your external file system for a suitable place to save the database.

After having adjusted all these settings, click **Next**, which opens the dialog seen in figure 10.11

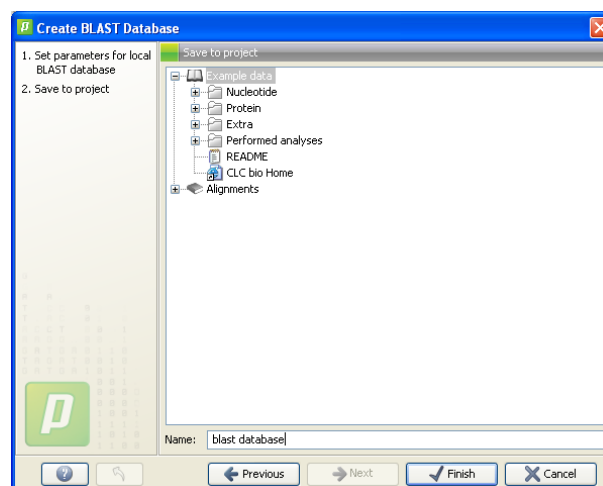


Figure 10.11: Choose where the access point to your local BLAST database is saved in the Navigation Area.

Click **Next** to complete the creation of the database.

Chapter 11

Viewing and editing sequences

Contents

11.1 View sequence	113
11.1.1 Sequence Layout in Side Panel	114
11.1.2 Selecting parts of the sequence	119
11.1.3 Editing the sequence	120
11.1.4 Adding and modifying annotations	120
11.1.5 Removing annotations	122
11.1.6 Sequence region types	122
11.2 Sequence information	123
11.2.1 Annotation map	124
11.3 View as text	124
11.4 Creating a new sequence	125
11.5 Sequence Lists	126
11.5.1 Graphical view of sequence lists	127
11.5.2 Sequence list table	128
11.5.3 Extract sequences	128
11.6 Circular DNA	128
11.6.1 Using split views to see details of the circular molecule	129
11.6.2 Mark molecule as circular and specify starting point	130

CLC Protein Workbench 2.0 offers three different ways of viewing and editing sequences as described in this chapter. Furthermore, this chapter also explains how to create a new sequence and how to assemble several sequences in a sequence list.

11.1 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 3.3 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 17.2).

11.1.1 Sequence Layout in Side Panel

Each view of a sequence has a **Side Panel** located at the right side of the 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 gray Side Panel button to the right to show**

When you open a view, the **Side Panel** has default settings which can be changed in the **User Preferences** (see chapter 4).

Below, each group of preferences 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 preferences.

Notice! 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:

- **Space every 10 residues.** Inserts a space every 10 residues - only visible when you zoom in to see the residues.
- **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 plus strand.** Whether to set the numbers relative to the positive or the negative strand in a nucleotide 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**.
- **Follow selection.** When viewing the same sequence in two separate views, "Follow selection" will automatically scroll the view in order to follow a selection made in the other view.
- **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).
 - Species.
 - Species (accession).
 - Common Species.
 - Common Species (accession).

Annotation Layout

Annotations are data attached to a specific part of a sequence. If the sequence is downloaded from a database it has annotations attached to it, e.g. the location of genes on a DNA sequence. If you have performed **Restriction Site** or **Proteolytic Cleavage** analysis, the cut sites can be displayed as annotations on the sequence. Other analyses also attach annotations on the sequence. See section 11.1.6 for more information about how to interpret the annotations. The annotations are shown as colored boxes along the sequence, and their appearance is determined in the **Annotation layout** preferences group:

- **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.
- **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.** Each annotation can be labelled with a name. 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.
- **Show arrows.** Toggles the display of arrow heads on the annotations.
- **Use gradients.** Fills the boxes with gradient color.

Annotation types

- **Annotation types.** This group lists all the types of annotations that are attached to the sequence that is viewed. For sequences with many annotations it can be easier to get an overview, if you deselect the annotation types that are not relevant. If you want to remove single annotations while preserving other annotations of the same type, see section [11.1.4](#).

It is possible to color the different annotations for better overview.

Color settings for an annotation can be done by clicking the colored square next to the relevant annotation type.

Many different settings can be set in the three layers: Swatches, HSB, and RGB. 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**.

Restriction sites

These preferences allow you to display restriction sites on the sequence. There is a list of enzymes which are represented by different colors. By selecting or deselecting the enzymes in the list, you can specify which enzymes' restriction sites should be displayed (see figure [11.1](#)).



Figure 11.1: Showing restriction sites of two restriction enzymes.

The color of the flag of the restriction site can be changed by clicking the colored box next to the enzyme's name.

The list of restriction enzymes contains per default ten of the most popular enzymes, but you can easily modify this list and add more enzymes. You have four ways of modifying the list:

- **Edit enzymes button.** This displays a dialog with the enzymes currently in the list shown at the bottom and a list of available enzymes at the top. To add more enzymes, select them in the upper list and press the **Add enzymes button** (↓). To remove enzymes, select them in the list below and click the **Remove enzymes button** (↑).
- **Load enzymes button.** If you have previously created an enzyme list, you can select this list by clicking the Load enzymes button. You can filter the enzymes in the same way as illustrated in figure [16.2](#).
- **Add enzymes cutting the selection to panel.** If you make a selection on the sequence, right-click, you find this option for adding enzymes. Based on the entire list of available enzymes, the enzymes cutting in the region you selected will be added to the list in the **Side Panel**.
- **Insert restriction site before/after selection.** If you make a selection on the sequence, right-click, you find this option for inserting a restriction site before or after the region you

selected. A dialog is shown where you can select an enzyme whose recognition sequence is inserted. If it was not already present in the list in the **Side Panel**, the enzyme will now be added and selected.

Finally, if you have selected a set of enzymes that you wish to keep for later use, you can click **Save enzymes** and the selected enzymes will be saved to an enzyme list. This list can then be used both when finding restriction sites from the **Toolbox** or when viewing another sequence.

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 polarity of amino acids.
 - **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.

Nucleotide info

These preferences only apply to nucleotide sequences.

- **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.
 - **Frame.** Determines where to start the translation.
 - * **+1 to -1.** Select one of the six reading frames.
 - * **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 [11.1.2](#).
 - * **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. For more about translation tables, see section 14.4.
- **Only AUG start codons.** For most genetic codes, a number of codons can be start codons. Selecting this option only colors the AUG codons green.
- **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.
 - * **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.

Hydrophobicity info

These preferences only apply to proteins and are described in section 15.5.2.

Search

The Search group is not a preferences group, but can be used for searching the sequence. Clicking the search button will search for the first occurrence of the search string. Clicking the search 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 to search for. The search function does not discriminate between lower and upper case characters.
- **Sequence search.** Search the nucleotides or amino acids. For nucleotides, all the standard IUPAC codes can be used, e.g. RT will find both GT and AT. RT will also find e.g. AN. The IUPAC codes are available from the **Help** menu under Background Information. For amino acids, the single letter abbreviations should be used for searching. Accordingly, N (for nucleotides) and X (for proteins) can be used as a wildcard character.

- **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.
- **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 and end number.
- **Include negative strand.** When searching the sequence for nucleotides or amino acids, you can search on both strands.

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 label and translations if relevant).

- **Text size.** Five different sizes.
- **Font.** Shows a list of Fonts available on your computer.
- **Bold residues.** Makes the residues bold.

11.1.2 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 search function described above.

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.

If you have made a selection, you can expand it by using **Shift** and **Ctrl** keys or by using the right-click menu:

right-click the selection | Expand Selection | Select the number of residues to expand the selection to both sides

To select the entire sequence:

right-click the sequence label to the left

To select a part of a sequence covered by an annotation:

right-click the annotation | Select annotation

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**.

The process described above is also the way to manually translate coding parts of sequences (CDS) into protein. You simply translate the new sequence into protein. This is done by:

right-click the tab of the new sequence | Toolbox | Nucleotide Analyses (📄) | Translate to Protein (🔍)

A selection can also be copied to the clipboard and pasted into another program:

make a selection | Ctrl + C (⌘ + C on Mac)

Notice! The annotations covering the selection will not be copied.

A selection of a sequence can be edited as described in the following section.

11.1.3 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

Another way to edit the sequence is by inserting a restriction site. See section [11.1.1](#).

11.1.4 Adding and modifying annotations

Most sequences carry different biological information. When retrieving sequences from various databases, the sequence often contains biological information by way of annotations. You can manually add annotations from a compiled annotation list. This list of annotations covers the most frequently used annotations in UniProt and GenBank. Annotations which have been added to a sequence can be removed at any time (see section [11.1.5](#)).

Annotations can be added to a sequence:

make a selection covering the part of the sequence you want to annotate | right-click the selection | Add Annotation

This will display a dialog like the one in figure [11.2](#).

The left-hand part of the dialog lists a number of **Annotation types**. When you have selected an annotation type, it appears in **Chosen type**. You can also select an annotation from the **Chosen type** list. Choosing an annotation type is mandatory.

The right-hand part of the dialog contains the following text fields:

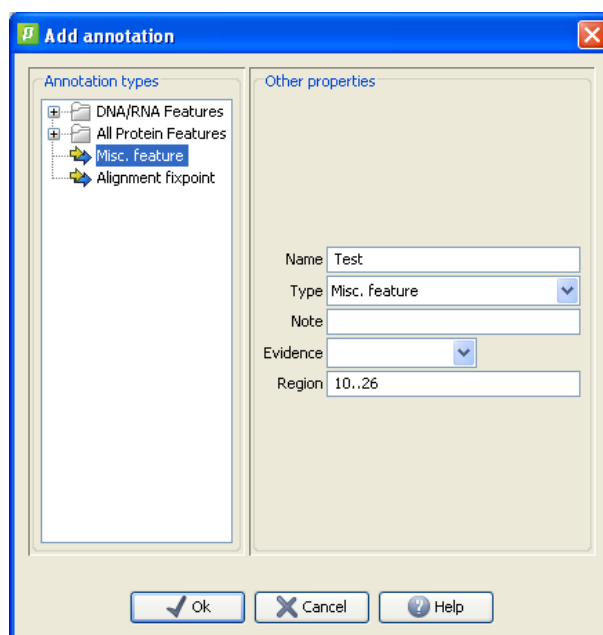


Figure 11.2: *The Add Annotation dialog.*

- **Name.** The name of the annotation which can be shown in the view. Whether the name is shown depends on the **Annotation Layout** preferences (see section 11.1.1).
- **Chosen type.** Reflects the left-hand part of the dialog as described above.
- **Note.** This is a field for entering notes about the annotation. The note will be displayed in a tooltip when you hold the mouse pointer over the sequence.
- **Evidence.** There are two options for the evidence supporting the annotation: experimental and non-experimental.
- **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 syntax of 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).

Click **OK** to add the annotation.

Notice! The annotation will be included if you export the sequence in GenBank, Swiss-Prot or CLC format.

To modify an existing annotation:

right-click the annotation | Edit Annotation

This will show the same dialog as in figure 11.2, with the exception that some of the fields are filled out depending on how much information the annotation contains.

11.1.5 Removing annotations

Annotations can be hidden using the **Annotation Types** preferences in the **Side Panel** to the right of the view (see section 11.1.1). 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 Annotations of This Type

If you want to remove all annotations from a sequence:

right-click an annotation | Delete All Annotations

The removal of annotations can be undone using Ctrl + Z or Undo () in the Toolbar.

11.1.6 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 11.3 is an example of three regions with separate colors.

Figure 11.4 shows an artificial sequence with all the different kinds of regions.

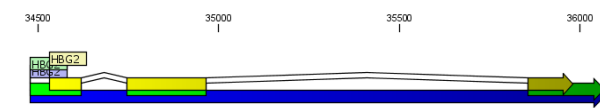


Figure 11.3: Three regions on a human beta globin DNA sequence (HUMHBB).

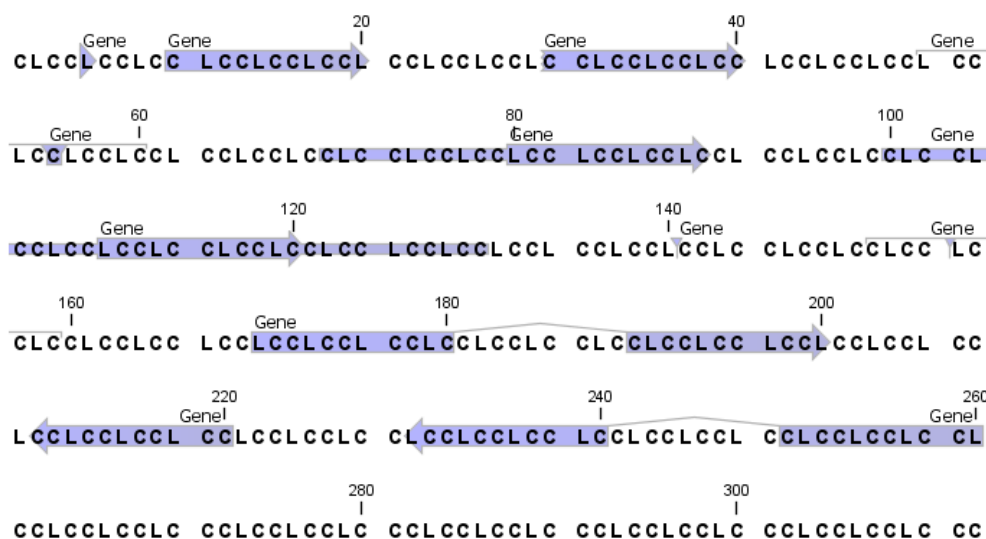


Figure 11.4: Region #1: A single residue, Region #2: A range of residues including both endpoints, Region #3: A range of residues starting somewhere before 30 and continuing up to and including 40, Region #4: A single residue somewhere between 50 and 60 inclusive, Region #5: A range of residues beginning 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.

11.2 Sequence 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 **Sequence info** function which also displays a textual overview of the annotations.

To view the sequence information:

select a sequence in the Navigation Area | Show () in the Toolbar | Sequence info ()

This will display a view similar to fig 11.5.

All the lines in the view are headings, and the corresponding text can be shown by clicking the text. The information available depends on the origin of the sequence. If the sequence is annotated, the annotations can be found under the heading **Annotation map**.

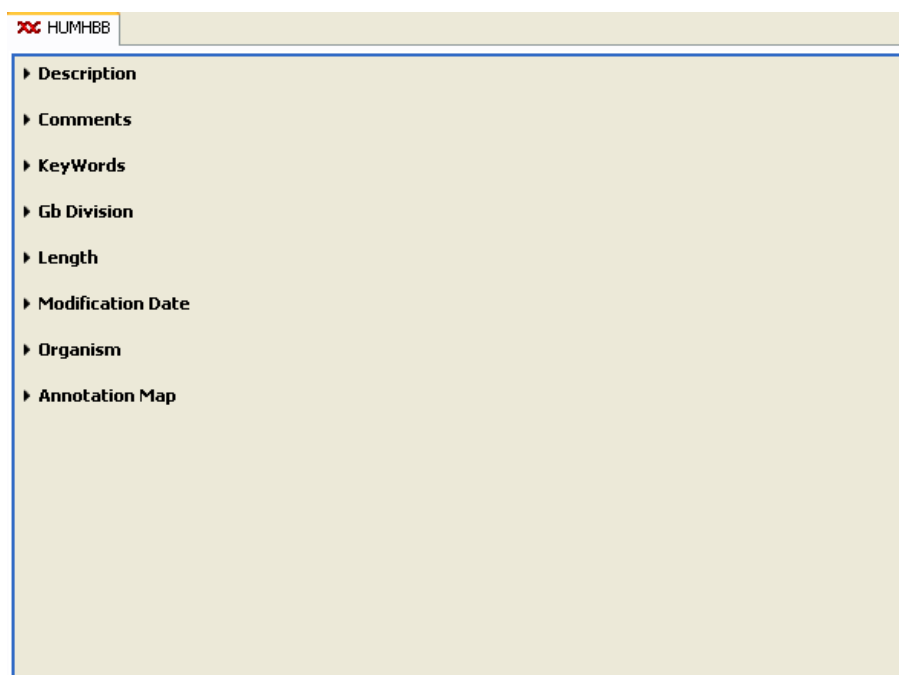


Figure 11.5: *The initial display of sequence info for the HUMHBB DNA sequence from the Example data.*

11.2.1 Annotation map

The **Annotation map** displays the various types of annotations that are attached to the sequence. Clicking on the name of a type of annotation will list the annotations of this type. If there are more annotations of the same kind, the blue arrows can be used to move up and down in the annotations of that type. In order to use the links, you have to open a second view of the sequence (double-click the sequence in the **Navigation Area**). If you have this view open, clicking one of the annotations in the **Annotation map** will make a selection in the other view corresponding to the annotation (see fig 11.6).

Annotations cannot be added or modified using the **Sequence info**. For adding and modifying annotations see section 11.1.4.

11.3 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 | Show in the Toolbar | As text

This way it is 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 Viewer** 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 11.2.)

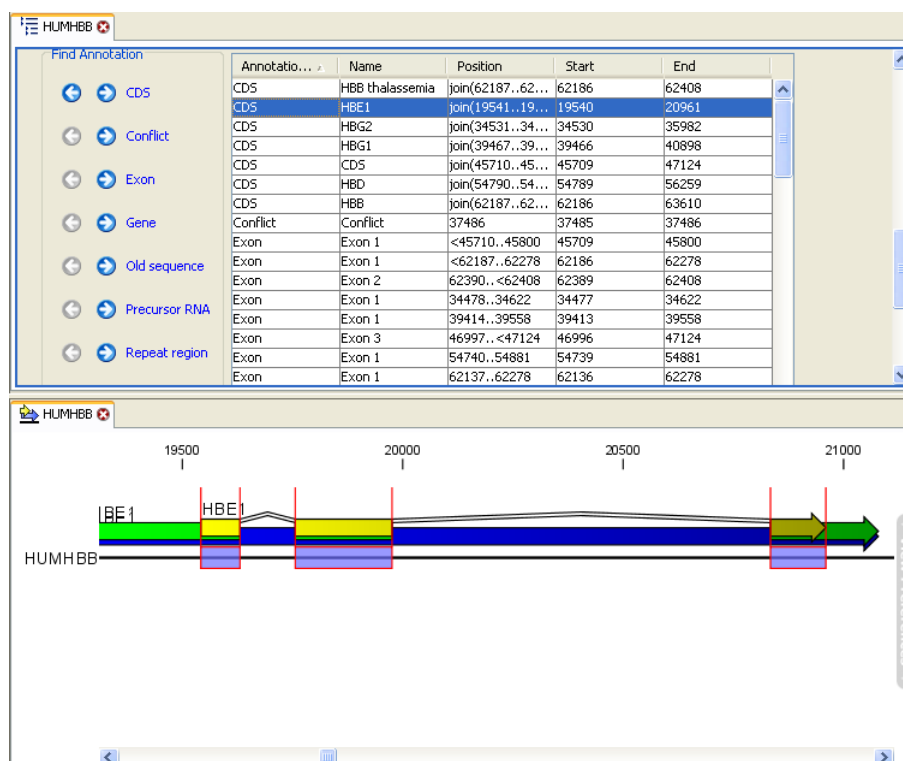


Figure 11.6: Clicking a sequence map annotation in the sequence information view, selects the annotation on the (normal) sequence view.

11.4 Creating a new sequence

A sequence can either be imported, downloaded from an online database or created in the *CLC Protein Workbench 2.0*. This section explains how to create a new sequence:

New(+) in the toolbar

Figure 11.7: Creating a sequence.

The **Create Sequence** dialog (figure 11.7) reflects the information needed in the GenBank format, but you are free to enter anything into the fields. The following description is a guideline for entering information about a sequence:

- **Name.** The name of the sequence. This is used for saving the sequence.
- **Common name.** A common name for the species.
- **Species.** The Latin name.
- **Type.** Select between DNA, RNA and protein.
- **Circular.** Specifies whether the sequence is circular. This will open the sequence in a circular view as default. (applies only to nucleotide sequences).
- **Description.** A description of the sequence.
- **Keywords.** A set of keywords separated by semicolons (;).
- **Comments.** Your own comments to the sequence.
- **Sequence.** Depending on the type chosen, this field accepts nucleotides or amino acids. Spaces and numbers can be entered, but they are ignored when the sequence is created. This allows you to paste in a sequence directly from a different source, even if the residue numbers are included. Characters that are not part of the IUPAC codes cannot be entered. At the top right corner of the field, the number of residues are counted. The counter does not count spaces or numbers.

Clicking Next will allow you to save the sequence to a project in the **Navigation Area**.

11.5 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. The sequence list may originate from an NCBI search (chapter 9.1). Moreover, if a multiple sequence fasta file is imported, it is possible to store the data in a sequences list. A **Sequence List** can also be generated using a dialog, which is described here:

select two or more sequences | right-click the elements | New | Sequence List 

This action opens 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.

After clicking "Next", you can choose where to save the list. Then click **Finish**.

Opening a Sequence list is done by:

right-click the sequence list in the Navigation Area | Show | click Graphical sequence list OR click Table

The two different views of the same sequence list are shown in split screen in figure 11.9.

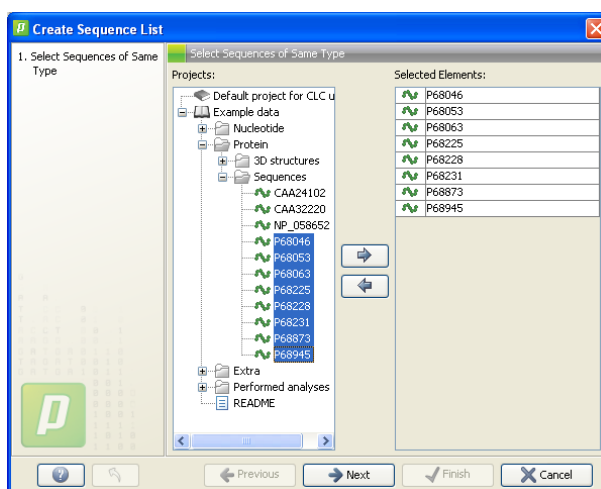


Figure 11.8: A Sequence List dialog.

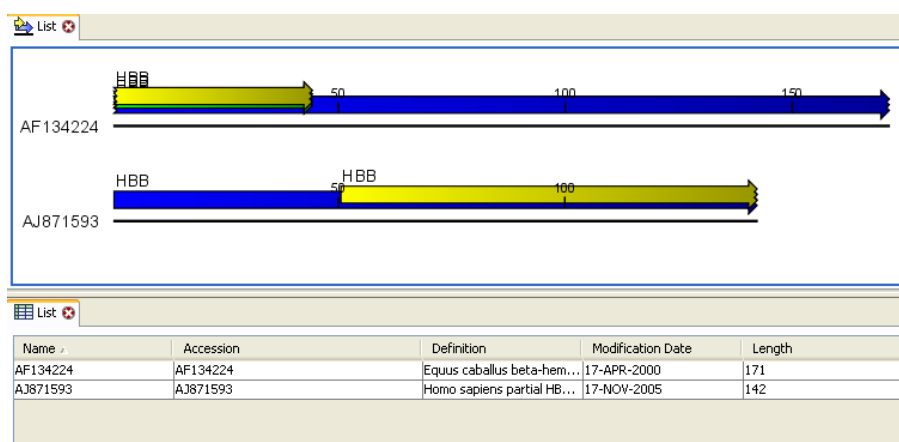


Figure 11.9: A sequence list of two sequences can be viewed in either a table or in a graphical sequence list.

11.5.1 Graphical view of sequence lists

The graphical view of sequence lists is almost identical to the view of single sequences (see section 11.1). 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 label and select **Delete Sequence**.
- To sort the sequences in the list, right-click the label 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 label of the sequence and select **Rename Sequence**.

11.5.2 Sequence list table

Each sequence in the table sequence list is displayed with:

- Name.
- Accession.
- Definition.
- Modification date.
- Length.

In the View preferences for the table view of the sequence list, columns can be excluded, and the view preferences can be saved in a style sheet. See section 4.5.

The sequences can be sorted by clicking the column headings. You can further refine the sorting by pressing Ctrl while clicking the heading of another column.

11.5.3 Extract sequences

It is possible to extract individual sequences from a sequence list in two ways. If the sequence list is opened in the tabular view, it is possible to drag (with the mouse) one or more sequences 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 to a preferred location in the **Navigation Area**:

right-click a sequence list in the Navigation Area | Extract Sequences

Select a location for the sequences and click OK. Copies of all the sequences in the list are now placed in the location you selected.

11.6 Circular DNA

A sequence can be shown as a circular molecule:

select a sequence in the Navigation Area | Show in the Toolbar | Circular()

This will open a view of the molecule similar to the one in figure 11.10.

This view of the sequence shares some of the properties of the linear view of sequences as described in section 11.1, but there are some differences. The similarities and differences are listed below:

- **Similarities:**
 - **Annotation Layout**, **Annotation Types** and **Text Format** preferences groups.
- **Differences:**
 - In the **Sequence Layout** preferences, only the following options are available in the circular view: **Ticks 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 (see below).

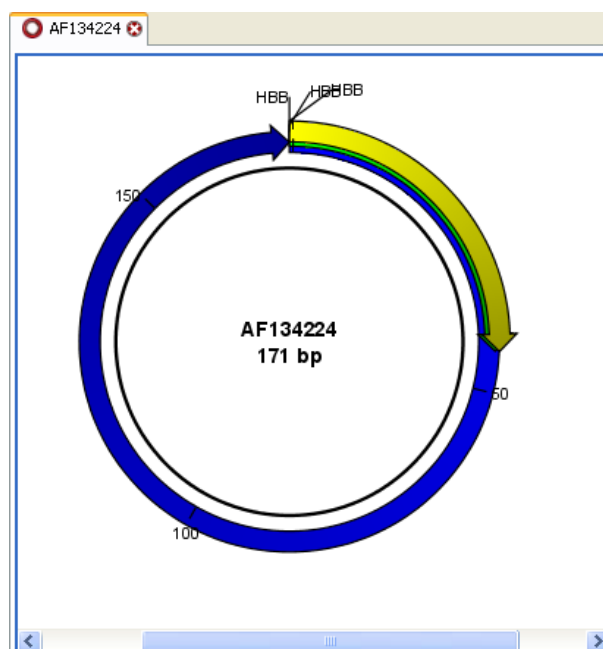


Figure 11.10: A molecule shown in a circular view.

11.6.1 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:

right-click the tab of the circular view of the sequence | Show | Sequence(👉👈)

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 11.11.

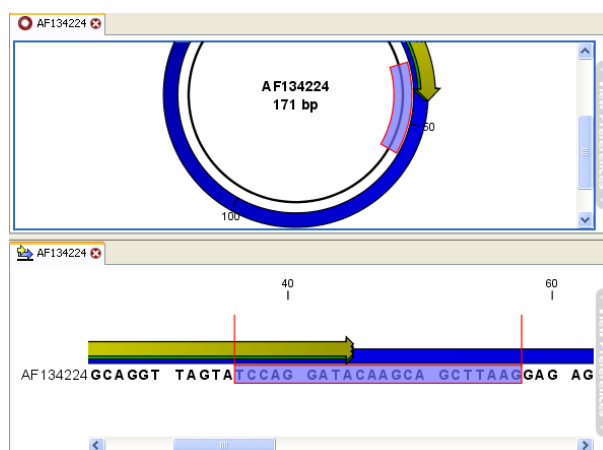


Figure 11.11: Two views showing the same sequence. The bottom view is zoomed in.

Notice! 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.

11.6.2 Mark molecule as circular and specify starting point

You can mark a DNA molecule as circular by right-clicking its label 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 ».

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

Notice! This can only be done for sequence that have been marked as circular.

Chapter 12

3D molecule viewing

Contents

12.1 Importing structure files	131
12.2 Viewing structure files	132
12.2.1 Moving and rotating	132
12.3 The structure table	133
12.3.1 Identification	133
12.3.2 Opening sequence information	133
12.3.3 Display and coloring options	134
12.4 Options through the preference panel	134
12.4.1 Atoms & Bonds	134
12.4.2 Backbone	134
12.4.3 Coloring	134
12.4.4 General settings	135
12.4.5 Selection scheme	136
12.5 3D Output	136

In order to understand protein function it is often valuable to see the actual three-dimensional structure of the protein. This is of course only possible if the structure of the protein has been resolved and published. *CLC Protein Workbench 2.0* has an integrated viewer of structure files. Structure files are usually deposited at the Protein DataBank (PDB) www.rcsb.org, where protein structure files can be searched and downloaded.

12.1 Importing structure files

In order to view the three-dimensional structure files there are different ways to import these. The supported file formats are PDB and mmCIF which both can be downloaded from the Protein DataBank (<http://www.rcsb.org>) and imported through the import menu. (see section 6.1.1.

Another way to import structure files is if a structure file is found either through a direct search at GenBank or by a BLAST search towards the PDB database. In the latter case, structure files can be directly downloaded to the navigation area by clicking the **download structure** button below all the BLAST hits. Downloading structure files from a conducted BLAST search is only possible

if the results are shown in a BLAST table. (See figure 12.1). How to conduct a BLAST search can be seen in section 10.1.

Summary of hits from query: CAA26204 Number of hits: 103

Query...	Hit	Descript...	E-value	Score	Hit start	Hit end	Query s...	Query end	Identity
CAA26204	1DXT-D	Chain D, H...	7.65273E-67	629.0	1	125	1	121	121
CAA26204	1Y85-D	Chain D, T...	2.90803E-66	624.0	2	125	1	120	120
CAA26204	2DN3-B	Chain B, I...	2.90803E-66	624.0	2	125	1	120	120
CAA26204	1O1N-D	Chain D, D...	6.47842E-66	621.0	2	125	1	120	119
CAA26204	1Y83-D	Chain D, T...	6.47842E-66	621.0	2	125	1	120	119
CAA26204	1YVT-B	Chain B, T...	8.46108E-66	620.0	2	125	1	120	119
CAA26204	1HDB-D	Chain D, A...	8.46108E-66	620.0	2	125	1	120	119
CAA26204	1MDD-D	Chain D, C...	8.46108E-66	620.0	2	125	1	120	119

Buttons: Download and Open, Download and Save, Open at NCBI, Open structure

Figure 12.1: It is possible to open a structure file directly from the output of a conducted BLAST search by clicking the Open Structure button.

12.2 Viewing structure files

The usual view area is used to display the actual structure (See figure 12.2 for an example of the structure view). At the bottom of the view area you will find a table displaying the polymer subunits of the structure along with additional compounds and in some cases water molecules. It is possible to copy polymer sequence information to the navigator area for further sequence analysis by the integrated workbench tools. To view the contents of a polymer subunit, right-click on the relevant table row and select **Open Sequence**. The newly opened view can be dragged onto the navigation area for further analysis.

Structures can be rotated and moved using the mouse and keyboard. **Pan mode** (🖱️) must be enabled in order to rotate and move the sequence. When changing to the 3D view a dialog box with the option of shifting to **pan mode** is displayed if **Selection** mode is enabled.

Notice! It is only possible to view one structure file at a time, in order to limit the amount of memory used.

12.2.1 Moving and rotating

Structure files are simply rotated by holding down the left mouse button while moving the mouse. This will rotate the structure in the direction the mouse is moved. The structures can be freely rotated in all directions.

Holding down the Ctrl key on the keyboard while dragging the mouse moves the structure in the direction the mouse is moved. This is particularly useful if the view is zoomed to cover only a small region of the protein structure.

Zoom in (🔍) and zoom out (🔍) on the structure is done by selecting the appropriate zoom tool in the toolbar and clicking with the mouse on the view area. The view can be restored to display the entire structure by clicking the **fit with** (📐) button on the toolbar.

- **Rotate mode**

The structure is rotated when the "Pan mode" (🖱️) is selected in the toolbar. If the "pan mode" is not enabled on the first view of a structure a warning is shown.

- **Zoom mode**

Use the zoom buttons on the toolbar to enable zoom mode. A single click with the mouse will zoom slightly on the structure. Moreover, it is possible to zoom in and out on the structure by keeping the left mouse button pressed while moving the mouse up and down.

- **Move mode**

It is possible to move the structure from side to side if the Ctrl key on the keyboard is pressed while dragging with the mouse.

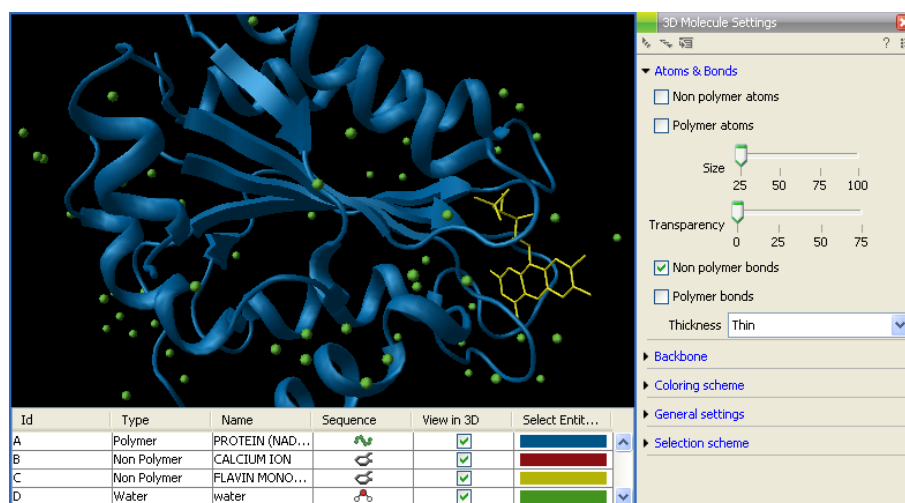


Figure 12.2: 3D view. Structure files can be opened, viewed and edited in several ways.

12.3 The structure table

Below the 3D image you will find a table presenting information on the protein subunits along with any compounds complexed with the protein in the resolved structure.

12.3.1 Identification

ID specifies an identifier for the subunit or compound as specified by the PDF or mmCIF record, while **Type** specifies the nature of the compound in question. Protein chains and RNA/DNA chains are specified as **Polymers**, while all other molecules, including water, are specified as **Non-Polymers**. The **Name** of the compound is also displayed as specified by the PDB or mmCIF record. The **ID** is appended to the structure identifier when opening sequence information (see below).

12.3.2 Opening sequence information

Only **Polymer** sequences may be opened in a sequence view. This is accomplished by rightclicking the appropriate table element and selecting **Open Sequence**. Editing a child sequence directly is not allowed in order to preserve consistency between the displayed 3D structure and the sequence. The sequence may however be copied to the navigation area by dragging the tab, after which editing is allowed. Changes to the copy are not reflected in the original child sequence.

The child sequence is named according to the parent structure, with the ID of the subunit appended. For example, the A chain of the structure with the ID 1A00 will be named 1A00-A. Brackets around the child name indicate the child-parent relationship.

12.3.3 Display and coloring options

Individual subunits, polymer as well as non-polymer, may be switched on and off in the 3D view using the **View in 3D** checkbox. Also, when using the Entity coloring mode (see below), the colors of individual subunits may be specified by the user using the **Select Entity Color** color choosers.

12.4 Options through the preference panel

The view of the structure can be changed in several ways. All graphical changes are carried out through the **Side Panel**. All options in the **Side Panel** are described below.

12.4.1 Atoms & Bonds

- **Non Polymer Atoms**

Show the individual atoms of non-polymer molecules as ball shaped structures. Atom size and transparency can be varied by using the sliders (see figure 12.2).

- **Polymer Atoms**

Show the individual atoms of the protein chain as ball shaped structures. Atom sizes and transparency can be varied by using the sliders (see figure 12.2).

- **Non Polymer Bonds**

Show bonds between atoms in non-polymer compounds. The width of the bond can be selected from the drop-down box.

- **Polymer Bonds**

Show bonds between polymer atoms. The width of the bond can be selected from the drop-down box.

12.4.2 Backbone

- **None**

The structure is displayed without any special indication of the backbone.

- **Cartoon**

Show the backbone on proteins as cartoon drawings. When using this view it is possible to see alpha helices and beta-sheets.

- **Backbone**

The alphacarbon atoms are connected by thick bonds.

12.4.3 Coloring

Atoms, bonds and cartoon elements are colored individually according to the list below. For the Atom Type scheme, the coloring scheme (CPK) is adapted from the visualization tool Rasmol.

- **Atom type**

Color the atoms individually.

- **Carbon:** Light grey
- **Oxygen:** Red
- **Hydrogen:** White
- **Nitrogen:** Light blue
- **Sulphur:** Yellow
- **Chlorine, Boron:** Green
- **Phosphorus, Iron, Barium:** Orange
- **Sodium:** Blue
- **Magnesium:** Forest green
- **Zn, Cu, Ni, Br:** Brown
- **Ca, Mn, Al, Ti, Cr, Ag:** Dark grey
- **F, Si, Au:** Goldenrod
- **Iodine:** Purple
- **Lithium:** firebrick
- **Helium:** Pink
- **Other:** Deep pink

- **Entities**

This will color protein subunits and additional structures individually. Using the view table, the user may select which colors are used to color subunits.

- **Rainbow**

This color mode will color the structure with rainbow colors along the sequence.

- **Secondary structure**

The structure is colored according to secondary structures. Alpha helices are colored light blue, while beta sheets are colored light green. All other atoms are colored grey.

12.4.4 General settings

- **Quality**

You may specify the image quality by using the dropdown list. Lower quality images render faster, but may not display well under high zoom factors.

- **Show table**

The table containing sequence information etc. may be turned off using this checkbox.

- **Show table**

The background color may be changed using this color chooser. Default color is black.

12.4.5 Selection scheme

When a polymer sequence from a structure is opened, selections made on the sequence will be mirrored by the atoms of the structure. The selection scheme specifies how atoms are highlighted.


- **Highlight**

Atoms retain their original color regardless of coloring scheme, but become more luminescent.

- **Inverse Transparency**

Nonselected atoms are rendered transparent while highlighted atoms will retain their original appearance. This scheme is useful for large, complex molecules, or for selections deep within the molecule. Note that the transparency slider is not functional when this scheme is set.

12.5 3D Output

The output of the 3D viewer is rendered on the screen in real time and changes to the preferences are visible immediately. From *CLC Protein Workbench 2.0* you can export the visible part of the 3D view to different graphic formats, by pressing the **Graphics** button () on the **Menu bar**. This will allow you to export in the following formats:

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

Printing is not fully implemented with the 3D editor. Should you wish to print a 3D view, this can be done by:

Windows:

- Adjust your 3D view in *CLC Protein Workbench*
- Press **Print Screen** on your keyboard (or **Alt + Print Screen**)
- Paste the result into an 'image editor' e.g. Paint or GIMP <http://www.gimp.org/>
- Crop (edit the screenshot)
- Save in your preferred file format and/or print

Mac:

- Set up your 3D view

- Press **⌘ + shift + 3** (or **⌘ + shift + 4**) (to take screen shot)
- Open the saved file (.pdf or .png) in a 'image editor' e.g. GIMP <http://www.gimp.org/>
- Crop (edit the screenshot)
- Save in your preferred file format and/or print

Linux:

- Set up your 3D view
- e.g. use GIMP to take the screen shot <http://www.gimp.org/>
- Crop (edit the screenshot)
- Save in your preferred file format and/or print

Chapter 13

General sequence analyses

Contents

13.1 Dot plots	138
13.1.1 Create dot plots	139
13.1.2 View dot plots	140
13.1.3 Bioinformatics explained: Dot plots	141
13.1.4 Bioinformatics explained: Scoring matrices	144
13.2 Shuffle sequence	148
13.3 Local complexity plot	149
13.3.1 Local complexity view preferences	150
13.4 Sequence statistics	151
13.4.1 Sequence statistics output	154
13.4.2 Bioinformatics explained: Protein statistics	154
13.5 Join sequences	158
13.6 Motif Search	159
13.6.1 Motif search parameter settings	161
13.6.2 Motif search output	162
13.7 Pattern Discovery	162
13.7.1 Pattern discovery search parameters	163
13.7.2 Pattern search output	164

CLC Protein Workbench 2.0 offers different kinds of sequence analyses, which apply to both protein and DNA. The analyses are described in this chapter.

13.1 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 adjust the view of the plot.

13.1.1 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, can various substitution matrices be applied in order to take the evolutionary distance of the two sequences into account.

To create a dot plot:

Toolbox | General Sequence Analyses (📁) | Create Dot Plot(🔍)

or **Select one or two sequences in the Navigation Area | Toolbox in the Menu Bar | General Sequence Analyses (📁) | Create Dot Plot (🔍)**

or **Select one or two sequences in the Navigation Area | right-click in the Navigation Area | Toolbox | General Sequence Analyses (📁) | Create Dot Plot(🔍)**

This opens the dialog shown in figure 13.1.

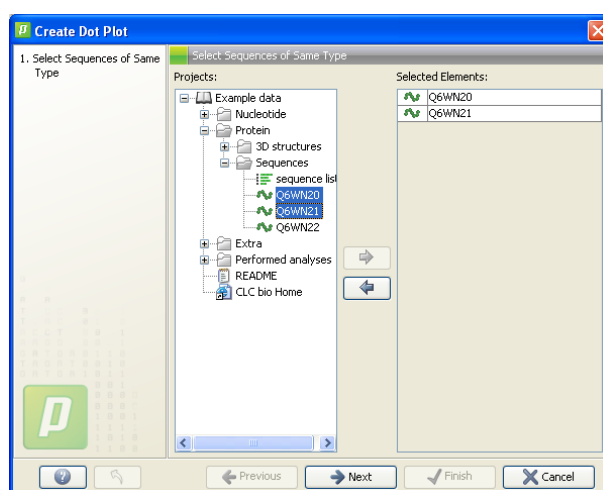


Figure 13.1: Selecting sequences for the dot plot.

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 **Project Tree**. Click **Next** to adjust dot plot parameters. Clicking **Next** opens the dialog shown in figure 13.2.

Notice that calculating dot plots take up a considerable amount of memory in the computer. Therefore, you see a warning 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, allowing you to save your work first. To avoid the Workbench shutting down you may choose to adjust the memory allocation to *CLC Protein Workbench*.

See section 1.8

Adjust dot plot parameters

There are two parameters for calculating 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 8.1). If not, click **Finish**.

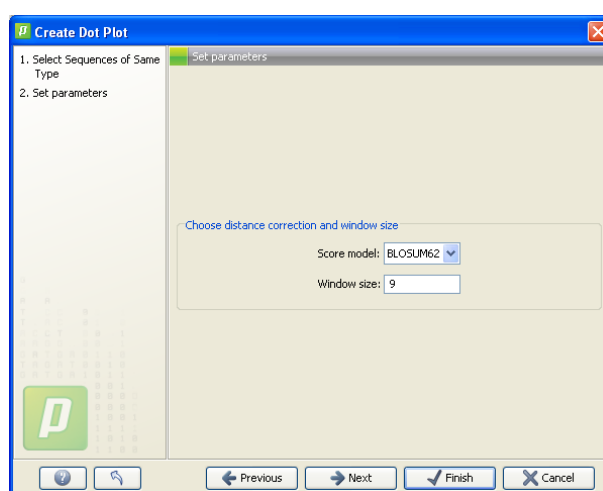


Figure 13.2: Setting the dot plot parameters.

13.1.2 View dot plots

A view of a dot plot can be seen in figure 13.3. You can select **Zoom in** (🔍) in the Toolbar and click the dot plot to zoom in to see the details of particular areas.

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). See figure 13.3.

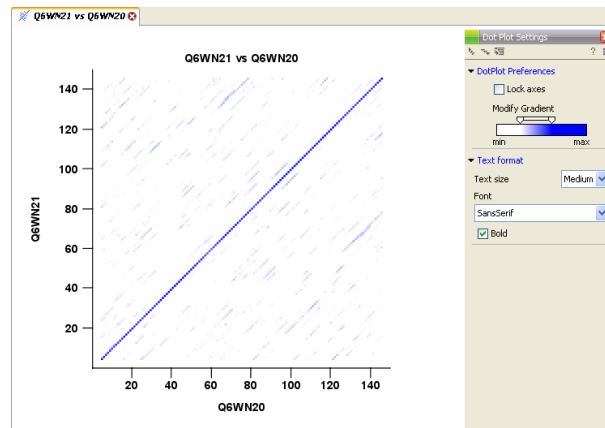


Figure 13.3: A view is opened showing the dot plot.

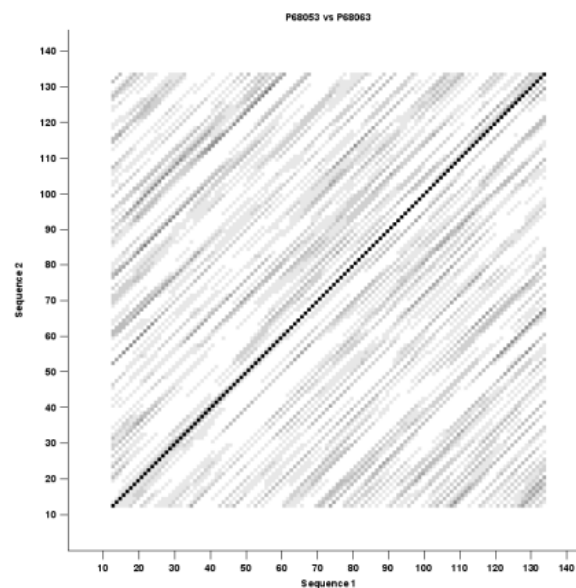


Figure 13.4: Dot plot with inverted colors, practical for printing.

13.1.3 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
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 13.5 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 lines parallel to the diagonal line.

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 13.7 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 13.8. In this figure, three frame shifts for the sequence on the y-axis are found.

1. Deletion of nucleotides
2. Insertion of nucleotides
3. Mutation (out of frame)

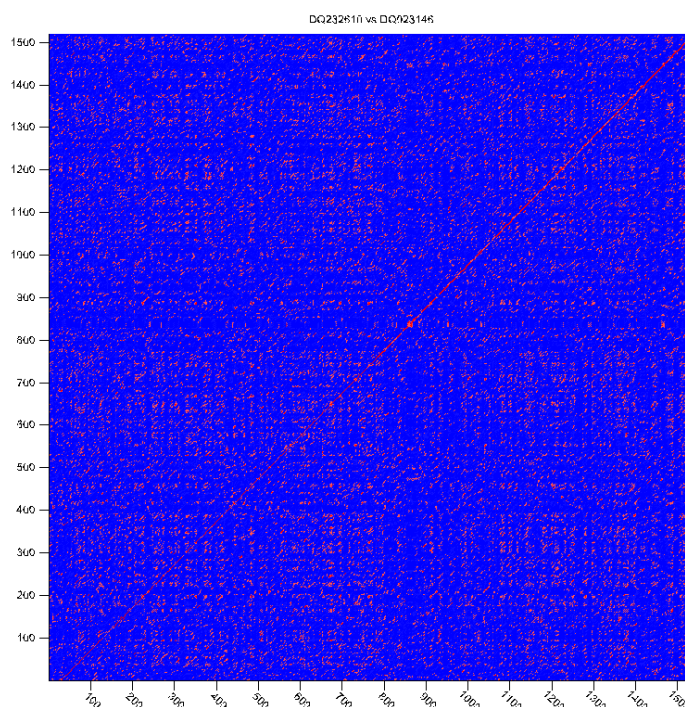


Figure 13.5: Dot plot of DQ232610 vs. DQ23146 (Influenza A virus nucleoproteins) showing and overall similarity



Figure 13.6: Direct and inverted repeats shown on an amino acid sequence generated for demonstration purposes.

Sequence inversions

In dot plots you can see an inversion of sequence as contrary diagonal to the diagonal showing similarity. In figure 13.9 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 13.10 is a square shows the low-complexity region of this sequence.

Creative Commons License

All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in it's original form and

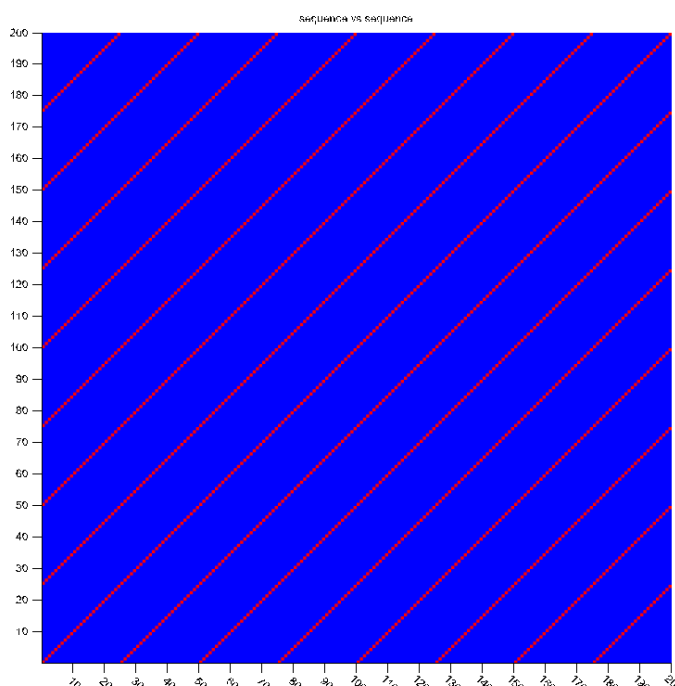


Figure 13.7: The dot plot of a sequence showing repeated elements. See also figure 13.6.

"CLC bio" has to be clearly labelled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, or build upon this work.



See <http://creativecommons.org/licenses/by-nc-nd/2.5/> for more about how you may use the contents.

13.1.4 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, 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

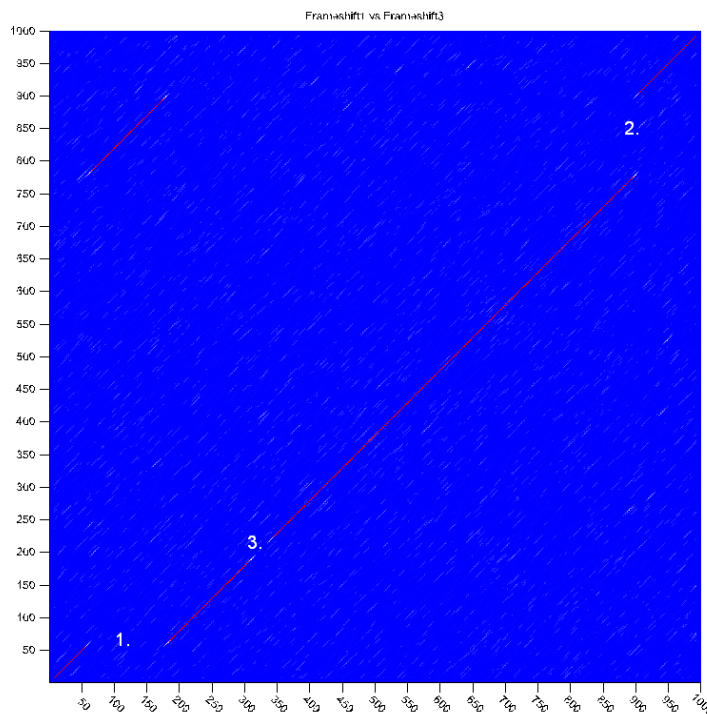


Figure 13.8: This dot plot show various frame shifts in the sequence. See text for details.

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 PAM256 matrix, gives the probability of 256 mutations in a 100 amino acids (see figure 13.11).

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

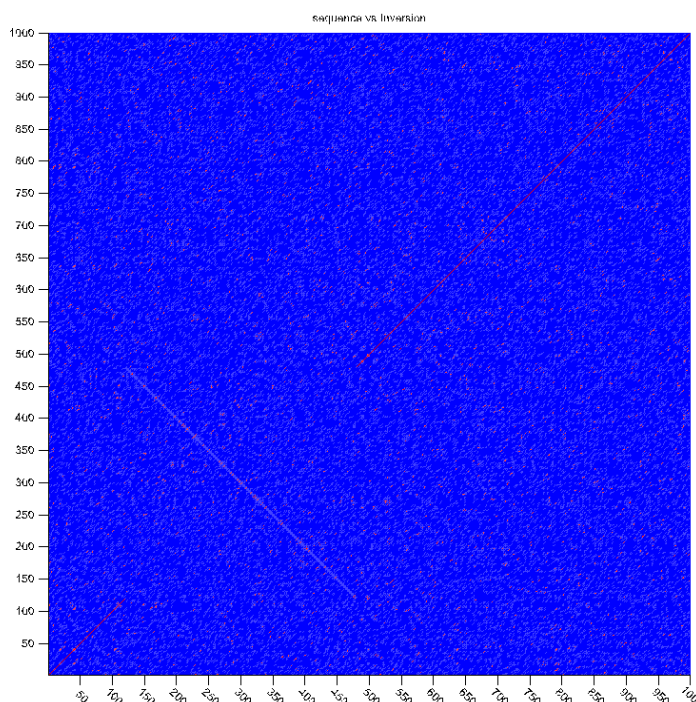


Figure 13.9: *The dot plot showing a inversion in a sequence. See also figure 13.6.*

sequences where none of the aligned sequences share less than 62% identity. This resulted in a scoring matrix called 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 to 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 probably 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 13.11) for correlations between the PAM and BLOSUM matrices. To summarize, if you

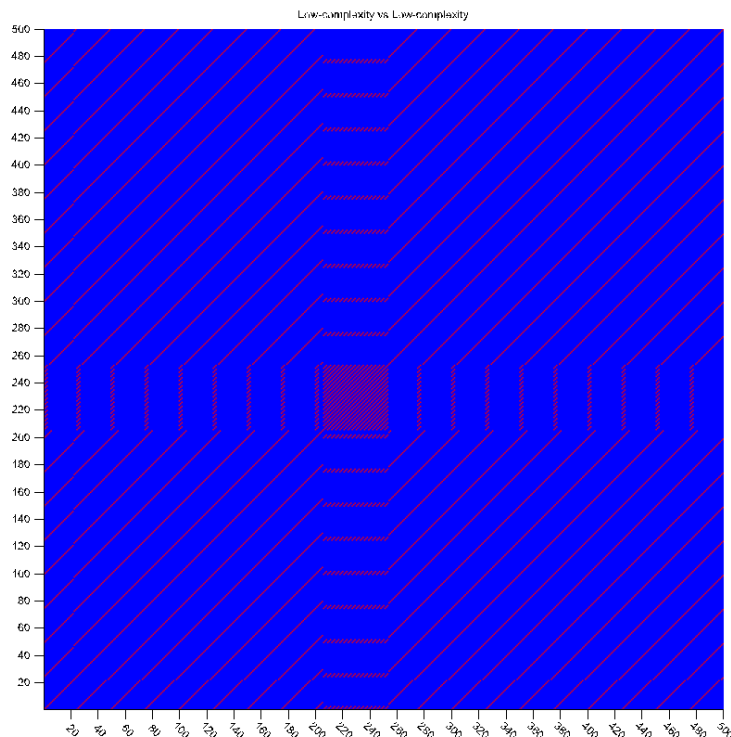


Figure 13.10: The dot plot showing a low-complexity region in the sequence. The sequence is artificial and low complexity regions does not always show as a square.

want to find distant related proteins to a sequence of interest using BLAST, you could benefit of using BLOSUM45 or similar matrices.

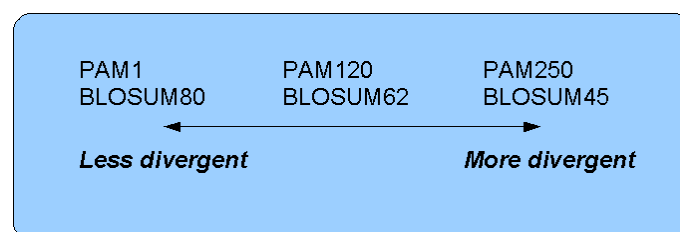


Figure 13.11: Relationship between scoring matrices. The BLOSUM62 has become a de facto standard scoring matrix for a wide range of alignment programs. It is the default matrix in BLAST.

Other useful resources

Calculate your own PAM matrix

<http://www.bioinformatics.nl/tools/pam.html>

BLOKS database

<http://blocks.fhcrc.org/>

NCBI help site

<http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Scoring2.html>

	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].

Creative Commons License

All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in it's original form and "CLC bio" has to be clearly labelled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, or build upon this work.



See <http://creativecommons.org/licenses/by-nc-nd/2.5/> for more about how you may use the contents.

13.2 Shuffle sequence

In some cases, it is beneficial to shuffle a sequence. This is an option in the **Toolbox** menu under **General Sequence Analyses**. It is normally used for statistical analyses, e.g. when comparing an alignment score with the distribution of scores of shuffled sequences. The shuffling is done without replacement, resulting in exactly the same number of the different residues as before the shuffling.

Shuffling a sequence removes all annotations that relate to the residues.

select sequence | **Toolbox in the Menu Bar** | **General Sequence Analyses** (🔧) | **Shuffle Sequence** (🔀)

or **right-click a sequence** | **Toolbox** | **General Sequence Analyses** (📁) | **Shuffle Sequence** (🔀)

This opens the dialog displayed in figure 13.12:

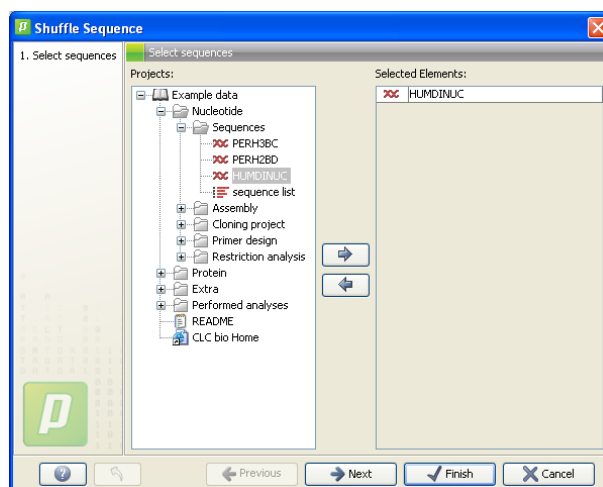


Figure 13.12: Choosing sequence for shuffling.

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 **Project Tree**.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**.

This will open a new view in the **View Area** displaying the shuffled sequence. The new sequence is not saved automatically. To save the protein sequence, drag it into the **Navigation Area** or press ctrl + S (⌘ + S on Mac) to activate a save dialog.

13.3 Local complexity plot

In *CLC Protein Workbench* it is possible to calculate local complexity for both DNA and protein sequences. The local complexity is a measure of the diversity in the composition of amino acids within a given range (window) of the sequence. The K2 algorithm is used for calculating local complexity [Wootton and Federhen, 1993]. To conduct a complexity calculation do the following:

Select sequences in Navigation Area | **Toolbox in Menu Bar** | **General Sequence Analyses** (📁) | **Create Complexity Plot** (📊)

This opens a dialog. In **Step 1** you can change, remove and add DNA and protein sequences.

When the relevant sequences are selected, clicking **Next** takes you to **Step 2**. This step allows you to adjust the window size from which the complexity plot is calculated. Default is set to 11 amino acids and the number should always be odd. The higher the number, the less volatile the graph.

Figure 13.13 shows an example of a local complexity plot.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**. The values of the complexity plot approaches 1.0 as the distribution of amino acids become more complex.

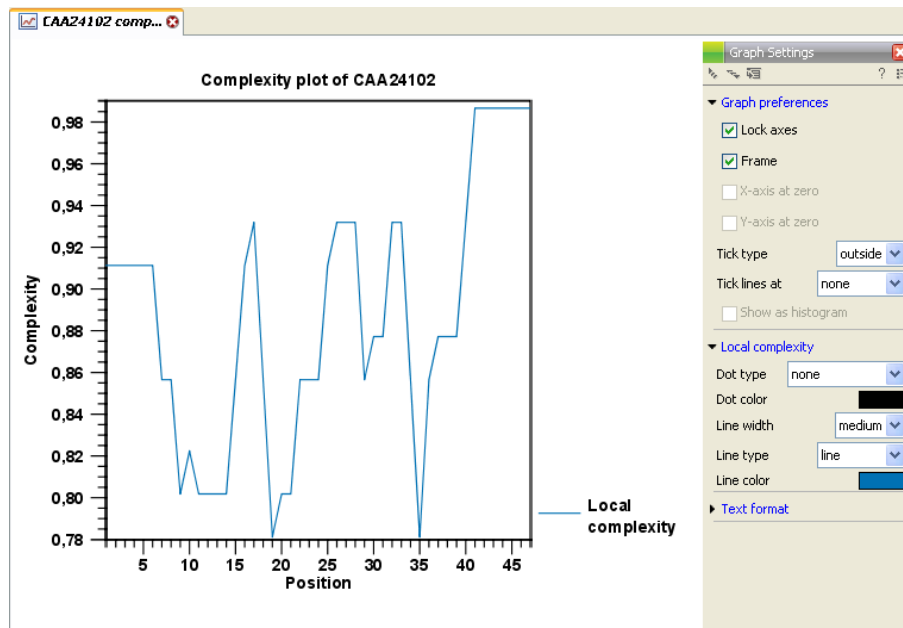


Figure 13.13: An example of a local complexity plot.

13.3.1 Local complexity view preferences

There are two groups of preferences for the local complexity view: Graph preferences and Local complexity preferences:

The **Graph preferences** apply to the whole graph:

- **Lock axis.**
This will always show the axis even though the plot is zoomed to a detailed level.
- **Frame.**
Toggles the frame of the graph.
- **X-axis at zero.**
Toggles the x-axis at zero.
- **Y-axis at zero.**
Toggles the y-axis at zero.
- **Tick type**
 - outside
 - inside
- **Tick lines at.**
Shows a grid behind the graph.
 - none
 - major ticks
- **Show as histogram**
For some data-series it is possible to see it as a histogram rather than a line plot.

The **Local complexity preferences** include:

- **Dot type**
 - none
 - cross
 - plus
 - square
 - diamond
 - circle
 - triangle
 - reverse triangle
 - dot
- **Dot color.** Allows you to choose between many different colors.
- **Line width**
 - thin
 - medium
 - wide
- **Line type**
 - none
 - line
 - long dash
 - short dash
- **Line color.** Allows you to choose between many different colors.

13.4 Sequence statistics

CLC Protein Workbench 2.0 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:

select sequence(s) | Toolbox in the Menu Bar | General Sequence Analyses (🔍) | Create Sequence Statistics (📊)

This opens a dialog where you can alter your choice of sequences which you want to create statistics for. You can also add sequence lists.

Notice! You cannot create statistics for DNA and protein sequences at the same time.

When the sequences are selected, click **Next**.

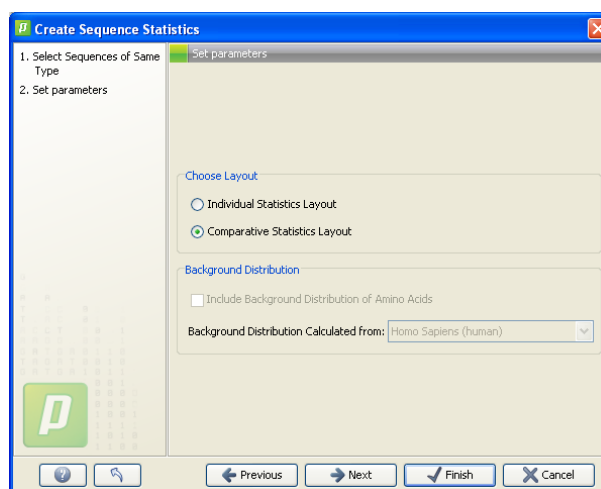


Figure 13.14: Setting parameters for the sequence statistics.

This opens the dialog displayed in figure 13.14.

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 8.1). If not, click **Finish**. An example of protein sequence statistics is shown in figure 13.15.

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.

Notice! 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
 - Locus
 - Description
 - Modification Date

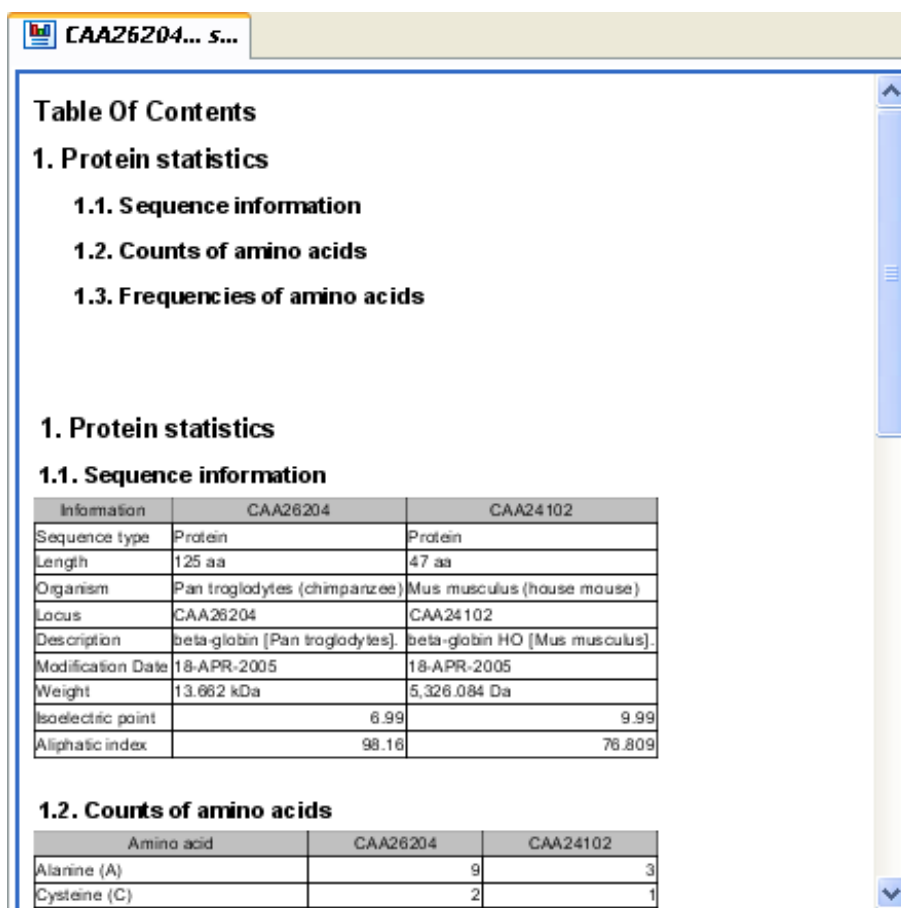


Table Of Contents

1. Protein statistics

1.1. Sequence information

1.2. Counts of amino acids

1.3. Frequencies of amino acids

1. Protein statistics

1.1. Sequence information

Information	CAA26204	CAA24102
Sequence type	Protein	Protein
Length	125 aa	47 aa
Organism	Pan troglodytes (chimpanzee)	Mus musculus (house mouse)
Locus	CAA26204	CAA24102
Description	beta-globin [Pan troglodytes]	beta-globin HO [Mus musculus]
Modification Date	18-APR-2005	18-APR-2005
Weight	13.662 kDa	5,326.084 Da
Isoelectric point	6.99	9.99
Aliphatic index	98.16	76.809

1.2. Counts of amino acids

Amino acid	CAA26204	CAA24102
Alanine (A)	9	3
Cysteine (C)	2	1

Figure 13.15: Comparative sequence statistics.

- Weight
- 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

The output of nucleotide sequence statistics include:

- General statistics:
 - Sequence type
 - Length
 - Organism
 - Locus
 - Description
 - Modification Date
 - Weight
- Atomic composition
- Nucleotide distribution table
- Nucleotide distribution histogram
- Annotation table
- Counts of di-nucleotides
- Frequency of di-nucleotides

A short description of the different areas of the statistical output is given in section [13.4.2](#).

13.4.1 Sequence statistics output

The entire statistical output can be printed. To do so, click the **Print** icon ().

13.4.2 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.

$$\text{Aliphatic index} = X(\text{Ala}) + a * X(\text{Val}) + b * X(\text{Leu}) + b * (X)\text{Ile} \quad (13.1)$$

$X(\text{Ala})$, $X(\text{Val})$, $X(\text{Ile})$ and $X(\text{Leu})$ 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].

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:

$$\text{Ext}(\text{Protein}) = \text{count}(\text{Cystine}) * \text{Ext}(\text{Cystine}) + \text{count}(\text{Tyr}) * \text{Ext}(\text{Tyr}) + \text{count}(\text{Trp}) * \text{Ext}(\text{Trp}), \quad (13.2)$$

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:

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*.

- 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:

$$Absorbance(Protein) = \frac{Ext(Protein)}{Molecular\ weight} \quad (13.3)$$

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.

Creative Commons License

All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in it's original form and "CLC bio" has to be clearly labelled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, or build upon this work.



See <http://creativecommons.org/licenses/by-nc-nd/2.5/> for more about how you may use the contents.

13.5 Join sequences

CLC Protein Workbench can join several nucleotide or protein sequences into one sequence. This feature can for example be used to construct "supergenes" for phylogenetic inference by joining several disjoint genes into one. Note, that when sequences are joined, all their annotations are carried over to the new spliced sequence.

Two (or more) sequences can be joined by:

select sequences to join | Toolbox in the Menu Bar | General Sequence Analyses | Join sequences (🔗)

or **select sequences to join | right-click either selected sequence | Toolbox | General Sequence Analyses | Join sequences (🔗)**

This opens the dialog shown in figure 13.16.

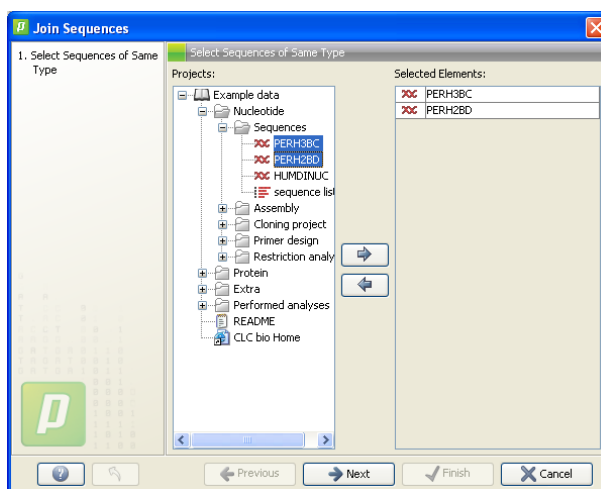


Figure 13.16: Selecting two alignments to be joined.

If you have selected some sequences 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 from the **Project Tree**. Click **Next** opens the dialog shown in figure 13.17.

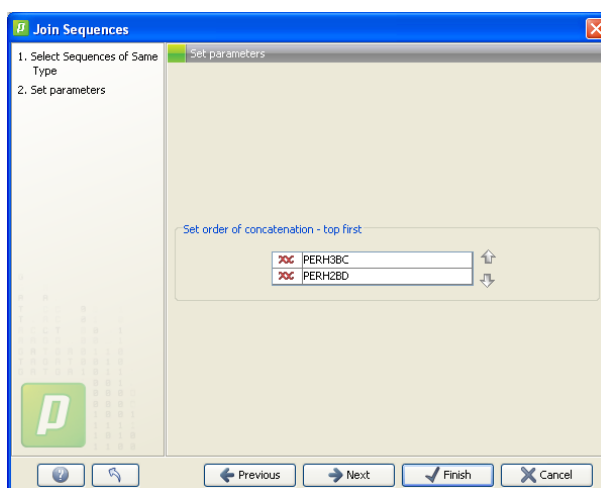


Figure 13.17: Setting the order in which sequences are joined.

In step 2 you can change the order in which the sequences will be joined. Select a sequence and

use the arrows to move the selected sequence up or down.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**.

The result is shown in figure 13.18.

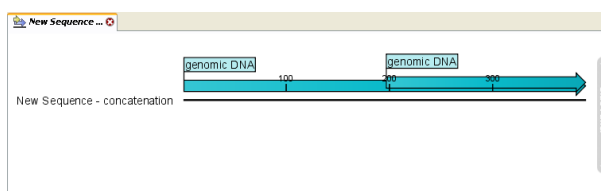


Figure 13.18: The result of joining sequences is a new sequence containing all the annotations of the joined sequences.

13.6 Motif Search

CLC Protein Workbench offers advanced and versatile options to search for unknown sequence patterns or known motifs represented either by a literal string or a regular expression. These advanced search capabilities are available for use in both DNA and protein sequences.

Difference between Motif Search and Pattern Discovery

In motif search (see 13.6), the user has some predefined knowledge about the pattern/motif of interest. This motif is defined by the user and the algorithm runs through the entire sequence and looks for identical or degenerate patterns. 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*.

Our pattern discovery algorithm (see 13.7) is based on proprietary hidden Markov models (HMM) and scans the entire sequence (one or more) for patterns which may be unknown to the user.

Motifs

If you have a known motif represented by a literal string or a sequence pattern of interest, you can search for them using the *CLC Protein Workbench*. Patterns and motifs can be searched with different levels of degeneracy in both DNA and protein sequences.

You can also search for matches with known motifs represented by a regular expression: A regular expression 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. You are limited to the following syntax rules (See the Java regular expression syntax):

$[A - Z]$ 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 .

$[\hat{A} - M]$ will match any character except those between A and M (Excluding). You can also put single characters between the brackets: The expression $[\hat{A}G]$ matches any character except A and G .

$[A - Z \& \& [\hat{M} - P]]$ 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 $[A - P \& \& [\hat{C}G]]$ 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 $ACGACG$.

$X\{n, m\}$ 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 , $ACTACT$ and $ACTACTACT$.

$X\{n, \}$ represents a repetition of an element at least n times. For example, $AC\{2, \}$ matches all strings $ACAC$, $ACACAC$, $ACACACAC$,...

The symbol \wedge restricts the search to the beginning of your sequence. For example, if you search through a sequence with the regular expression $\wedge AC$, 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 $[ACG][\hat{A}C]G\{2\}$ 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 $G.[\hat{A}]\$$ 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 .

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.

In order to search for a known motif:

Select DNA or protein sequence(s) | Toolbox in the Menu Bar | General Sequence Analyses  | **Motif Search** 

or **Right-click DNA or protein sequence(s) | Toolbox | General Sequence Analyses**  | **Motif Search** 

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 **Project Tree**.

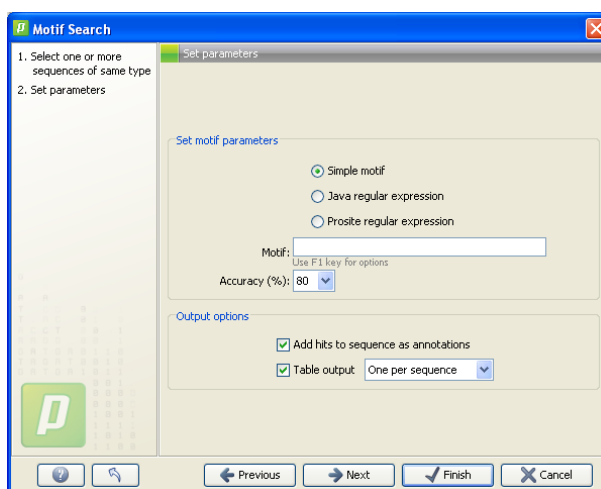


Figure 13.19: Setting parameters for the motif search. See text for details.

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 in the sequences and open a new view for each of the sequences.

Click **Next** to adjust parameters (see figure 13.19).

13.6.1 Motif search parameter settings

Various parameters can be set prior to the motif search. The parameters are listed below and a screen shot of the parameter settings can be seen in figure 13.19.

- **Motif types**

You can choose literal string (simple motif) or Java regular expression as your motif type. For proteins, you can choose to search with a Prosite regular expression.

- **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 your motif type is Java regular expression, you should enter a regular expression according to the syntax rules above. Press **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 search string to the match on the sequence.

- **Table output**

Opens the motifs or patterns found in a table view. It is possible to see one table per sequence but it is also possible with one table for multiple sequences.

- **Add motif to sequence as annotation**

Check this box to add search strings found as annotations on the sequence.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**. This will open a view showing the motifs or patterns found as annotations on the original

sequence (see figure 13.20). If you have selected several sequences, a corresponding number of views will be opened.



Figure 13.20: Sequence view displaying the pattern found. The search string was 'QRQXRXXXXQQ'.

13.6.2 Motif 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. If wanted, annotations on patterns found can be added to all the sequences. Each pattern found will be represented as an annotation of the type **Region**. More information on each motif or pattern found is available through the tooltip, including detailed information on the position of the pattern and how similar it was to the search string.

It is also possible to get a tabular view of all motifs or patterns found in either one combined table or in individual tables if multiple sequences were selected. Then each pattern found will be represented with its position in the sequence and the obtained accuracy score.

13.7 Pattern Discovery

With *CLC Protein 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:

Select DNA or protein sequence(s) | Toolbox in the Menu Bar | General Sequence Analyses (📁) | Pattern Discovery (🔍)

or **right-click DNA or protein sequence(s) | Toolbox | General Sequence Analyses (📁) | 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 **Project Tree**.

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 13.21).

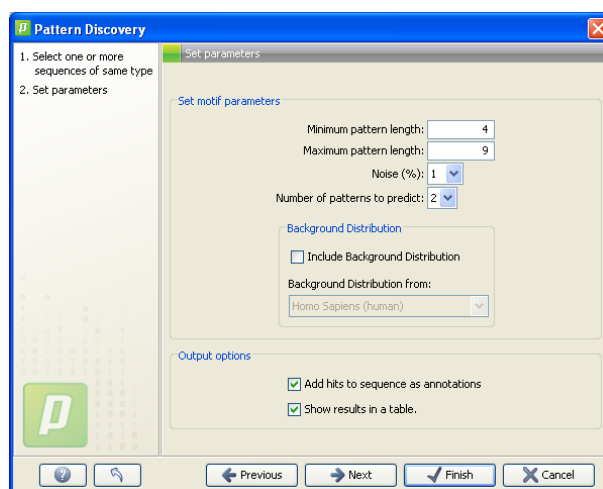


Figure 13.21: Setting parameters for the pattern discovery. See text for details.

13.7.1 Pattern discovery search parameters

Various parameters can be set prior to the pattern discovery. The parameters are listed below and a screen shot of the parameter settings can be seen in figure 13.21.

- **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.
- **Show result of patterns discovery in a table**
Generate a tabular output which displays patterns found.
- **Include Background Distribution of Amino Acids**
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 8.1). If not, click **Finish**. This will open a view showing the patterns found as annotations on the original sequence (see figure 13.22). If you have selected several sequences, a corresponding number of views will be opened.



Figure 13.22: Sequence view displaying two discovered patterns.

13.7.2 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 tooltip, 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.

Chapter 14

Nucleotide analyses

Contents

14.1 Convert DNA to RNA	165
14.2 Convert RNA to DNA	166
14.3 Reverse complements of sequences	167
14.4 Translation of DNA or RNA to protein	168
14.4.1 Translate part of a nucleotide sequence	169
14.5 Find open reading frames	169
14.5.1 Open reading frame parameters	170

CLC Protein Workbench 2.0 offers different kinds of sequence analyses, which only apply to DNA and RNA.

14.1 Convert DNA to RNA

CLC Protein Workbench 2.0 lets you convert a DNA sequence into RNA, substituting the T residues (Thymine) for U residues (Urasil):

select a DNA sequence in the Navigation Area | Toolbox in the Menu Bar | Nucleotide Analyses (📁) | Convert DNA to RNA (🔗)

or **right-click a sequence in Navigation Area | Toolbox | Nucleotide Analyses (📁) | Convert DNA to RNA (🔗)**

This opens the dialog displayed in figure 14.1:

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 **Project Tree**.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**.

Notice! You can select multiple DNA sequences and sequence lists at a time. If the sequence list contains RNA sequences as well, they will not be converted.

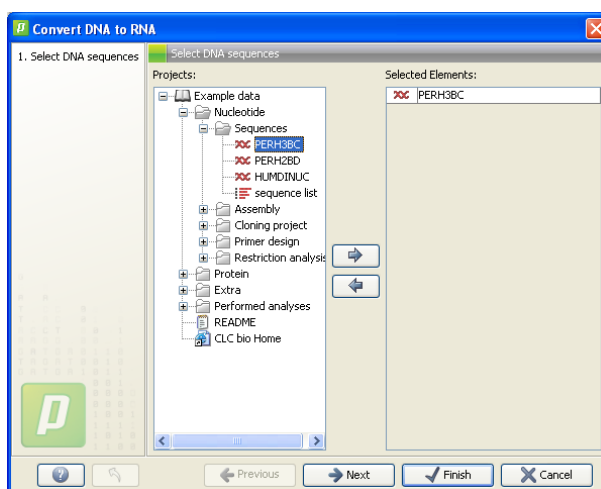


Figure 14.1: Translating DNA to RNA.

14.2 Convert RNA to DNA

CLC Protein Workbench 2.0 lets you convert an RNA sequence into DNA, substituting the U residues (Urasil) for T residues (Thymine):

select an RNA sequence in the Navigation Area | Toolbox in the Menu Bar | Nucleotide Analyses (🔍) | Convert RNA to DNA (🔄)

or **right-click a sequence in Navigation Area | Toolbox | Nucleotide Analyses (🔍) | Convert RNA to DNA (🔄)**

This opens the dialog displayed in figure 14.2:

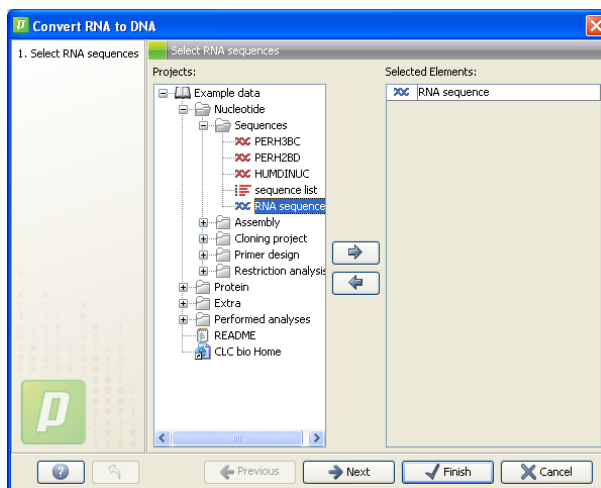


Figure 14.2: Translating RNA to DNA.

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 **Project Tree**.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**.

This will open a new view in the **View Area** displaying the new DNA sequence. The new sequence is not saved automatically. To save the protein sequence, drag it into the **Navigation Area** or

press Ctrl + S (⌘ + S on Mac) to activate a save dialog.

Notice! You can select multiple RNA sequences and sequence lists at a time. If the sequence list contains DNA sequences as well, they will not be converted.

14.3 Reverse complements of sequences

CLC Protein Workbench 2.0 is able to create the reverse complement of a nucleotide sequence. By doing that, a new sequence is created which also has all the annotations reversed since they now occupy the opposite strand of their previous location.

To quickly obtain the reverse complement of a sequence or part of a sequence, you may select a region on the negative strand and open it in a new view:

right-click a selection on the negative strand | Open selection in a new view

By doing that, the sequence will be reversed. This is only possible when the double stranded view option is enabled. It is possible to copy the selection and paste it in a word processing program or an e-mail. To obtain a reverse complement of an entire sequence:

select a sequence in the Navigation Area | Toolbox in the Menu Bar | Nucleotide Analyses (📄) | Create Reverse Complement (🔄)

or **right-click a sequence in Navigation Area | Toolbox | Nucleotide Analyses (📄) | Create Reverse Complement (🔄)**

This opens the dialog displayed in figure 14.3:

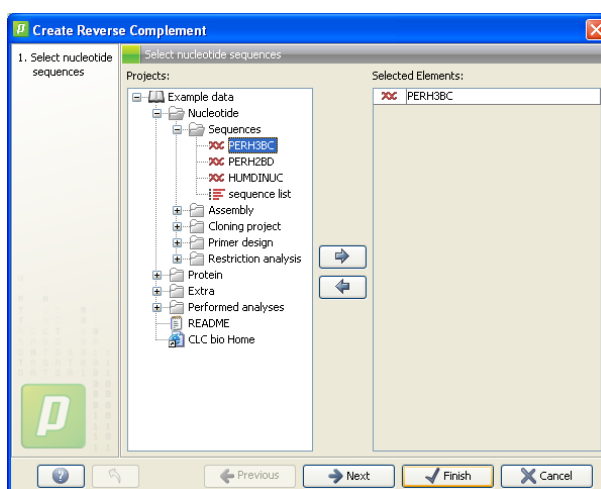


Figure 14.3: Creating a reverse complement sequence.

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 **Project Tree**.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**.

This will open a new view in the **View Area** displaying the reverse complement of the selected sequence. The new sequence is not saved automatically. To save the protein sequence, drag it into the **Navigation Area** or press Ctrl + S (⌘ + S on Mac) to activate a save dialog.

14.4 Translation of DNA or RNA to protein

In *CLC Protein Workbench 2.0* you can translate a nucleotide sequence into a protein sequence using the **Toolbox** tools. Usually, you use the +1 reading frame which means that the translation starts from the first nucleotide. Stop codons result in an asterisk being inserted in the protein sequence at the corresponding position. It is possible to translate in any combination of the six reading frames in one analysis. To translate:

select a nucleotide sequence | Toolbox in the Menu Bar | Nucleotide Analyses (📁) | Translate to Protein (📄)

or **right-click a nucleotide sequence | Toolbox | Nucleotide Analyses (📁) | Translate to Protein (📄)**

This opens the dialog displayed in figure 14.4:

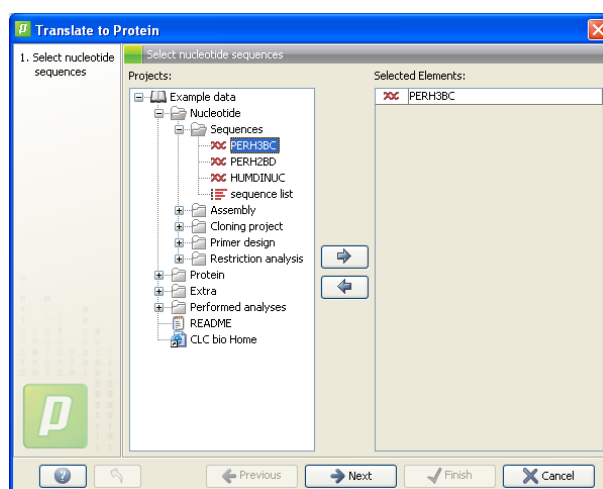


Figure 14.4: Choosing sequences for translation.

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 **Project Tree**.

Click **Next** to set reading frames, select if you want to translate all coding regions of the sequence and choose translation tables. Clicking **Next** generates the dialog seen in figure 14.5:

The translation tables in *CLC Protein Workbench* are updated regularly from NCBI. Therefore the tables are not available in this printable version of the user manual. Instead the tables are included in the **Help**-menu in the **Menu Bar** under **Background Information**.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**. The newly created protein is shown, but is not saved automatically.

There are also new views of proteins for every **CDS** or **ORF** annotation if you have selected to translate all coding regions.

To save a protein sequence, drag it into the **Navigation Area** or press Ctrl + S (⌘ + S on Mac) to activate a save dialog.

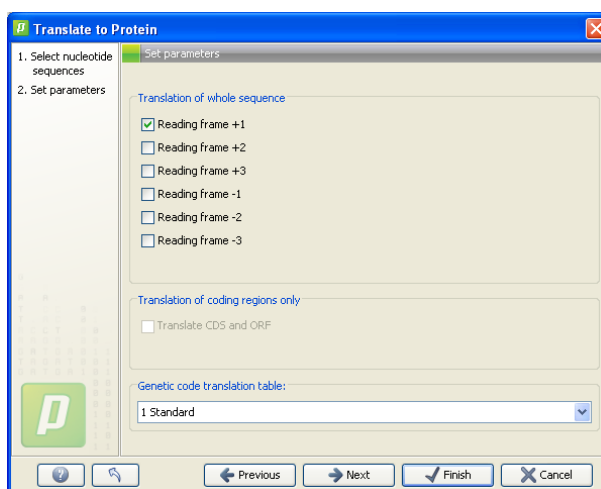


Figure 14.5: Choosing +1 and +3 reading frames, and the standard translation table.

14.4.1 Translate part of a nucleotide sequence

If you want to make separate translations of *all* the coding regions of a nucleotide sequence, you can check the option: "Translate CDS and ORF" in the translation dialog (see figure 14.5).

If you want to translate a *specific* coding region, which is annotated on the sequence, use the following procedure:

Open the nucleotide sequence | right-click the ORF or CDS annotation | Translate CDS/ORF | choose a translation table | OK (📄)

If the annotation contains information about the translation, this information will be used, and you do not have to specify a translation table.

The CDS and ORF annotations are colored yellow as default.

14.5 Find open reading frames

CLC Protein Workbench 2.0 has a basic functionality for gene finding in the form of open reading frame (ORF) determination. The ORFs will be shown as annotations on the sequence. You have the option of choosing translation table, start codons, minimum length and other parameters for finding the ORFs. These parameters will be explained in this section.

To find open reading frames:

select a nucleotide sequence | Toolbox in the Menu Bar | Nucleotide Analyses (📄) | Find Open Reading Frames (🔍)

or **right-click a nucleotide sequence | Toolbox | Nucleotide Analyses (📄) | Find Open Reading Frames (🔍)**

This opens the dialog displayed in figure 14.6:

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 **Project Tree**.

If you want to adjust the parameters for finding open reading frames click **Next**.

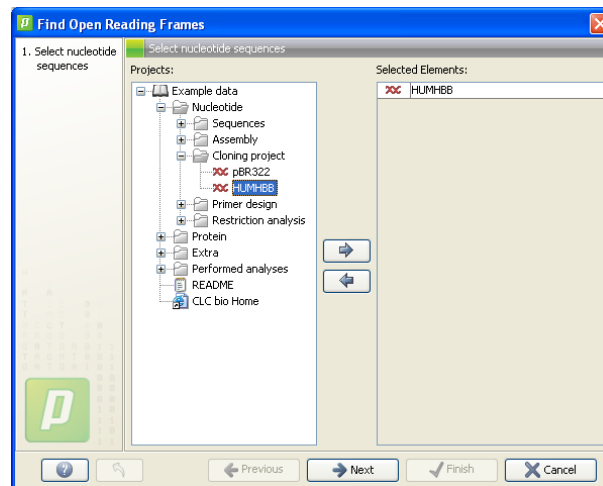


Figure 14.6: Create Reading Frame dialog.

14.5.1 Open reading frame parameters

This opens the dialog displayed in figure 14.7:

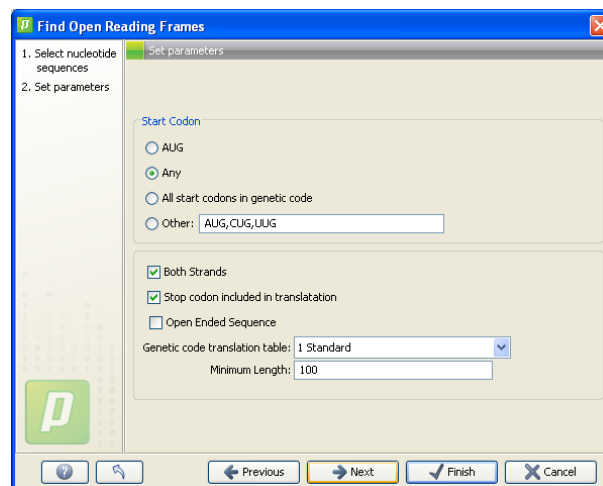


Figure 14.7: Create Reading Frame dialog.

The adjustable parameters for the search are:

- **Start Codon:**
 - **AUG.** Most commonly used start codon.
 - **Any.**
 - **All start codons in genetic code.**
 - **Other.** Here you can specify a number of start codons separated by commas.
- **Both Strands.** Finds reading frames on both strands.
- **Stop Codon included in Annotation** The ORFs will be shown as annotations which can include the stop codon if this option is checked.

- **Open Ended Sequence.** Allows the ORF to start or end outside the sequence. If the sequence studied is a part of a larger sequence, it may be advantageous to allow the ORF to start or end outside the sequence.
- **Genetic code translation table.** The translation tables are occasionally updated from NCBI. The tables are not available in this printable version of the user manual. Instead, the tables are included in the **Help**-menu in the **Menu Bar** under **Background Information**.
- **Minimum Length.** Specifies the minimum length for the ORFs to be found.

Using open reading frames for gene finding is a fairly simple approach which is likely to predict genes which are not real. Setting a relatively high minimum length of the ORFs will reduce the number of false positive predictions, but at the same time short genes may be missed (see figure 14.8).

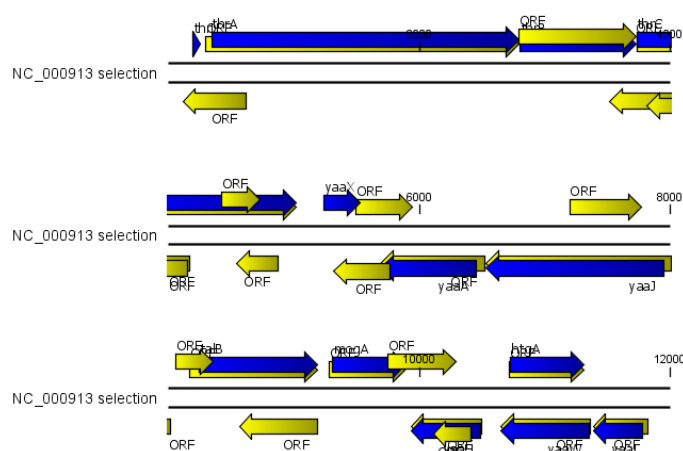


Figure 14.8: The first 12,000 positions of the E. coli sequence NC_000913 downloaded from GenBank. The blue (dark) annotations are the genes while the yellow (brighter) annotations are the ORFs with a length of at least 100 amino acids. On the positive strand around position 11,000, a gene starts before the ORF. This is due to the use of the standard genetic code rather than the bacterial code. This particular gene starts with CTG, which is a start codon in bacteria. Two short genes are entirely missing, while a handful of open reading frames do not correspond to any of the annotated genes.

Finding open reading frames is often a good first step in annotating sequences such as cloning vectors or bacterial genomes. For eukaryotic genes, ORF determination may not always be very helpful since the intron/exon structure is not part of the algorithm.

Chapter 15

Protein analyses

Contents



15.1 Signal peptide prediction	173
15.1.1 Signal peptide prediction parameter settings	173
15.1.2 Signal peptide prediction output	174
15.1.3 Bioinformatics explained: Prediction of signal peptides	174
15.2 Protein charge	179
15.2.1 Modifying the layout	180
15.3 Transmembrane helix prediction	181
15.4 Antigenicity	183
15.4.1 Plot of antigenicity	183
15.5 Hydrophobicity	185
15.5.1 Hydrophobicity plot	185
15.5.2 Hydrophobicity graphs along sequence	187
15.5.3 Bioinformatics explained: Protein hydrophobicity	188
15.6 Pfam domain search	190
15.6.1 Pfam search parameters	191
15.6.2 Download and installation of additional Pfam databases	192
15.7 Secondary structure prediction	193
15.8 Protein report	194
15.8.1 Protein report output	196
15.9 Reverse translation from protein into DNA	197
15.9.1 Reverse translation parameters	197
15.9.2 Bioinformatics explained: Reverse translation	198
15.10 Proteolytic cleavage detection	201
15.10.1 Proteolytic cleavage parameters	201
15.10.2 Bioinformatics explained: Proteolytic cleavage	203

CLC Protein Workbench 2.0 offers a number of analyses of proteins as described in this chapter.

15.1 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, *CLC Protein Workbench* uses SignalP version 3.0 [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 paper [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 15.1.3). This score has been shown to be the most accurate [Klee and Ellis, 2005] in an evaluation study of signal peptide predictors.

Select a protein sequence | Toolbox in the Menu Bar | Protein Analyses  | Predict signal peptide 

or **right-click a protein sequence | Toolbox | Protein Analyses  | Predict signal peptide **

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 **Project Tree**.

Click **Next** to set parameters for the SignalP analysis.

15.1.1 Signal peptide prediction parameter settings

It is possible to set different options prior to running the analysis (see figure 15.1). An organism type should be selected. The default is eukaryote.

- Eukaryote (default)
- Gram-negative bacteria
- Gram-positive bacteria

The predictions obtained can either be shown as annotations on the sequence or be shown as the detailed and full text output from the SignalP method. This can be used to interpret borderline predictions.

- Put annotation on sequence (default)
- Text output

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.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**.

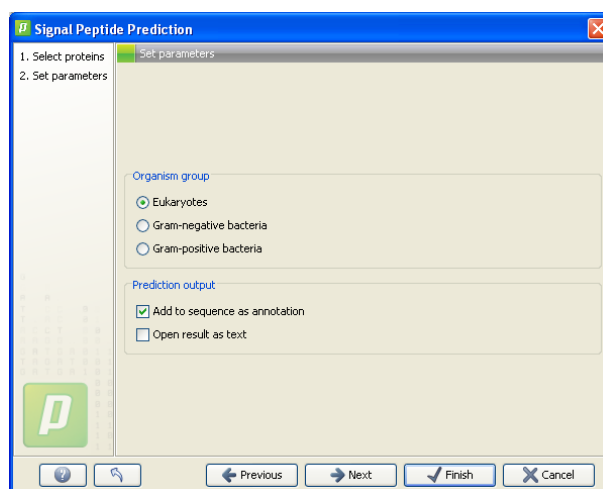


Figure 15.1: Setting the parameters for signal peptide prediction.

15.1.2 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 15.2).

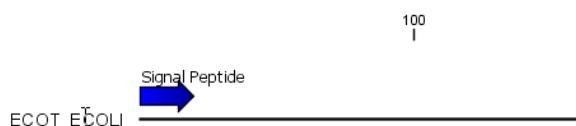


Figure 15.2: N-terminal signal peptide shown as annotation on the sequence.

Each annotation will carry a tooltip note saying that the corresponding annotation is predicted with SignalP version 3.0. Additional notes can be added through the **Add annotation** right-click mouse menu. See section 11.1.4.

Undesired annotations can be removed through the **Edit annotation** right-click mouse menu. See section 11.1.5.

15.1.3 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 PubMed¹ 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.

¹<http://www.ncbi.nlm.nih.gov/entrez/>

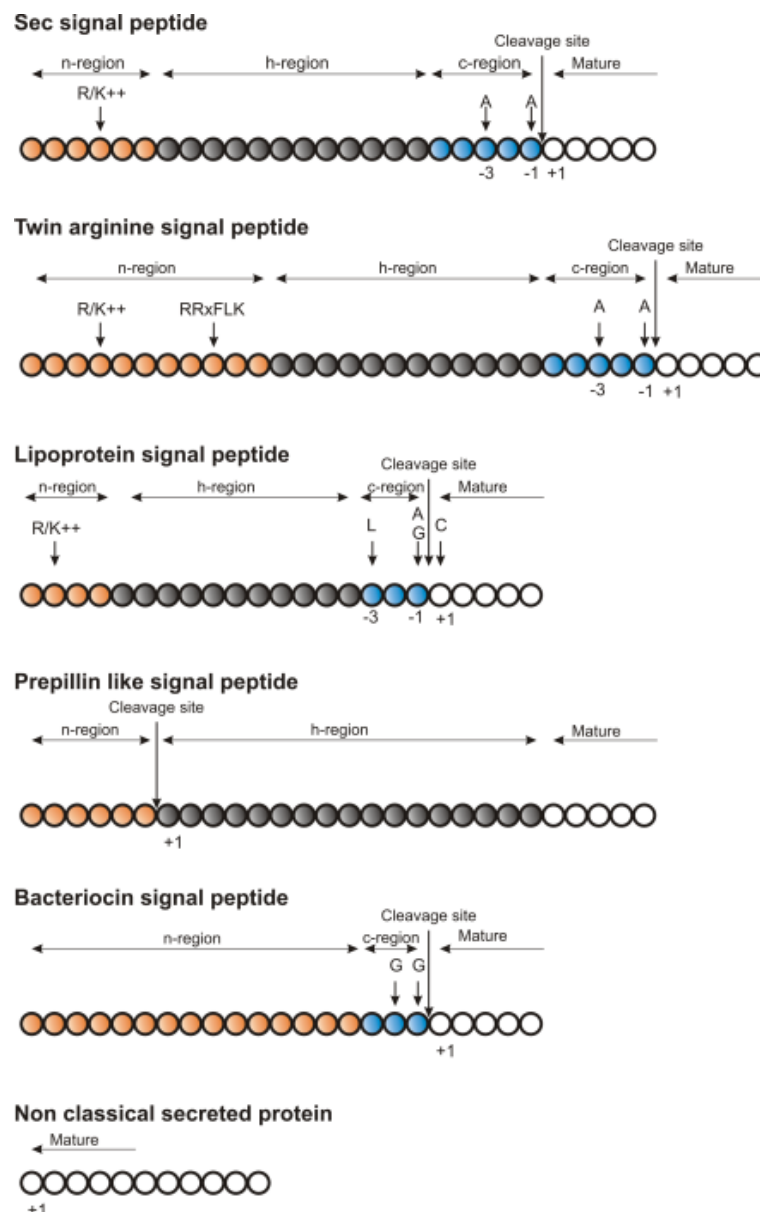


Figure 15.3: Schematic representation of various signal peptides. Red color indicates n-region, gray color indicates h-region, cyan indicates c-region. All white circles are part of the mature protein. +1 indicates the first position of the mature protein. The length of the signal peptides is not drawn to scale.

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

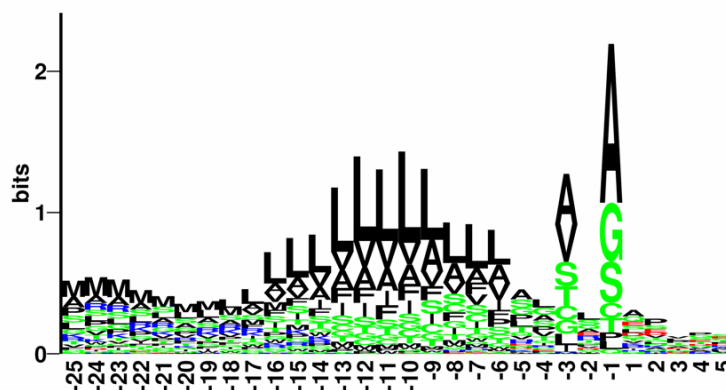


Figure 15.4: Sequence logo of eukaryotic signal peptides, showing conservation of amino acids in bits [Schneider and Stephens, 1990]. Polar and hydrophobic residues are shown in green and black, respectively, while blue indicates positively charged residues and red negatively charged residues. The logo is based on an ungapped sequence alignment fixed at the -1 position of the signal peptides.

protein and as such cannot be described as a new group of signal peptides.

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 15.3.

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].

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.

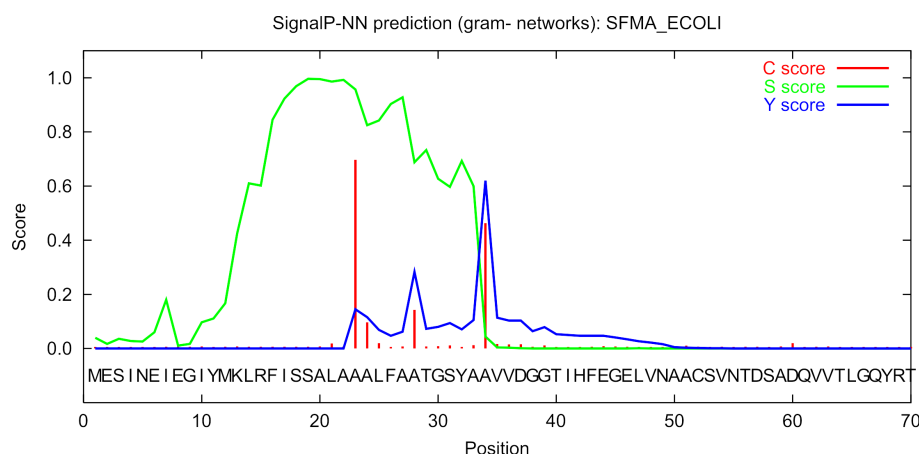


Figure 15.5: Graphical output from the SignalP method of Swiss-Prot entry *SFMA_ECOLI*. Initially this seemed like a borderline prediction, but closer inspection of the sequence revealed an internal methionine at position 12, which could indicate a erroneously annotated start of the protein. Later this protein was reannotated by Swiss-Prot to start at the M in position 12. See the text for description of the scores.

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. 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].

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 15.5.

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 entry for the original paper.

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=15223320&dopt=Citation

Creative Commons License

All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in its original form and "CLC bio" has to be clearly labelled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, or build upon this work.



See <http://creativecommons.org/licenses/by-nc-nd/2.5/> for more about how you may use the contents.

15.2 Protein charge

In *CLC Protein Workbench* you can create a graph in the electric charge of a protein as a function of pH. This is particularly useful for finding the net charge of the protein at a given pH. This knowledge can be used e.g. in relation to isoelectric focusing on the first dimension of 2D-gel electrophoresis. The isoelectric point (pI) is found where the net charge of the protein is zero. The calculation of the protein charge does not include knowledge about any potential post-translational modifications the protein may have.

In order to calculate the protein charge:

Select a protein sequence | Toolbox in the Menu Bar | Protein Analyses (📁) | Create Protein Charge Plot (📈)

or **right-click a protein sequence | Toolbox | Protein Analyses (📁) | Create Protein Charge Plot (📈)**

This opens the dialog displayed in figure 15.6:

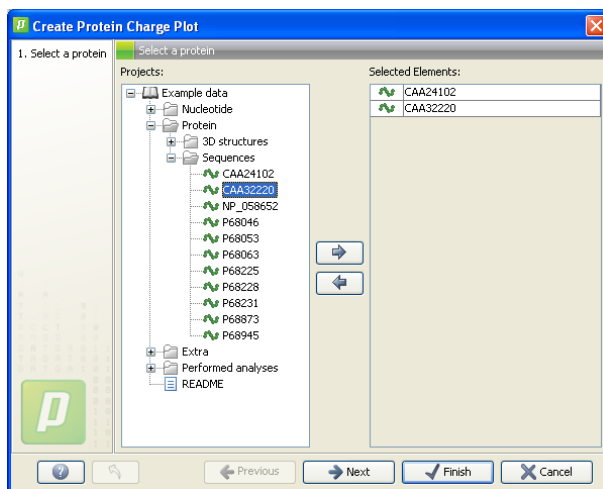


Figure 15.6: Choosing protein sequences to calculate protein charge.

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 **Project Tree**.

You can perform the analysis on several protein sequences at a time. This will result in one output graph showing protein charge graphs for the individual proteins.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**.

15.2.1 Modifying the layout

Figure 15.7 shows the electrical charges for three proteins. In the **Side Panel** to the right, you can modify the layout of the graph.

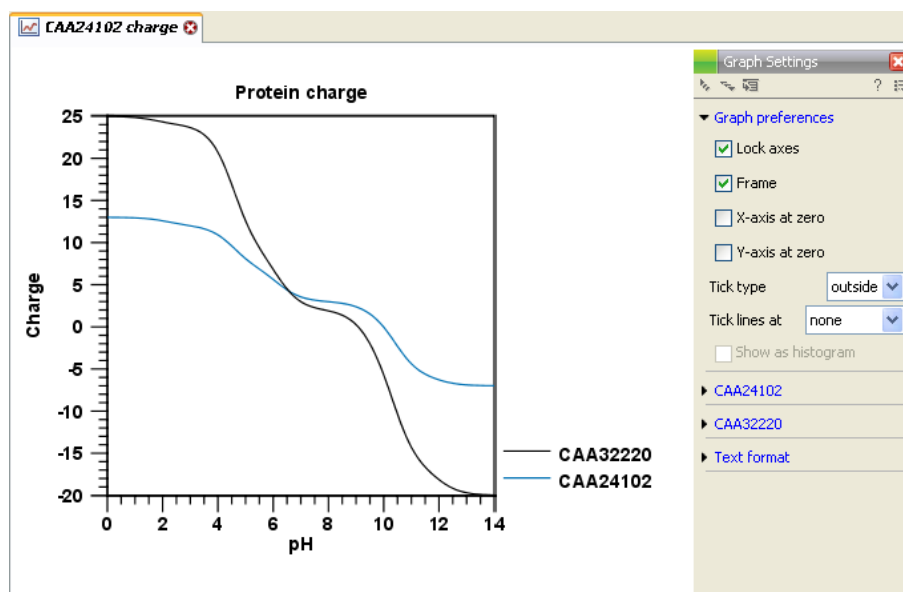


Figure 15.7: View of the protein charge.

Graph preferences

The **Graph preferences** apply to the whole graph:

- **Lock axis.**
This will always show the axis even though the plot is zoomed to a detailed level.
- **Frame.**
Toggles the frame of the graph.
- **X-axis at zero.**
Toggles the x-axis at zero.
- **Y-axis at zero.**
Toggles the y-axis at zero.
- **Tick type**
 - outside
 - inside
- **Tick lines at.**
Shows a grid behind the graph.
 - none
 - major ticks
- **Show as histogram**
For some data-series it is possible to see it as a histogram rather than a line plot.

Preferences for each protein

Underneath the **Graph preferences** you will find is a set of preferences for each protein in the graph. These preferences only apply to the curve for the specific protein.

- **Dot type**

- none
- cross
- plus
- square
- diamond
- circle
- triangle
- reverse triangle
- dot

- **Dot color.** Allows you to choose between many different colors.

- **Line width**

- thin
- medium
- wide

- **Line type**

- none
- line
- long dash
- short dash

- **Line color.** Allows you to choose between many different colors.

These settings will apply to both the curve and the legend.

Modifying labels and legends

Click the title of the graph, the axis-titles or the legend to edit the text.

15.3 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 Protein 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 TMHMM 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].

Select a protein sequence | Toolbox in the Menu Bar | Protein Analyses (📁) | Transmembrane Helix Prediction (🔍)

or **right-click a protein sequence | Toolbox | Protein Analyses (📁) | 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 **Project Tree**.

Click **Next** to set parameters for the TMHMM analysis.

The predictions obtained can either be shown as annotations on the sequence or be shown as the detailed and text output from the TMHMM method. These options are chosen in **Step 2** (see figure 15.8).

- Add transmembrane helices as annotation on the sequence
- Open result as text

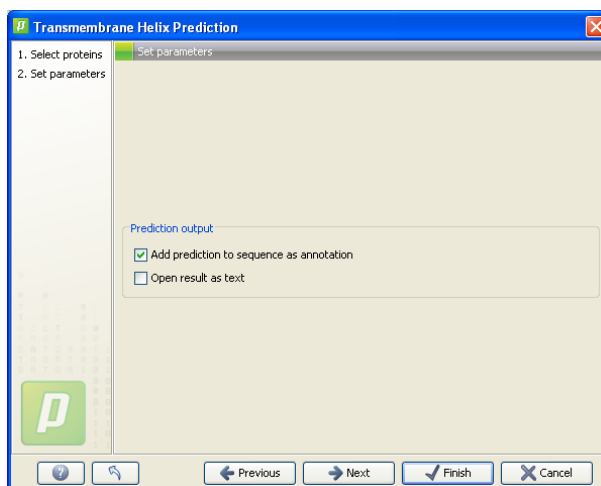


Figure 15.8: Choosing one or more protein sequences for transmembrane helix prediction.

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 transmembrane helix is found. If a transmembrane helix is not found a dialog box will be presented.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**.

After running the prediction as described above, the protein sequence will show predicted transmembrane helices as annotations on the original sequence (see figure 15.9). Moreover, annotations showing the topology will be shown. That is, which part the proteins is located on the inside or on the outside.

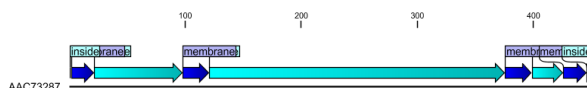


Figure 15.9: Transmembrane segments shown as annotation on the sequence and the topology.

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 **Add annotation** right-click mouse menu. See section 11.1.4.

Undesired annotations can be removed through the **Edit annotation** right-click mouse menu. See section 11.1.5.

15.4 Antigenicity

CLC Protein Workbench can help to identify antigenic regions in protein sequences in different ways, using different algorithms. The algorithms provided in the Workbench, merely plot an index of antigenicity over the sequence.

Two different methods are available.

[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.

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%.

Note! Similar results from the two method can not always be expected as the two methods are based on different training sets.

(See section 15.4). Furthermore, antigenicity of sequences can be displayed as antigenicity plots and as graphs along sequences. Finally, *CLC Protein Workbench 2.0* can calculate antigenicity for several sequences at the same time, and for alignments.

15.4.1 Plot of antigenicity

Displaying the antigenicity for a protein sequence in a plot is done in the following way:

select a protein sequence in Navigation Area | Toolbox in the Menu Bar | Protein Analyses (📁) | Create Antigenicity Plot (📈)

This opens a dialog. The first step allows you to add or remove sequences. Clicking **Next** takes you through to **Step 2**, which is displayed in figure 15.10.

The **Window size** is the width of the window where, the antigenicity is calculated. The wider the window, the less volatile the graph. You can chose from a number of antigenicity scales. Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**. The result can be seen in figure 15.11.

CLC Protein Workbench 2.0 offers some **View Preferences** for the view of the antigenicity plot.

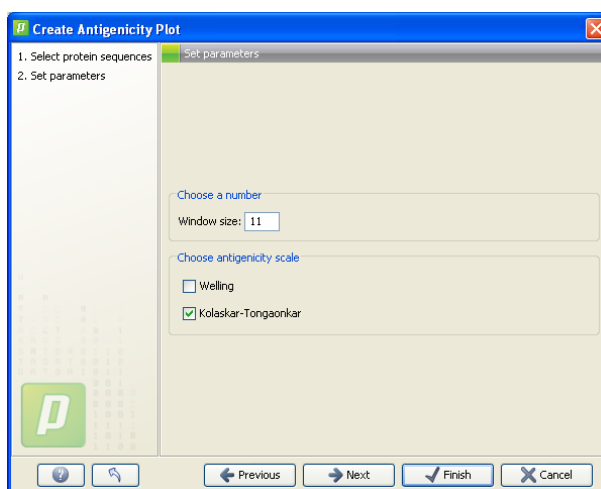


Figure 15.10: Step two in the Antigenicity Plot allows you to choose different antigenicity scales and the window size.

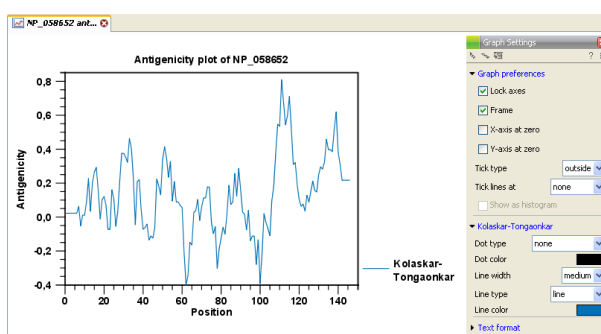


Figure 15.11: The result of the antigenicity plot calculation and the associated Side Panel.

The drop down menus are opened by clicking the black triangular arrows.

There are two kinds of view preferences. The graph preferences and preferences for the kind of hydrophobicity scale used to calculate the graph, e.g. Welling.

The **Graph preferences** include:

- **Lock axis.**
This will always show the axis even though the plot is zoomed to a detailed level.
- **Frame.**
Toggles the frame of the graph.
- **X-axis at zero.**
Toggles the x-axis at zero.
- **Y-axis at zero.**
Toggles the y-axis at zero.
- **Tick type**
 - outside
 - inside

- **Tick lines at.**

Shows a grid behind the graph.

- none
- major ticks

- **Show as histogram**

For some data-series it is possible to see it as a histogram rather than a line plot.

The preferences for the different scales are identical and include the following:

- **Dot type.** Lets you choose the marking of dots in the graph.
- **Dot color.** Lets you choose the color of the dots.
- **Line width.** Applies to the line connecting the dots.
- **Line type.** Applies to the line connecting the dots.
- **Line color.** Applies to the line connecting the dots.

The level of antigenicity is calculated on the basis of the different scales. The different scales add different values to each type of amino acid. The antigenicity 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 antigenicity scores.

15.5 Hydrophobicity

CLC Protein Workbench can calculate the hydrophobicity of protein sequences in different ways, using different algorithms. (See section 15.5.3). Furthermore, hydrophobicity of sequences can be displayed as hydrophobicity plots and as graphs along sequences. In addition, *CLC Protein Workbench 2.0* can calculate hydrophobicity for several sequences at the same time, and for alignments.

15.5.1 Hydrophobicity plot

To display the hydrophobicity for a protein sequence in a plot is done in the following way:

select a protein sequence in Navigation Area | Toolbox in the Menu Bar | Protein Analyses  **| Create Hydrophobicity Plot** 

This opens a dialog. The first step allows you to add or remove sequences. Clicking **Next** takes you through to **Step 2**, which is displayed in figure 15.12.

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 choose from a number of hydrophobicity scales which are further explained in section 15.5.3 Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**. The result can be seen in figure 15.13.

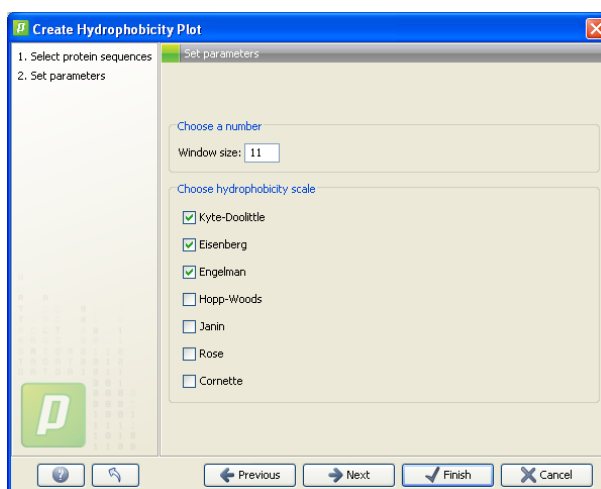


Figure 15.12: Step two in the Hydrophobicity Plot allows you to choose hydrophobicity scale and the window size.

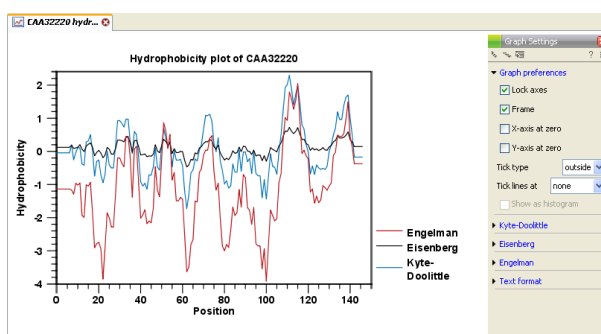


Figure 15.13: The result of the hydrophobicity plot calculation and the associated Side Panel.

In *CLC Protein Workbench 2.0* it is possible to change the layout of the hydrophobicity plot through the **Side Panel**. The drop down menus are opened by clicking the black triangular arrows.

There are two kinds of view preferences. The graph preferences and preferences for the kind of hydrophobicity scale used to calculate the graph, e.g. Kyte-Doolittle.

The **Graph preferences** include:

- **Lock axis.**
This will always show the axis even though the plot is zoomed to a detailed level.
- **Frame.**
Toggles the frame of the graph.
- **X-axis at zero.**
Toggles the x-axis at zero.
- **Y-axis at zero.**
Toggles the y-axis at zero.
- **Tick type**
 - outside
 - inside

- **Tick lines at.**

Shows a grid behind the graph.

- none
- major ticks

- **Show as histogram**

For some data-series it is possible to see it as a histogram rather than a line plot.

The preferences for the different scales are identical and include the following:

- **Dot type.**

Lets you choose the marking of dots in the graph.

- **Dot color.**

Lets you choose the color of the dots.

- **Line width.**

Applies to the line connecting the dots.

- **Line type.**

Applies to the line connecting the dots.

- **Line color.**

Applies to the line connecting the dots.

15.5.2 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 Hydrophobicity info in Side Panel

or **double-click protein sequence in Navigation Area | Show | Sequence | open Hydrophobicity info in Side Panel**

These actions result in the view displayed in figure 15.14.



Figure 15.14: The different available scales in Hydrophobicity info in **CLC Protein Workbench 2.0**.

The level of hydrophobicity is calculated on the basis of the different scales. The different scales 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 15.5.3).

In the following we will focus on the different ways that *CLC Protein Workbench 2.0* 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 15.15).

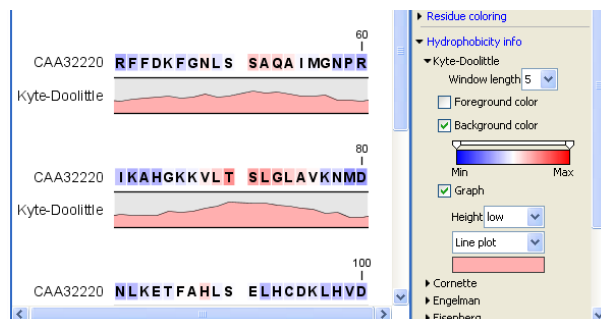


Figure 15.15: The different ways of displaying the hydrophobicity scores, using the Kyte-Doolittle scale.

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 15.15. Notice that you can choose the height of the graphs underneath the sequence.

15.5.3 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.

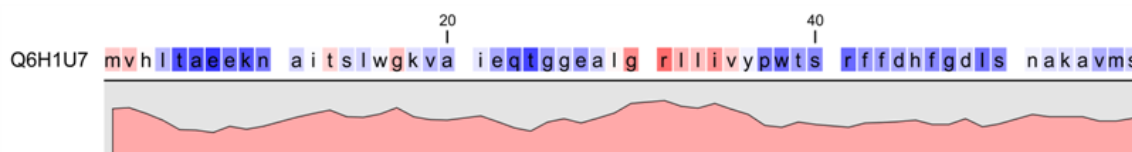


Figure 15.16: Plot of hydrophobicity along the amino acid sequence. Hydrophobic regions on the sequence have higher numbers according to the graph below the sequence, furthermore hydrophobic regions are colored on the sequence. Red indicates regions with high hydrophobicity and blue indicates regions 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 15.16).

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].

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>

Creative Commons License

All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in it's original form and "CLC bio" has to be clearly labelled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, or build upon this work.

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 15.1: *Hydrophobicity scales. This table shows seven different hydrophobicity scales which are generally used for prediction of e.g. transmembrane regions and antigenicity.*



See <http://creativecommons.org/licenses/by-nc-nd/2.5/> for more about how you may use the contents.

15.6 Pfam domain search

With *CLC Protein Workbench* you can perform a search for Pfam domains on protein sequences. The Pfam database at <http://www.sanger.ac.uk/Software/Pfam/> is a large collection of multiple sequence alignments that covers approximately 8000 protein domains and protein families [Bateman et al., 2004]. Based on the individual domain alignments, profile HMMs have been developed. These profile HMMs can be used to search for domains in unknown sequences.

Many proteins have a unique combination of domains which can be responsible, for instance, for the catalytic activities of enzymes. Pfam was initially developed to aid the annotation of the *C. elegans* genome. Annotating unknown 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]. An unknown protein may be annotated wrongly, for instance, as an enzyme if the pairwise alignment only finds a regulatory domain.

Using the Pfam search option in *CLC Protein Workbench*, you can search for domains in sequence data which otherwise do not carry any annotation information. The Pfam search option adds all found domains onto the protein sequence which was used for the search. If domains of no

relevance are found they can easily be removed as described in section 11.1.5. Setting a lower cutoff value will result in fewer domains.

In *CLC Protein Workbench* we have implemented our own HMM algorithm for prediction of the Pfam domains. Thus, we do not use the original HMM implementation, HMMER <http://hmmer.wustl.edu/> for domain prediction. We find the most probable state path/alignment through each profile HMM by the Viterbi algorithm and based on that we derive a new null model by averaging over the emission distributions of all *M* and *I* states that appear in the state path (*M* is a match state and *I* is an insert state). From that model we now arrive at an additive correction to the original bit-score, like it is done in the original HMMER algorithm.

In order to conduct the Pfam search:

Select a protein sequence | Toolbox in the Menu Bar | Protein Analyses (📁) | Pfam Domain Search (🔍)

or **right-click a protein sequence | Toolbox | Protein Analyses (📁) | Pfam Domain Search (🔍)**

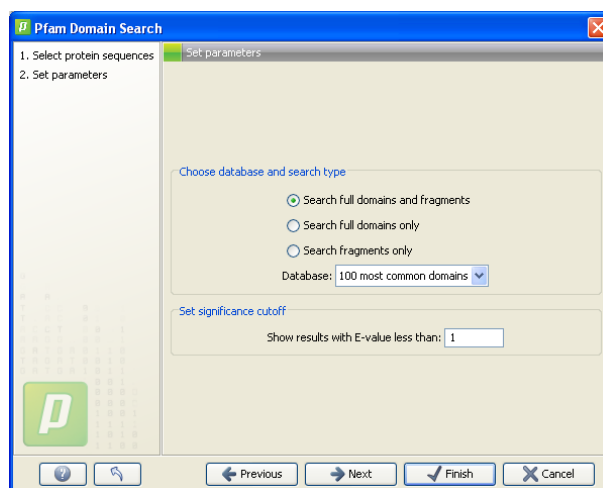


Figure 15.17: Setting parameters for Pfam domain search.

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 **Project Tree**.

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. Click **Next** to adjust parameters (see figure 15.17).

15.6.1 Pfam search parameters

- **Choose database and search type**

When searching for Pfam domains it is possible to choose different databases and specify the search for full domains or fragments of domains. Only the 100 most frequent domains are included as default in *CLC Protein Workbench*. Additional databases can be downloaded directly from CLC bio's website at www.clcbio.com.

- **Search full domains and fragments**

This option allows you to search both for full domain but also for partial domains. This

could be the case if a domain extends beyond the ends of a sequence.

- **Search full domains only**

Selecting this option only allows searches for full domains.

- **Search fragments only**

Only partial domains will be found.

- **Database**

Only the 100 most frequent domains are included as default in *CLC Protein Workbench*, but additional databases can be downloaded and installed as described in section 15.6.2.

- **Set significance cutoff**

The E-value (expectation value) is the number of hits that would be expected to have a score equal to or better than this value, by chance alone. This means that a good E-value which gives a confident prediction is much less than 1. E-values around 1 is what is expected by chance. Thus, the lower the E-value, the more specific the search for domains will be. Only positive numbers are allowed.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**. This will open a view showing the found domains as annotations on the original sequence (see figure 15.18). If you have selected several sequences, a corresponding number of views will be opened.

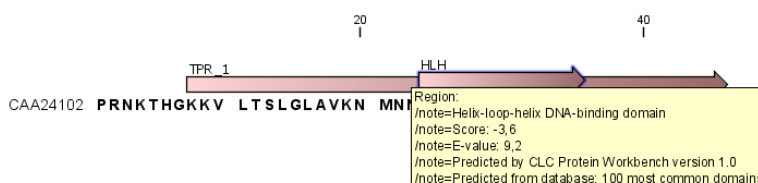


Figure 15.18: Domains annotations based on Pfam.

Each found domain will be represented as an annotation of the type **Region**. More information on each found domain is available through the tooltip, including detailed information on the identity score which is the basis for the prediction.

For a more detailed description of the provided scores through the tooltip look at <http://www.sanger.ac.uk/Software/Pfam/help/scores.shtml>.

15.6.2 Download and installation of additional Pfam databases

Additional databases can be downloaded from <http://www.clcbio.com> under the software | download sections. Here are databases containing the 100 most frequent domains, the 500 most frequent domains, and the complete database of approximately 8000 domains. This site also includes descriptions (.pdf) of the databases.

When you have downloaded the database to your computer, e.g. to your desktop, do the following to install the database:

Mac OS X:

1. Go to your 'Applications' folder and locate the 'CLC Protein Workbench' application

2. Right-click (ctrl-click) the application and choose 'Show Package Contents'
3. Navigate to Contents/Resources/app/data/databases/pfam/
4. Copy the downloaded db-file into this directory
5. Start or restart CLC Protein Workbench

Windows:

1. Locate the installation directory for CLC Protein Workbench, typically C:/Program Files/CLC Protein Workbench
2. Navigate to data/databases/pfam/
3. Copy the downloaded db-file into this directory
4. Start or restart CLC Protein Workbench

Linux:

1. Locate the installation directory for CLC Protein Workbench, typically /opt/clcproteinwb/
2. Navigate to data/databases/pfam/
3. Copy the downloaded db-file into this directory
4. Start or restart CLC Protein Workbench

15.7 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 rodlike structure whereas a beta-sheet shows 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 Protein 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:

Select a protein sequence | Toolbox in the Menu Bar | Protein Analyses (📁) | Predict secondary structure (🌀)

or **right-click a protein sequence | Toolbox | Protein Analyses (📁) | Predict secondary structure (🌀)**

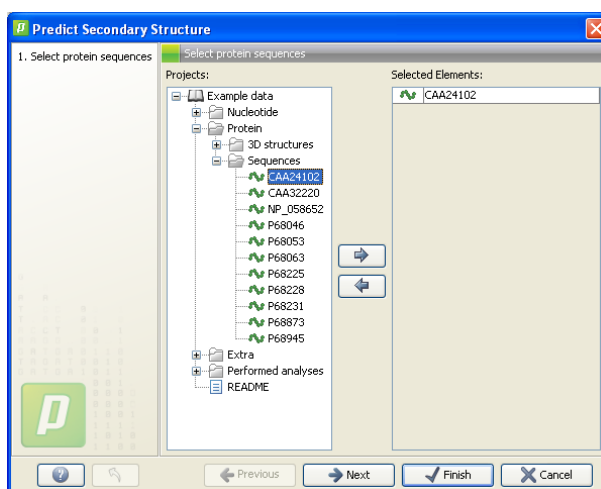


Figure 15.19: Choosing one or more protein sequences for secondary structure prediction.

This opens the dialog displayed in figure 15.19:

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 **Project Tree**.

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.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). 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 15.20).

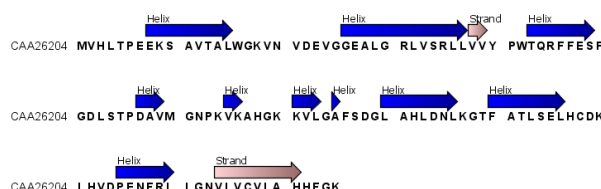


Figure 15.20: Alpha-helices and beta-strands shown as annotations on the sequence.

Each annotation will carry a tooltip note saying that the corresponding annotation is predicted with *CLC Protein Workbench*. Additional notes can be added through the **Add annotation** right-click mouse menu. See section 11.1.4.

Undesired alpha-helices or beta-sheets can be removed through the **Edit annotation** right-click mouse menu. See section 11.1.5.

15.8 Protein report

CLC Protein Workbench is able to produce protein reports, that allow you to easily generate different kinds of information regarding a protein.

Actually a protein report is a collection of some of the protein analyses which are described elsewhere in this manual.

To create a protein report do the following:

Right-click protein in Navigation Area | Toolbox | Protein Analyses (📁) | Create Protein Report (📄)

This opens dialog **Step 1**, where you can choose which protein to create a report for. Only one protein can be chosen. When the correct one is chosen, click **Next**.

In dialog **Step 2** you can choose which analyses you want to include in the report. The following list shows which analyses are available and explains where to find more details.

- **Sequence statistics.** See section 13.4 for more about this topic.
- **Plot of charge as function of pH.** See section 15.2 for more about this topic.
- **Plot of hydrophobicity.** See section 15.5 for more about this topic.
- **Plot of local complexity.** See section 13.3 for more about this topic.
- **Dot plot against self.** See section 13.1 for more about this topic.
- **Secondary structure prediction.** See section 15.7 for more about this topic.
- **Pfam domain search.** See section 15.6 for more about this topic.
- **SignalP signal peptide prediction.** See section 15.1 for more about this topic.
- **TMHMM transmembrane helix prediction.** See section 15.3 for more about this topic.
- **BLAST against local database.** See section 10.2 for more about this topic.
- **BLAST against NCBI databases.** See section 10.1 for more about this topic.

When you have selected the relevant analyses, click **Next**. **Step 3 to Step 7** (if you select all the analyses in **Step 2**) are adjustments of parameters for the different analyses. The parameters are mentioned briefly in relation to the following steps, and you can turn to the relevant chapters or sections (mentioned above) to learn more about the significance of the parameters.

In **Step 3** you can adjust parameters for sequence statistics:

- **Individual Statistics Layout.** **Comparative** is disabled because reports are generated for one protein at a time.
- **Include Background Distribution of Amino Acids.** Includes distributions from different organisms. Background distributions are calculated from UniProt www.uniprot.org version 6.0, dated September 13 2005.

In **Step 4** you can adjust parameters for hydrophobicity plots:

- **Window size.** Width of window on sequence (odd number).
- **Hydrophobicity scales.** Lets you choose between different scales.

In **Step 5** you can adjust a parameter for complexity plots:

- **Window size.** Width of window on sequence (must be odd).

In **Step 6** you can adjust parameters for dot plots:

- **Score model.** Different scoring matrices.
- **Window size.** Width of window on sequence.

In **Step 7** you can adjust parameters for BLAST search:

- **Program.** Lets you choose between different BLAST programs.
- **Database.** Lets you limit your search to a particular database.

15.8.1 Protein report output

An example of Protein report output can be seen in figure 15.21.

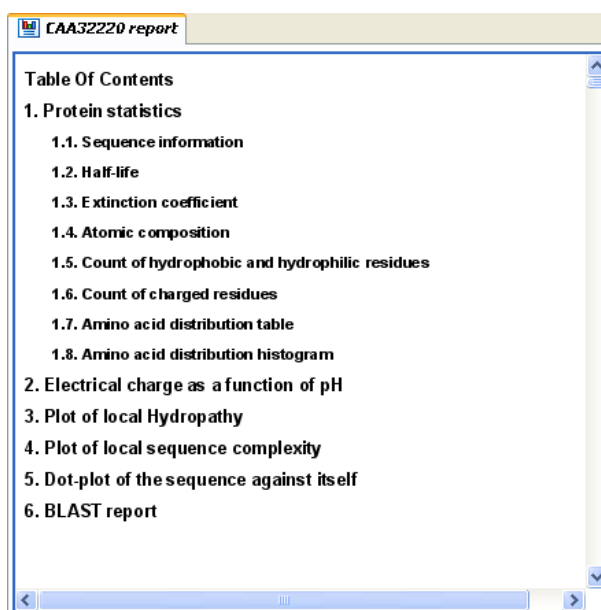


Figure 15.21: A protein report output. The TOC in the Side Panel allows you to easily browse the report.

By double clicking a graph in the output, this graph is shown in a different view (CLC Protein Workbench 2.0 generates another tab). The report output and the new graph views can be saved by dragging the tab into the **Navigation Area**.

The content of the tables in the report can be copy/pasted out of the program and e.g. into Microsoft Excel. To do so:

Select content of table | Right-click the selection | Copy

15.9 Reverse translation from protein into DNA

A protein sequence can be back-translated into DNA using *CLC Protein Workbench 2.0*. Due to degeneracy of the genetic code every amino acid could translate into several different codons (only 20 amino acids but 64 different codons). Thus, the program offers a number of choices for determining which codons should be used. These choices are explained in this section.

In order to make a reverse translation:

Select a protein sequence | Toolbox in the Menu Bar | Protein Analyses (🌿) | Reverse Translate (🔄)

or **right-click a protein sequence | Toolbox | Protein Analyses (🌿) | Reverse translate (🔄)**

This opens the dialog displayed in figure 15.22:

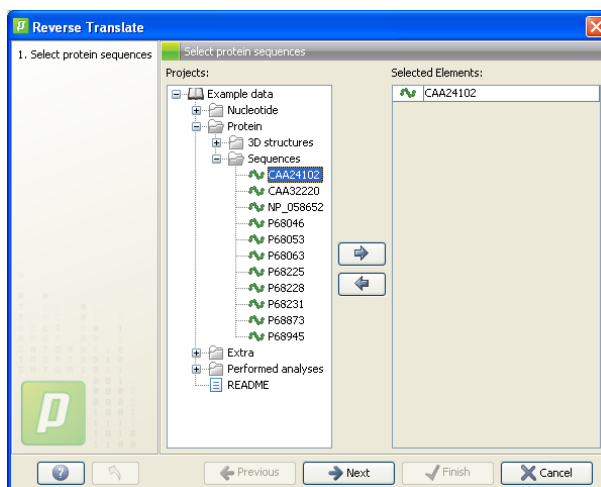


Figure 15.22: Choosing a protein sequence for reverse translation.

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 **Project Tree**. You can translate several protein sequences at a time.

Click **Next** to adjust the parameters for the translation.

15.9.1 Reverse translation parameters

Figure 15.23 shows the choices for making the translation.

- **Most frequently used codon.** On the basis of the selected translation table, this parameter/option will assign the codon that occurs most often. When choosing this option, the results of performing several reverse translations will always be the same, contrary to the following two options.
- **Uniform distribution.** This parameter/option will randomly back-translate an amino acid to a codon without using the translation tables. Every time you perform the analysis you will get a different result.
- **Distribution according to frequency.** This option is a mix of the other two options. The selected translation table is used to attach weights to each codon based on its frequency.

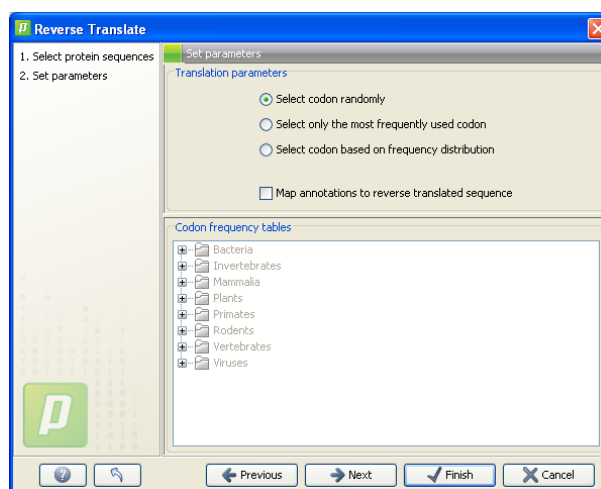


Figure 15.23: Choosing parameters for the reverse translation.

The codons are assigned randomly with a probability given by the weights. A more frequent codon has a higher probability of being selected. Every time you perform the analysis, you will get a different result. This option yields a result that is closer to the translation behavior of the organism (assuming you chose an appropriate codon frequency table).

- **Map annotations to reverse translated sequence.** If this checkbox is checked, then all annotations on the protein sequence will be mapped to the resulting DNA sequence. In the tooltip on the transferred annotations, there is a note saying that the annotation derives from the original sequence.

The **Codon Frequency Table** is used to determine the frequencies of the codons. Select a frequency table from the list that fits the organism you are working with. A translation table of an organism is created on the basis of counting all the codons in the coding sequences. Every codon in a **Codon Frequency Table** has its own count, frequency (per thousand) and fraction which are calculated in accordance with the occurrences of the codon in the organism.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**. The newly created nucleotide sequence is shown, and if the analysis was performed on several protein sequences, there will be a corresponding number of views of nucleotide sequences. The new sequence is not saved automatically. To save the sequence, drag it into the **Navigation Area** or press Ctrl + S (⌘ + S on Mac) to show the save dialog.

15.9.2 Bioinformatics explained: Reverse translation

In all living cells containing hereditary material such as DNA, a transcription to mRNA and subsequent a translation to proteins occur. This is of course simplified but is in general what is happening in order to have a steady production of proteins needed for the survival of the cell. In bioinformatics analysis of proteins it is sometimes useful to know the ancestral DNA sequence in order to find the genomic localization of the gene. Thus, the translation of proteins back to DNA/RNA is of particular interest, and is called reverse translation or back-translation.

The Genetic Code

In 1968 the Nobel Prize in Medicine was awarded to Robert W. Holley, Har Gobind Khorana and Marshall W. Nirenberg for their interpretation of the Genetic Code

(<http://nobelprize.org/medicine/laureates/1968/>). The Genetic Code represents translations of all 64 different codons into 20 different amino acids. Therefore it is no problem to translate a DNA/RNA sequence into a specific protein. But due to the degeneracy of the genetic code, several codons may code for only one specific amino acid. This can be seen in figure 15.24. After the discovery of the genetic code it has been concluded that different organism (and organelles) have genetic codes which are different from the "standard genetic code". Moreover, the amino acid alphabet is no longer limited to 20 amino acids. The 21st amino acid, selenocysteine, is encoded by an 'UGA' codon which is normally a stop codon. The discrimination of a selenocysteine over a stop codon is carried out by the translation machinery. Selenocysteines are very rare amino acids.

The figure 15.24 and 15.25 represents the Standard Code which is the default translation table.

```
AAs = FFLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIMTTTNNKKSS**VVVAAAADDEEGGGG
Starts = -----MMM-----M-----
Base1 = TTTT TTTT TTTT TTTT TTTT TTTT TTTT TTTT TTTT TTTT TTTT TTTT TTTT TTTT TTTT TTTT
Base2 = TTTT TTTT TTTT TTTT TTTT TTTT TTTT TTTT TTTT TTTT TTTT TTTT TTTT TTTT TTTT TTTT
Base3 = TCAG TCAG TCAG TCAG TCAG TCAG TCAG TCAG TCAG TCAG TCAG TCAG TCAG TCAG TCAG TCAG
```

Figure 15.24: The Standard Code for translation.

Second base in codon					
First base in codon					Third base in codon
	U	C	A	G	
	U	Phe	Ser	Tyr	Cys
		Phe	Ser	Tyr	Cys
		Leu	Ser	STOP	STOP
		Leu	Ser	STOP	Trp
	C	Leu	Pro	His	Arg
		Leu	Pro	His	Arg
		Leu	Pro	Gln	Arg
		Leu	Pro	Gln	Arg
	A	Ile	Thr	Asn	Ser
		Ile	Thr	Asn	Ser
		Ile	Thr	Lys	Arg
		Met	Thr	Lys	Arg
	G	Val	Ala	Asp	Gly
		Val	Ala	Asp	Gly
		Val	Ala	Glu	Gly
		Val	Ala	Glu	Gly

Figure 15.25: The standard genetic code showing amino acids for all 64 possible codons.

Challenge of reverse translation

A particular protein follows from the translation of a DNA sequence whereas the reverse translation need not have a specific solution according to the Genetic Code. The Genetic Code is degenerate

which means that a particular amino acid can be translated into more than one codon. Hence there are ambiguities of the reverse translation.

Solving the ambiguities of reverse translation

In order to solve these ambiguities of reverse translation you can define how to prioritize the codon selection, e.g:

- Choose a codon randomly.
- Select the most frequent codon in a given organism.
- Randomize a codon, but with respect to its frequency in the organism.

As an example we want to translate an alanine to the corresponding codon. Four different codons can be used for this reverse translation; GCU, GCC, GCA or GCG. By picking either one by random choice we will get an alanine.

The most frequent codon, coding for an alanine in *E. coli* is GCG, encoding 33.7% of all alanines. Then comes GCC (25.5%), GCA (20.3%) and finally GCU (15.3%). The data are retrieved from the Codon usage database, see below. Always picking the most frequent codon does not necessarily give the best answer.

By selecting codons from a distribution of calculated codon frequencies, the DNA sequence obtained after the reverse translation, holds the correct (or nearly correct) codon distribution. It should be kept in mind that the obtained DNA sequence is not necessarily identical to the original one encoding the protein in the first place, due to the degeneracy of the genetic code.

In order to obtain the best possible result of the reverse translation, one should use the codon frequency table from the correct organism or a closely related species. The codon usage of the mitochondrial chromosome are often different from the native chromosome(s), thus mitochondrial codon frequency tables should only be used when working specifically with mitochondria.

Other useful resources

The Genetic Code at NCBI:

<http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c>

Codon usage database:

<http://www.kazusa.or.jp/codon/>

Wikipedia on the genetic code

http://en.wikipedia.org/wiki/Genetic_code

Creative Commons License

All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in it's original form and "CLC bio" has to be clearly labelled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, or build upon this work.



See <http://creativecommons.org/licenses/by-nc-nd/2.5/> for more about how you may use the contents.

15.10 Proteolytic cleavage detection

CLC Protein Workbench 2.0 offers to analyze protein sequences with respect to cleavage by a selection of proteolytic enzymes. This section explains how to adjust the detection parameters and offers basic information on proteolytic cleavage in general.

15.10.1 Proteolytic cleavage parameters

Given a protein sequence, *CLC Protein Workbench 2.0* detects proteolytic cleavage sites in accordance with detection parameters and shows the detected sites as annotations on the sequence and in textual format in a table below the sequence view.

Detection of proteolytic cleavage sites is initiated by:

right-click a protein sequence in Navigation Area | Toolbox | Protein Analyses (📁)
| Proteolytic Cleavage, (✂️)

This opens the dialog shown in figure 15.26:

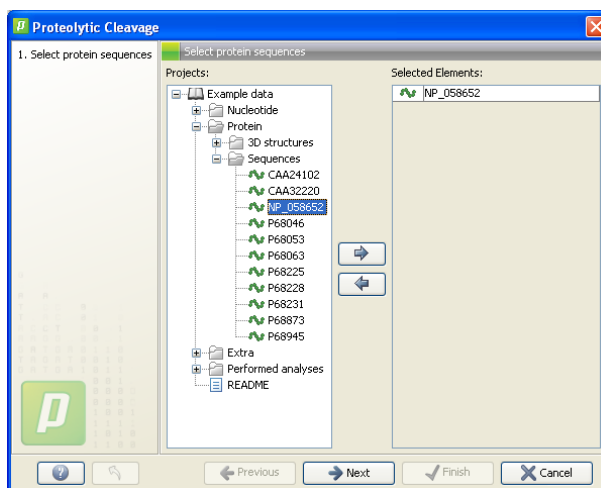


Figure 15.26: Choosing sequence CAA32220 for proteolytic cleavage.

CLC Protein Workbench 2.0 allows you to detect proteolytic cleavages for several sequences at a time. Correct the list of sequences by selecting a sequence and clicking the arrows pointing left and right. Then click **Next** to go to **Step 2**.

In **Step 2** you can select proteolytic cleavage enzymes. The list of available enzymes will be expanded continuously. Presently, the list contains the enzymes shown in figure 15.27. The full list of enzymes and their cleavage patterns can be seen in Appendix, section C.

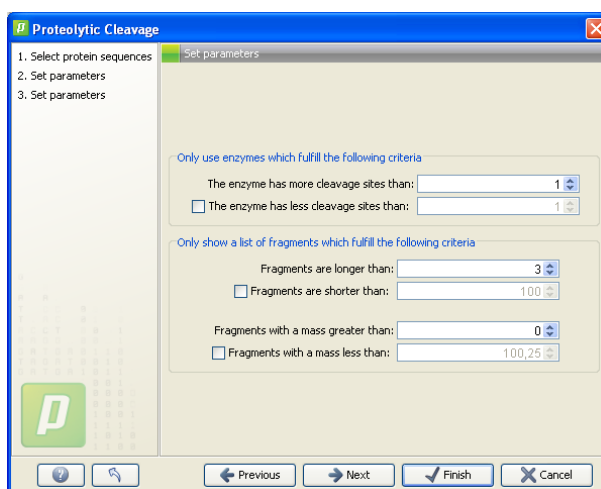


Figure 15.27: Setting parameters for proteolytic cleavage detection.

Select the enzymes you want to use for detection. When the relevant enzymes are chosen, click **Next**.

In **Step 3** you can set parameters for the detection. This limits the number of detected cleavages. Figure 15.28 shows an example of how parameters can be set.

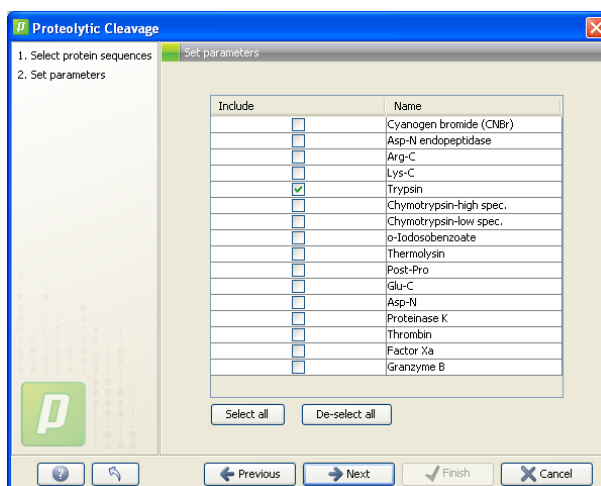


Figure 15.28: Setting parameters for proteolytic cleavage detection.

- **Exclude enzymes based on the number of matches.** Certain proteolytic enzymes cleave at many positions in the amino acid sequence. For instance proteinase K cleaves at nine different amino acids, regardless of the surrounding residues. Thus, it can be very useful to limit the number of actual cleavage sites before running the analysis.
- **Exclude fragments based on length** Likewise, it is possible to limit the output to only display sequence fragments between a chosen length. Both a lower and upper limit can be chosen.
- **Exclude fragments based on mass** The molecular weight is not necessarily directly correlated to the fragment length as amino acids have different molecular masses. For that reason it is also possible to limit the search for proteolytic cleavage sites to mass-range.

Example!: If you have one protein sequence but you only want to show which enzymes cut between two and four times. Then you should select "The enzymes has more cleavage sites than 2" and select "The enzyme has less cleavage sites than 4". In the next step you should simply select all enzymes. This will result in a view where only enzymes which cut 2,3 or 4 times are presented.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**.

The result of the detection is displayed in figure 15.29.

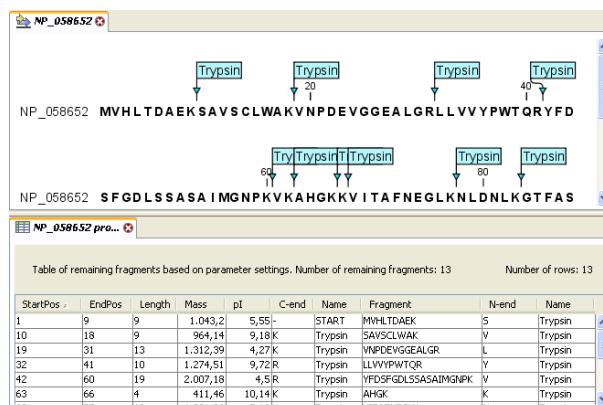


Figure 15.29: The result of the proteolytic cleavage detection.

Depending on the settings in the program, the output of the proteolytic cleavage site detection will display two views on the screen. The top view shows the actual protein sequence with the predicted cleavage sites indicated by small arrows. If no labels are found on the arrows they can be enabled by setting the labels in the "annotation layout" in the preference panel. The bottom view shows a text output of the detection, listing the individual fragments and information on these.

15.10.2 Bioinformatics explained: Proteolytic cleavage

Proteolytic cleavage is basically the process of breaking the peptide bonds between amino acids in proteins. This process is carried out by enzymes called peptidases, proteases or proteolytic cleavage enzymes.

Proteins often undergo proteolytic processing by specific proteolytic enzymes (proteases/peptidases) before final maturation of the protein. Proteins can also be cleaved as a result of intracellular processing of, for example, misfolded proteins. Another example of proteolytic processing of proteins is secretory proteins or proteins targeted to organelles, which have their signal peptide removed by specific signal peptidases before release to the extracellular environment or specific organelle.

Below a few processes are listed where proteolytic enzymes act on a protein substrate.

- N-terminal methionine residues are often removed after translation.
- Signal peptides or targeting sequences are removed during translocation through a membrane.
- Viral proteins that were translated from a monocistronic mRNA are cleaved.

- Proteins or peptides can be cleaved and used as nutrients.
- Precursor proteins are often processed to yield the mature protein.

Proteolytic cleavage of proteins has shown its importance in laboratory experiments where it is often useful to work with specific peptide fragments instead of entire proteins.

Proteases also have commercial applications. As an example proteases can be used as detergents for cleavage of proteinaceous stains in clothing.

The general nomenclature of cleavage site positions of the substrate were formulated by Schechter and Berger, 1967-68 [Schechter and Berger, 1967], [Schechter and Berger, 1968]. They designate the cleavage site between P1-P1', incrementing the numbering in the N-terminal direction of the cleaved peptide bond (P2, P3, P4, etc..). On the carboxyl side of the cleavage site the numbering is incremented in the same way (P1', P2', P3' etc.). This is visualized in figure 15.30.

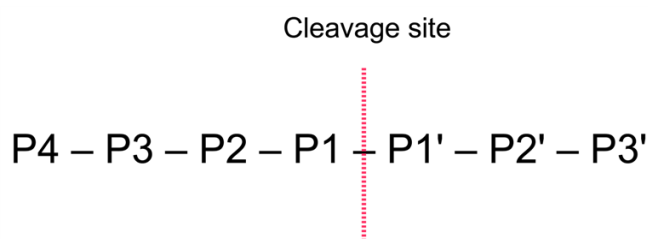


Figure 15.30: *Nomenclature of the peptide substrate. The substrate is cleaved between position P1-P1'.*

Proteases often have a specific recognition site where the peptide bond is cleaved. As an example trypsin only cleaves at lysine or arginine residues, but it does not matter (with a few exceptions) which amino acid is located at position P1'(carboxyterminal of the cleavage site). Another example is trypsin which cleaves if an arginine is found in position P1, but not if a D or E is found in position P1' at the same time. (See figure 15.31).

Bioinformatics approaches are used to identify potential peptidase cleavage sites. Fragments can be found by scanning the amino acid sequence for patterns which match the corresponding cleavage site for the protease. When identifying cleaved fragments it is relatively important to know the calculated molecular weight and the isoelectric point.

Other useful resources

The Peptidase Database: <http://merops.sanger.ac.uk/>

Creative Commons License

All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in it's original form and "CLC bio" has to be clearly labelled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, or build upon this work.

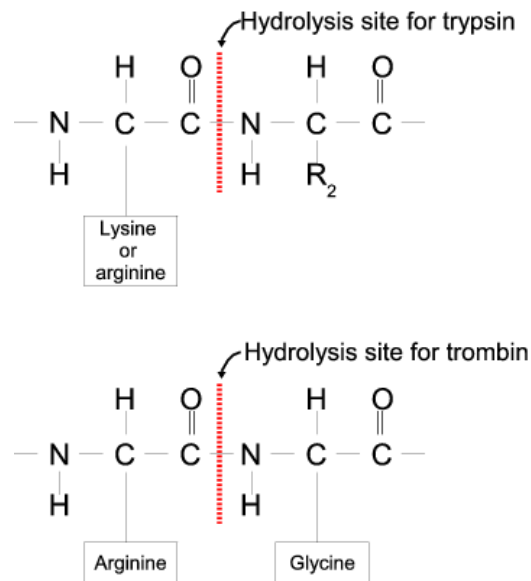


Figure 15.31: Hydrolysis of the peptide bond between two amino acids. Trypsin cleaves unspecifically at lysine or arginine residues whereas trombin cleaves at arginines if asparate or glutamate is absent.



See <http://creativecommons.org/licenses/by-nc-nd/2.5/> for more about how you may use the contents.

Chapter 16

Restriction site analyses

Contents

16.1 Restriction sites and enzyme lists	206
16.2 Restriction site analysis	206
16.2.1 Restriction site parameters	206
16.3 Restriction enzyme lists	210
16.3.1 Create enzyme list	210
16.3.2 Modify enzyme list	211
16.4 Gel electrophoresis	212
16.4.1 Separate sequences on gel	212
16.4.2 Separate fragments of sequences using restriction enzymes	213
16.4.3 Gel view	213

16.1 Restriction sites and enzyme lists

CLC Protein Workbench 2.0 offers the opportunity to detect restriction sites. First the restriction site analysis is described and next, the functionalities regarding enzyme lists are explained.

16.2 Restriction site analysis

This section explains how to adjust the detection parameters and offers basic information with respect to restriction site algorithms.

16.2.1 Restriction site parameters

Given a DNA sequence, *CLC Protein Workbench 2.0* detects restriction sites in accordance with detection parameters and shows the detected sites as annotations on the sequence or in textual format in a table.

To detect restriction sites:

select sequence | **Toolbox in the Menu Bar** | **Restriction Site Analyses** (🔧) | **Restriction sites** (✂️)

or **right-click sequence** | **Toolbox** | **Restriction Site Analyses** (🔧) | **Restriction sites** (✂️)

The result of these steps can be seen in figure 16.1.

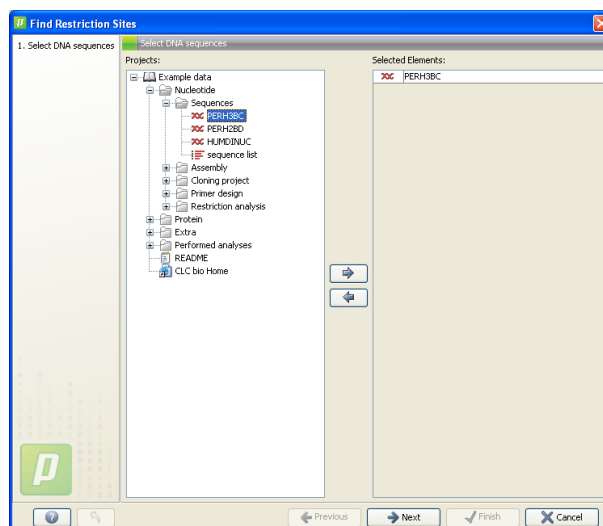


Figure 16.1: Choosing sequence PERH3BC.

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 **Project Tree**.

Clicking **Next** generates the dialog shown in figure 16.2.

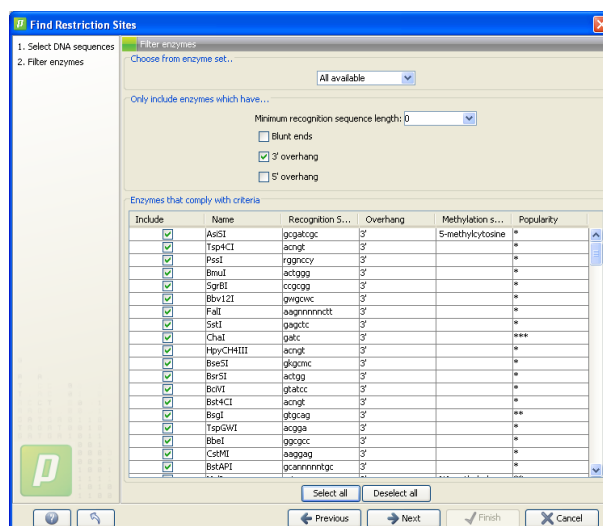


Figure 16.2: Selecting enzymes.

In **Step 2** you can adjust which enzymes to use. **Choose from enzyme set...**, allows you to select an enzyme list which is stored in the **Navigation Area**. See section 16.3 for more about creating and modifying enzyme lists.

Only include enzymes which have... In this part of the dialog, you can limit the number of

enzymes included in the list below. You can choose a minimum length of the recognition sequence, and you can choose whether to include enzymes with Blunt ends, 3' overhang, and/or 5' overhang.

Having adjusted the parameters in **Choose from enzyme set...** and **Only include enzymes which have...** the total list of enzymes is shown in the table. The enzymes can be sorted by clicking the column headings, and you can select which enzymes to include in the search by inserting / removing check marks next to the enzymes.

Clicking **Next** confirms the list of enzymes which will be included in the analysis, and takes you to **Step 3**.

In **Step 3** you can limit which enzymes' cut sites should be included in the output. See figure 16.3.

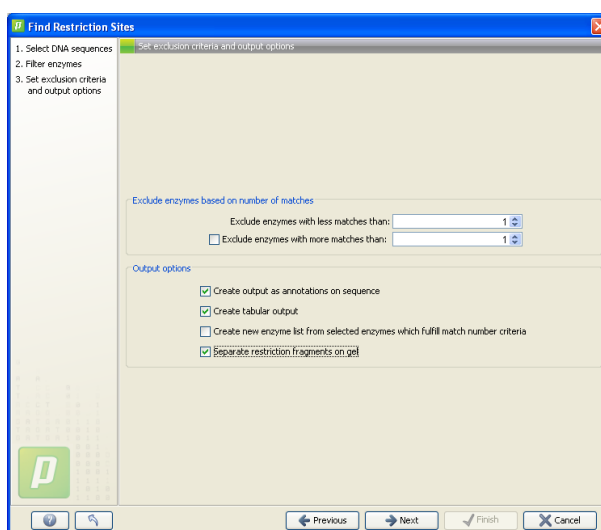


Figure 16.3: Exclusion criteria and output options.

The default setting **Exclude enzymes with less than 1** (matches), means that enzymes which do not match at all, are not included in the output. If e.g. you only want to see enzymes, which match exactly once, you can check the **Exclude enzymes with more than 1**.

The remaining options relate to the output of the analysis:

- Create output as annotations on sequence
- Create text output
- Create new enzyme list from selected enzymes which fulfill match number criteria
- Separate restriction fragments on gel

If you select the last output option (Separate restriction fragments on gel), there will be one more step. If you have chosen this option, click **Next** to see the dialog shown in figure 16.4.

Here you have four different ways of simulating a gel electrophoresis using the selected restriction enzymes:

- **Cut with selected enzymes and run in one lane.** This will display one lane with a number of bands corresponding to the number of fragments from cutting with the selected enzymes.

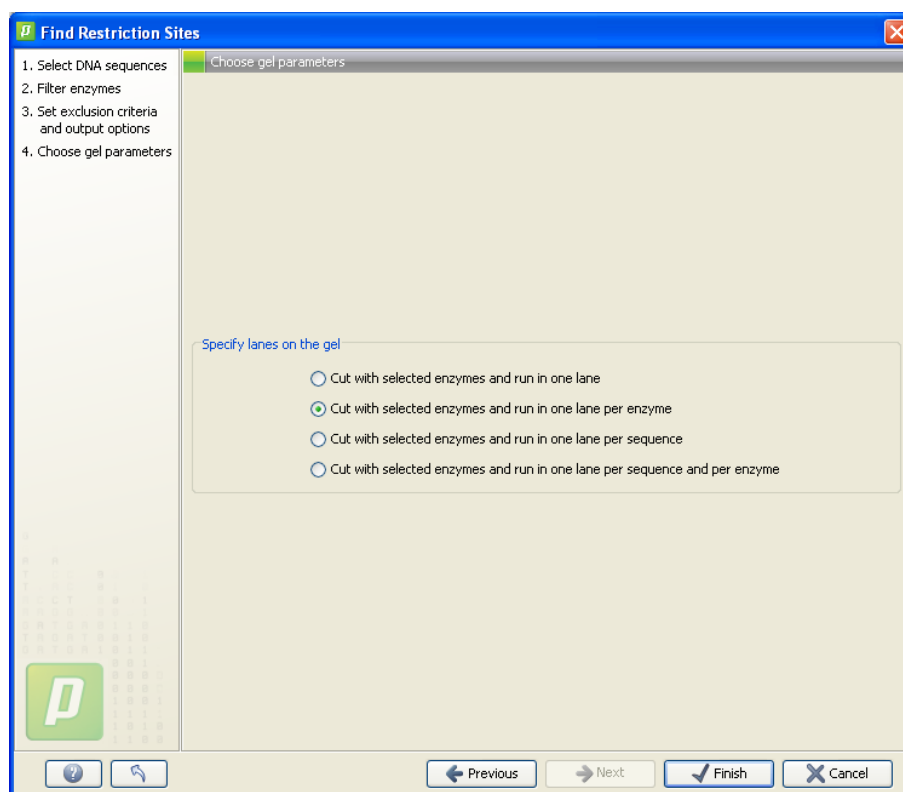


Figure 16.4: Choosing from four different ways of doing gel electrophoresis.

- **Cut with selected enzymes and run in one lane per enzyme.** For each of the enzymes selected, there will be a lane displaying the bands of the fragments from cutting just with this enzyme.
- **Cut with selected enzymes and run in one lane per sequence.** If you have selected more than one sequence, this option will display one lane per sequence in the same way as the first option
- **Cut with selected enzymes and run in one lane per sequence and per enzyme.** This will display a number of lanes equalling the number of selected sequences multiplied by the number of selected enzymes, thus combining the functionality of option number two and three.

For more information about gel electrophoresis, see section [16.4](#).

In order to complete the analysis click **Finish**. The result is shown in figure [16.5](#).

Choosing the textual output option will open a new view containing a table with an overview of restriction sites. Choosing the graphical output option will add restriction site annotations to the selected sequence.

If too many restriction sites are found, a dialog will ask if you want to proceed or show the restriction sites only in a table format. Showing too many restriction sites as annotations on the sequence will take up a lot of your computer's processing power.

Notice! The text is not automatically saved.

To save the result:

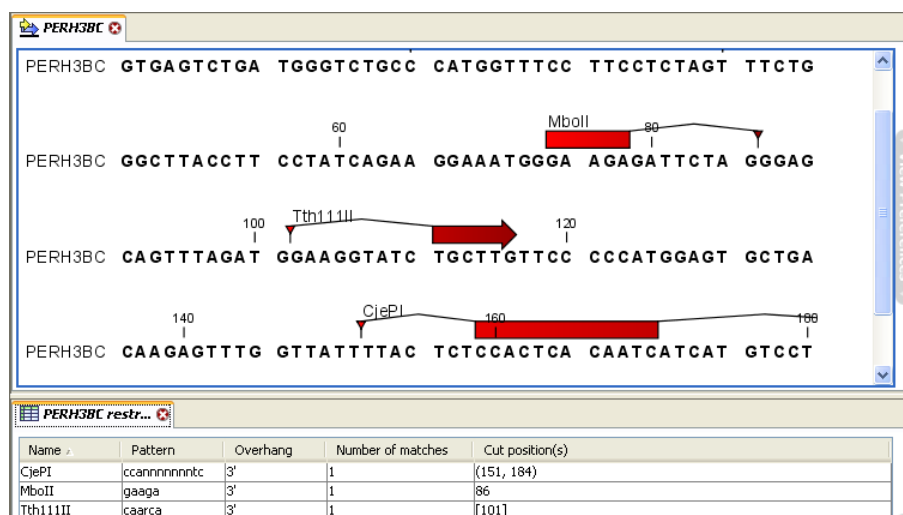


Figure 16.5: The result of the restriction site detection is displayed as text, and in this example the View Shares the View Area with a View of the PERH3BC sequence displaying the restriction sites (split-screen-view).

Right-click the tab | File | Save()

The textual output mentioned above will list all the cut positions where the sequence is restricted. This list may be very long, and hence it might not be possible for *CLC Protein Workbench* to display all cut positions in one cell. If you want to see the entire list of cut positions:

select the table line with the relevant enzyme | Ctrl + C (⌘ + C on Mac) | open a word processing program | Ctrl + V (⌘ + V on Mac)

16.3 Restriction enzyme lists

CLC Protein Workbench includes all the restriction enzymes available in the **REBASE** database. However, when performing restriction site analyses, it is often an advantage to use a customized list of enzymes. In this the user can create special lists containing e.g. all enzymes available in the laboratory freezer, all enzymes used to create a given restriction map or all enzymes that are available from the preferred vendor.

This section describes how you can create an enzyme list, and how you can modify it.

16.3.1 Create enzyme list

CLC Protein Workbench 2.0 uses enzymes from the **REBASE** restriction enzyme database at <http://rebase.neb.com>.

To start creating a sequence list:

right-click in the Navigation Area | New | Enzyme list()

This opens the dialog shown in figure 16.6

Step 1 includes two tables. The top table is a list of all the enzymes available in the **REBASE** database. Different information is available for the enzymes, and by clicking the column headings the list can be sorted.

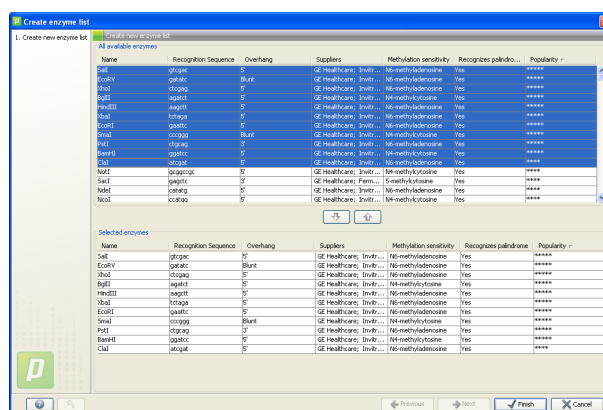


Figure 16.6: Choosing enzymes for the new enzyme list.

The sequence list is created by adding enzymes to the bottom table. To create sequence list:

Select sequences from top table (hold ctrl (⌘ on Mac)) | click down-arrow

When the desired enzymes have been chosen, click **Next**.

Choose where to save your enzyme list and name the sequence list. Click **Finish**, to see the enzyme list. In the View preferences it is possible to choose which column to display.

16.3.2 Modify enzyme list

If you want to make changes to an existing enzyme list:

select an enzyme list | Toolbox in the Menu Bar | Restriction Site Analyses (🔍) | Modify Enzyme List (🔧)

Select the Enzyme list and click **Next**. This opens the dialog shown in figure 16.7.

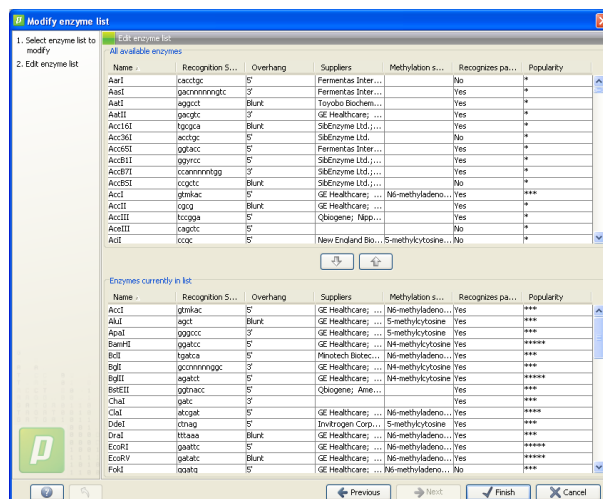


Figure 16.7: Adding and removing enzymes in the existing enzyme list.

Select sequences in either top or bottom table (see 16.3.1). Use the arrows to add and remove sequences. Click **Finish** to see the modified list.

16.4 Gel electrophoresis

CLC Protein Workbench enables the user to simulate the separation of nucleotide sequences on a gel. This feature is useful when e.g. designing an experiment which will allow the differentiation of a successful and an unsuccessful cloning experiment on the basis of a restriction map.

There are two main ways to simulate gel separation of nucleotide sequences:

- A number of existing sequences can be separated on a gel.
- One or more sequences can be digested with restriction enzymes and the resulting fragments can be separated on a gel.

There are several ways to apply these functionalities as described below.

16.4.1 Separate sequences on gel

This section explains how to simulate a gel electrophoresis of one or more existing sequences without restriction enzymes digestion:

select one or more sequences | Toolbox | Restriction Site Analyses () | Separate Sequences on Gel ()

This opens the dialog shown in figure 16.8.

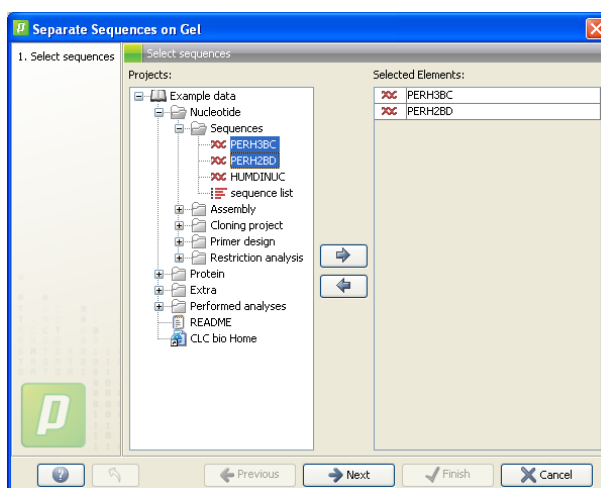


Figure 16.8: Select one or more sequences to separate on a gel.

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 **Project Tree**.

Clicking **Next** generates the dialog shown in figure 16.9.

In this dialog, you can choose from two different ways of simulating the gel electrophoresis:

- **Run each sequence in a separate lane.** This will create a new lane for each of the selected sequences. As a result, there will only be one band on each lane.
- **Run all sequences in same lane.** This will create only one lane in which each of the selected sequences will be represented by a band.



Figure 16.9: Choosing how to display the lanes.

The difference between these two options is shown in figure 16.10. Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**.

For more information about the view of the gel, see section 16.4.3.

16.4.2 Separate fragments of sequences using restriction enzymes

This section explains how to simulate a gel electrophoresis of one or more sequences which are digested with restriction enzymes. There are two ways to do this:

- When performing the **Restriction Sites** analysis from the **Toolbox**, you can choose to separate the resulting fragments on a gel. This is explained in section 16.2.1.
- From all the graphical views of sequences, you can right-click the label of the sequence and choose: **Digest Sequence with Selected Enzymes and Run on Gel**. The views where this option is available are listed below:
 - Circular view (see section 11.6).
 - Ordinary sequence view (see section 11.1).
 - Graphical view of sequence lists (see section 11.5).

Furthermore, you can also right-click an empty part of the view of the graphical view of sequence lists and choose **Digest All Sequences with Selected Enzymes and Run on Gel**. This opens a dialog with functionalities similar to the one in figure 16.4.

Notice! When using the right-click options, the sequence will be digested with the enzymes that are selected in the **Side Panel**. This is explained in section 11.1.1.

16.4.3 Gel view

In figure 16.11 you can see a simulation of a gel with its **Side Panel** to the right. This view will be explained in this section.

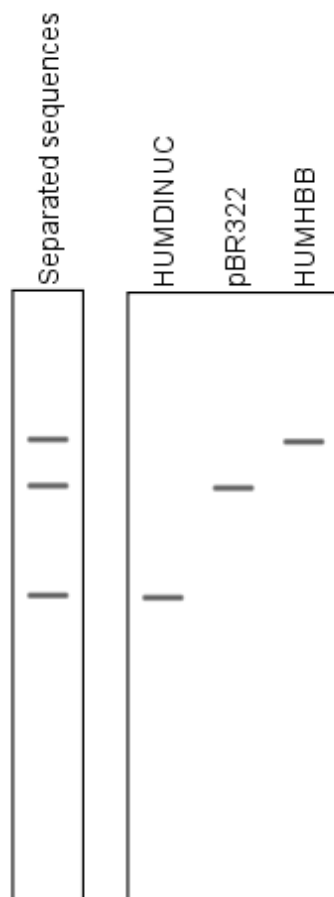


Figure 16.10: Gel electrophoresis of three sequences. The left side shows the sequences together in one lane, each represented by a band. The right side shows a lane for each sequence.

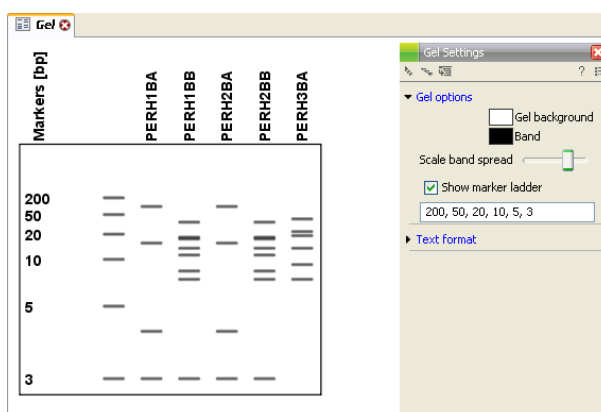


Figure 16.11: Five lanes showing fragments of five sequences cut with restriction enzymes.

Information on bands and fragment size

You can get information about the individual bands by hovering the mouse cursor on the band of interest. This will display a tool tip with information about the fragment size, and for lanes comparing whole sequences, you will also see the sequence name.



Notice! You have to be in **Selection** () or **Pan** () mode in order to get this information.

It can be useful to add markers to the gel which enables you to compare the sizes of the bands. This is done by clicking **Show marker ladder** in the **Side Panel**.

You enter the markers by writing them in the text field, separated by commas.

Modifying the layout

The background of the lane and the colors of the bands can be changed in the **Side Panel**. Click the colored box to display a dialog for picking a color. The slider **Scale band spread** can be used to adjust the effective time of separation on the gel, i.e. how much the bands will be spread over the lane. In a real electrophoresis experiment this property will be determined by several factors including time of separation, voltage and gel density.

You can also modify the layout of the view by zooming in or out. Click **Zoom in** () or **Zoom out** () in the Toolbar and click the view.

Finally, you can modify the format of the text heading each lane in the **Text format** preferences in the **Side Panel**.

Chapter 17

Sequence alignment

Contents

17.1 Create an alignment	217
17.1.1 Gap costs	218
17.1.2 Fast or accurate alignment algorithm	218
17.1.3 Aligning alignments	219
17.1.4 Fixpoints	220
17.2 View alignments	222
17.2.1 Sequence logo	223
17.2.2 Conservation	224
17.2.3 Gap fraction	225
17.3 Edit alignments	225
17.3.1 Move residues and gaps	225
17.3.2 Insert gap columns	225
17.3.3 Delete residues and gaps	226
17.3.4 Copy annotations to other sequences	226
17.3.5 Move sequences up and down	226
17.3.6 Delete and add sequences	226
17.3.7 Realign selection	227
17.4 Join alignments	227
17.4.1 How alignments are joined	228
17.5 Bioinformatics explained: Multiple alignments	229
17.5.1 Use of multiple alignments	229
17.5.2 Constructing multiple alignments	230

CLC Protein Workbench 2.0 can align nucleotides and proteins using a *progressive alignment* algorithm (see section 17.5 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.

17.1 Create an alignment

Alignments can be created from sequences, sequence lists (see section 11.5), existing alignments and from any combination of the three.

To create an alignment in *CLC Protein Workbench 2.0*:

select elements to align | Toolbox in the Menu Bar | Alignments and Trees(📁) | Create Alignment (🔍)

or **select elements to align | right-click either selected sequence | Toolbox | Alignments and Trees(📁) | Create Alignment (🔍)**

This opens the dialog shown in figure 17.1.

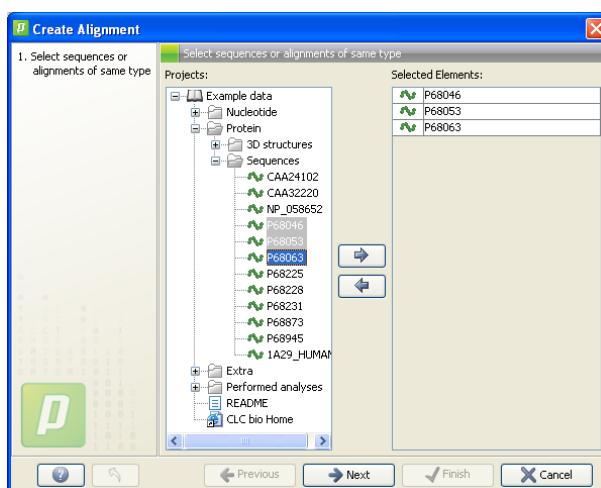


Figure 17.1: Creating an alignment.

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 **Project Tree**. Click **Next** to adjust alignment algorithm parameters. Clicking **Next** opens the dialog shown in figure 17.2.

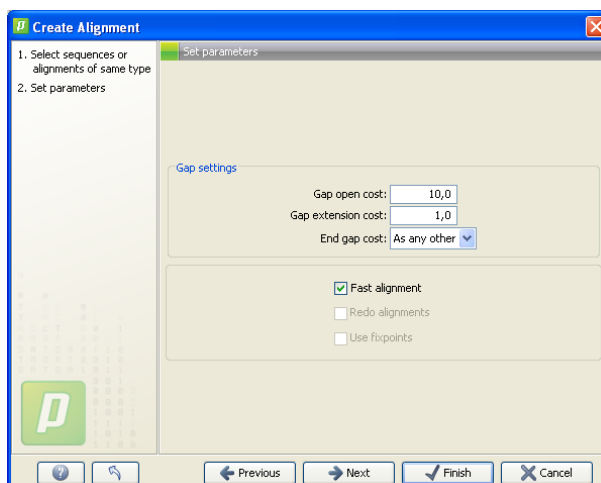


Figure 17.2: Adjusting alignment algorithm parameters.

17.1.1 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 Protein Workbench 2.0* 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 17.3 and 17.4 illustrate the differences between the different gap scores at the sequence ends.

17.1.2 Fast or accurate alignment algorithm

CLC Protein Workbench has two algorithms for calculating alignments:

- **Accurate alignment.** This is the recommended choice unless you find the processing time too long.
- **Fast alignment.** This allows for use of an optimized alignment algorithm which is very fast. The fast option is particularly useful for datasets with very long sequences.

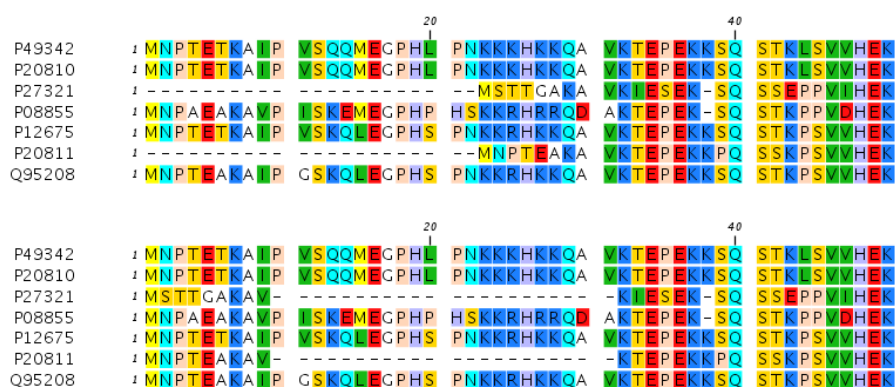


Figure 17.3: The first 50 positions of two different alignments of seven calpastatin sequences. The top alignment is made with cheap end gaps, while the bottom alignment is made with end gaps having the same price as any other gaps. In this case it seems that the latter scoring scheme gives the best result.

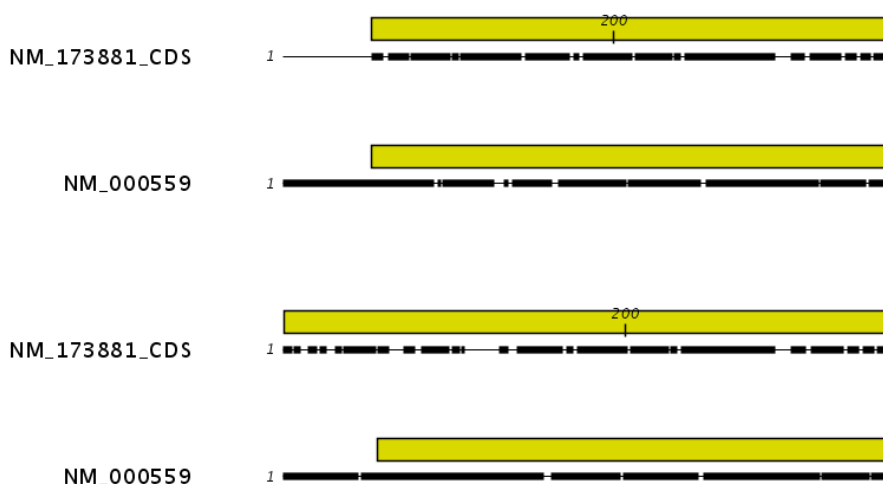


Figure 17.4: The alignment of the coding sequence of bovine myoglobin with the full mRNA of human gamma globin. The top alignment is made with free end gaps, while the bottom alignment is made with end gaps treated as any other. The yellow annotation is the coding sequence in both sequences. It is evident 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.

For a comprehensive explanation of the alignment algorithms see section 17.5.

17.1.3 Aligning alignments

If you have selected an existing alignment in the first step (17.1), 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 17.5.

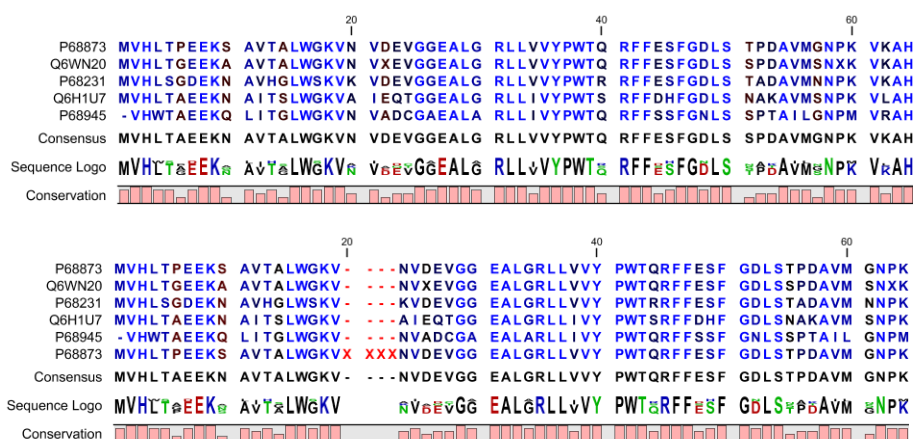


Figure 17.5: The top figures shows the original alignment. In the bottom panel a single sequence with four inserted X's are aligned to the original alignment. This introduces gaps in all sequences of the original alignment. All other positions in the original alignment are fixed.

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.

17.1.4 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.

Fixpoints are added to sequences or alignments before clicking "Create alignment". 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 17.6). Use this procedure to add fixpoints to the other sequence(s) that should be forced to align to each other.

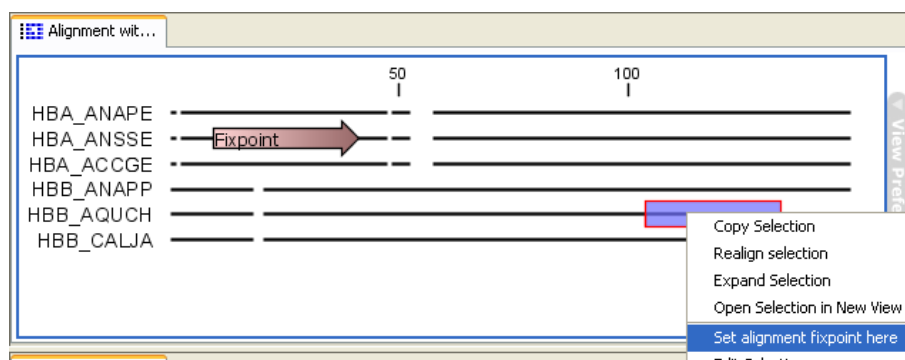


Figure 17.6: Adding a fixpoint to a sequence in an existing alignment. At the top you can see a fixpoint that has already been added.

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 17.7 the result of an alignment using fixpoints is illustrated.

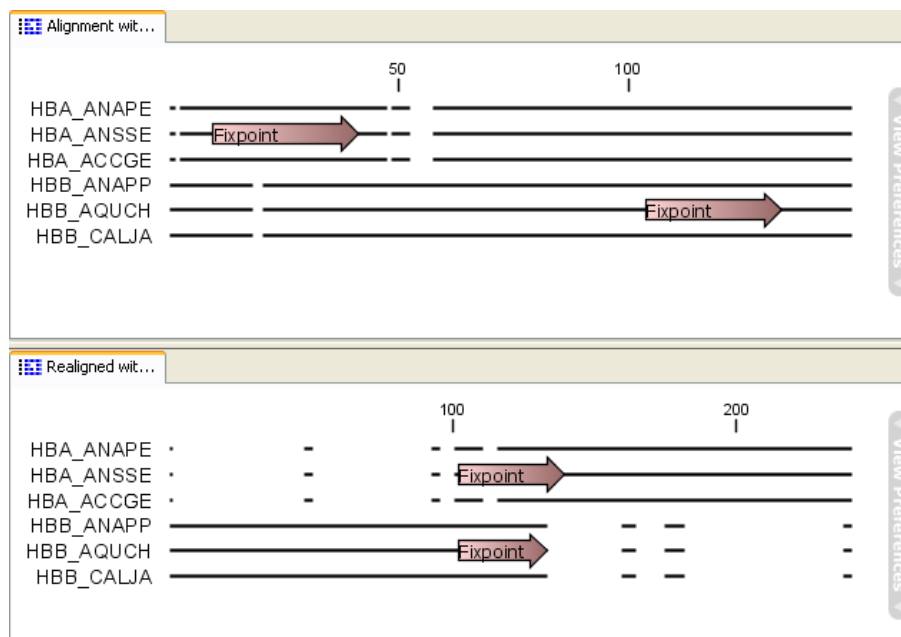


Figure 17.7: *Realigning using fixpoints.* In the top view, fixpoints have been added to two of the sequences. In the view below, the alignment has been realigned using the fixpoints. The three top sequences are very similar, and therefore they follow the one sequence (number two from the top) that has a fixpoint.

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 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

17.2 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 [11.1](#) for an explanation of these basic options.

However, there are a number of alignment-specific view options in the **Alignment info** preference group in the **Side Panel** to the right of the view. These preferences relate to each column in the alignment. Below is more information on these view options.

- **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 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**.

- **Limit.** This option determines how conserved the sequences must be in order to agree on a consensus.
- **No gaps.** Checking this option will not show gaps in the consensus.
- **Ambiguous symbol.** Select how ambiguities should be displayed in the consensus line.
- **Sequence logo.** See section [17.2.1](#) for more details.
 - **Foreground color.** Colors the letters using a gradient according to the information content of the alignment column.
 - **Background color.** Sets a background color of the residues using a gradient in the same way as described above.
 - **Graph.** 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.
- **Conservation.** Displays the level of conservation at each position in the alignment.
 - **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.

- * **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.
 - **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.
 - * **Height.** Specifies the height of the graph.
 - * **Type.** The type of the graph.
 - * **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.

17.2.1 Sequence logo

Below the alignment there is an option of displaying a "sequence logo" (shown as default). The sequence logo displays the information content of all positions in the alignment as residues or nucleotides stacked on top of each other (see figure 17.8). The sequence logo provides a far more detailed view of the alignment than the conservation view (see section 17.2.2). Sequence logos can aid to identify protein binding sites on DNA sequences but 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, show 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 for instance an alignment where in one position a particular residue is found in 70% of the sequences. If a consensus sequence would be defined it typically only displays the single residue with 70% coverage. In figure 17.8 and ungapped 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 the looking at the sequence logo it is seen that a GTG is also allowed as a start codon.

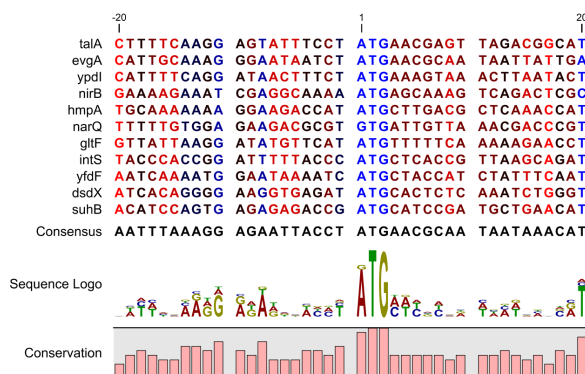


Figure 17.8: Ungapped sequence alignment of eleven *E. coli* sequences defining a start codon. The start codons start at position 1. Below the alignment is shown the corresponding sequence logo. As seen, a GTG start codon and the usual ATG start codons are present in the alignment. This can also be visualized 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 an 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 height of 0.1.

Other useful resources

The website of Tom Schneider

<http://www-lmm.b.ncifcrf.gov/~toms/>

WebLogo

<http://weblogo.berkeley.edu/> [Crooks et al., 2004]

17.2.2 Conservation

The conservation view is very simplified view compared to the sequence logo view as described above. The bar (default view) show the conservation of all sequence positions. The height of

the bars in the view reflects how conserved that particular position is in the alignment. If one position is 100% conserved the bar will be shown in full height.

17.2.3 Gap fraction

The gap fraction view show if any gaps are present in the alignment. If a gap is present in the majority of sequences this will be represented in the view.

17.3 Edit alignments

17.3.1 Move residues and gaps

The placement of gaps in the alignment can be changed by modifying the parameters when creating the alignment (see section 17.1). 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 17.9).

Notice! Residues can only be moved when they are next to a gap.

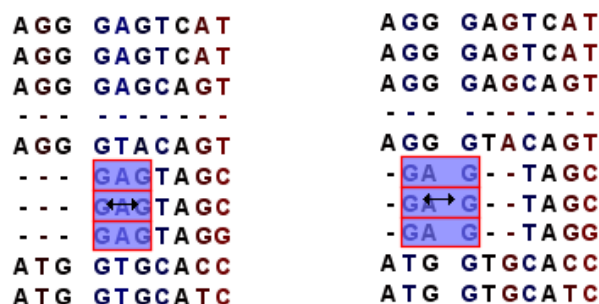


Figure 17.9: Moving a part of an alignment. Notice the change of mouse pointer to a horizontal arrow.

17.3.2 Insert gap columns

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 gap columns (i.e. gaps in all the sequences):

select a part of the alignment | right-click the selection | Add gap columns before/after

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.

17.3.3 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.

To delete entire columns:

select the part of the alignment you want to delete | right-click the selection | Delete columns

The selection may cover one or more sequences, but the **Delete columns** function will always apply to the entire alignment.

17.3.4 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.

17.3.5 Move sequences up and down

Sequences can be moved up and down in the alignment:

drag the label 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 label of a sequence | Sort Sequences Alphabetically

If you change the Sequence label (in the **Sequence Layout** view preferences), you will have to ask the program to sort the sequences again.


The sequences can also be sorted by similarity, grouping similar sequences together:

Right-click the label of a sequence | Sort Sequences by Similarity

17.3.6 Delete 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.

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 17.1).

The same procedure can be used for joining two alignments.

17.3.7 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 17.1).

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.

17.4 Join alignments

CLC Protein Workbench can join several alignments into one. This feature can for example be used to construct "supergenes" for phylogenetic inference by joining alignments of several disjoint genes into one spliced alignment. Note, that when alignments are joined, all their annotations are carried over to the new spliced alignment.

Alignments can be joined by:

select alignments to join | Toolbox in the Menu Bar | Alignments and Trees() | Join Alignments ()

or **select alignments to join | right-click either selected alignment | Toolbox | Alignments and Trees() | Join Alignments ()**

This opens the dialog shown in figure 17.10.

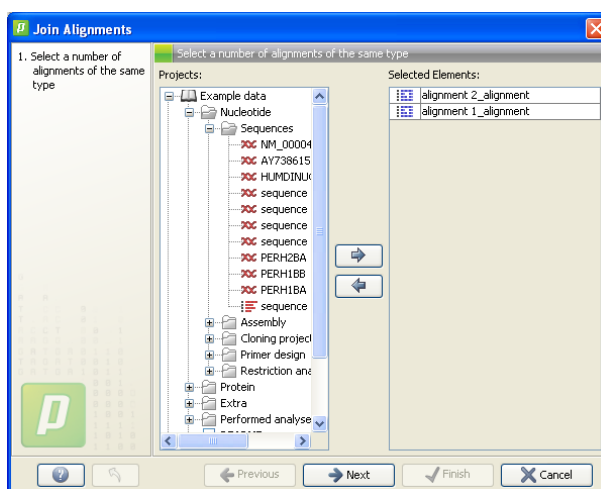


Figure 17.10: Selecting two alignments to be joined.

If you have selected some alignments before choosing the Toolbox action, they are now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove alignments from the **Project Tree**. Click **Next** opens the dialog shown in figure 17.11.

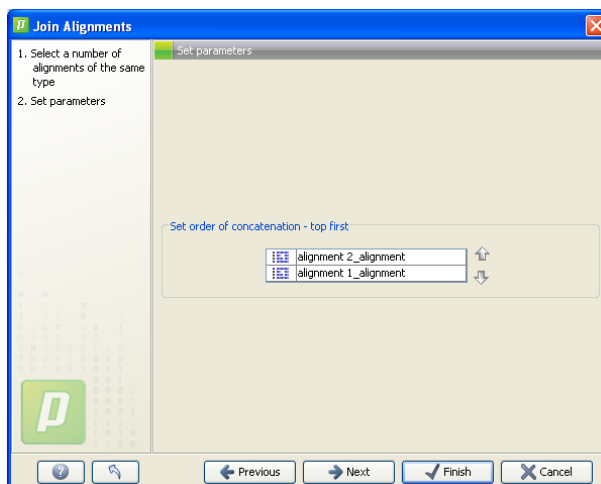


Figure 17.11: Selecting order of concatenation.

To adjust the order of concatenation, click the name of one of the alignments, and move it up or down using the arrow buttons.

Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**. The result is seen in figure 17.12.

17.4.1 How alignments are joined

Alignments are joined by considering the sequence names in the individual alignments. If two sequences from different alignments have identical names, they are considered to have the same origin and are thus joined. Consider the joining of alignments A and B. If a sequence named "in-A-and-B" is found in both A and B, the spliced alignment will contain a sequence named "in-A-and-B" which represents the characters from A and B joined in direct extension of each other. If a sequence with the name "in-A-not-B" is found in A but not in B, the spliced alignment will contain a sequence named "in-A-not-B". The first part of this sequence will contain

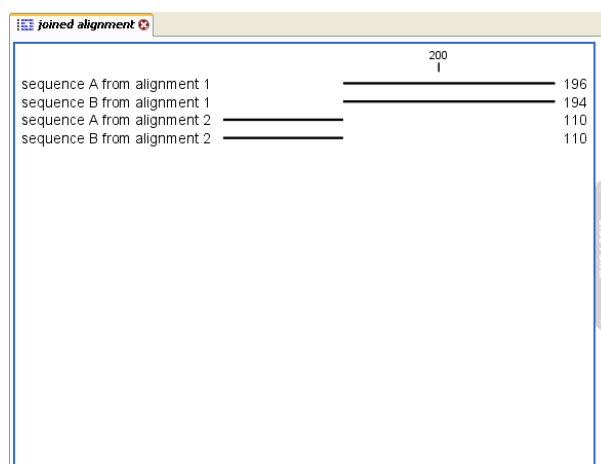


Figure 17.12: The joining of the alignments result in one alignment containing rows of sequences corresponding to the number of uniquely named sequences in the joined alignments.

the characters from A, but since no sequence information is available from B, a number of gap characters will be added to the end of the sequence corresponding to the number of residues in B. Note, that the function does not require that the individual alignments contain an equal number of sequences.

17.5 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 17.13) 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.

17.5.1 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 17.13: 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.

17.5.2 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.

Creative Commons License

All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-

NoDerivs 2.5 License. You are free to to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in it's original form and "CLC bio" has to be clearly labelled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, or build upon this work.



See <http://creativecommons.org/licenses/by-nc-nd/2.5/> for more about how you may use the contents.

Chapter 18

Phylogenetic trees

Contents

18.1 Inferring phylogenetic trees	232
18.1.1 Phylogenetic tree parameters	232
18.1.2 Tree View Preferences	234
18.2 Bioinformatics explained: phylogenetics	235
18.2.1 The phylogenetic tree	236
18.2.2 Modern usage of phylogenies	236
18.2.3 Reconstructing phylogenies from molecular data	237
18.2.4 Interpreting phylogenies	238

CLC Protein Workbench 2.0 offers different ways of inferring phylogenetic trees. The first part of this chapter will briefly explain the different ways of inferring trees in *CLC Protein Workbench 2.0*. The second part, "Bioinformatics explained", will give a more general introduction to the concept of phylogeny and the associated bioinformatics methods.

18.1 Inferring phylogenetic trees

For a given set of aligned sequences (see chapter 17) it is possible to infer their evolutionary relationships. In *CLC Protein Workbench 2.0* this is done by creating a phylogenetic tree:

Toolbox in the Menu Bar | Alignments and Trees | Create Tree ()

or **right-click alignment in Navigation Area | Toolbox | Alignments and Trees | Create Tree ()**

This opens the dialog displayed in figure 18.1:

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.

18.1.1 Phylogenetic tree parameters

Figure 18.2 shows the parameters that can be set:

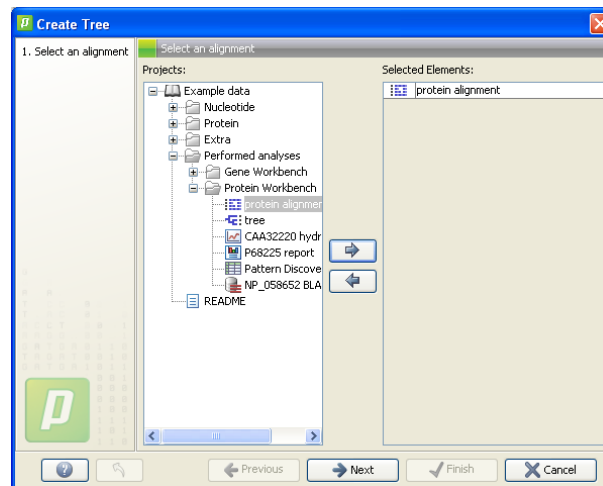


Figure 18.1: Creating a Tree.

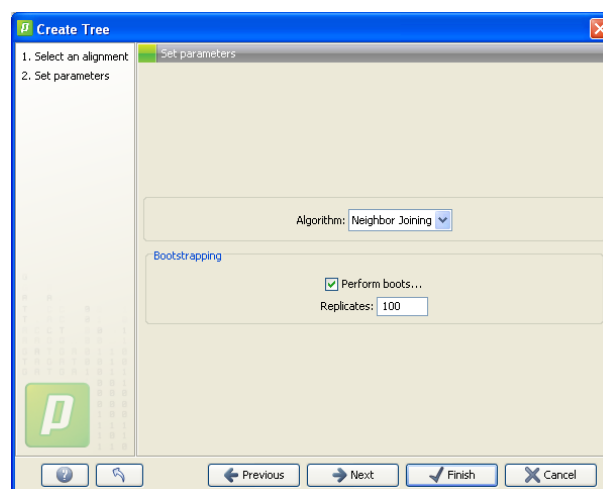


Figure 18.2: Adjusting parameters.

- Algorithms
 - The **UPGMA** method assumes that evolution has occurred at a constant rate in the different lineages. This means that a root of the tree is also estimated.
 - The **neighbor joining** method builds a tree where the evolutionary rates are free to differ in different lineages. *CLC Protein Workbench 2.0* always draws trees with roots for practical reasons, but with the neighbor joining method, no particular biological hypothesis is postulated by the placement of the root. Figure 18.3 shows the difference between the two methods.
- To evaluate the reliability of the inferred trees, *CLC Protein Workbench 2.0* allows the option of doing a **bootstrap** analysis. A bootstrap value will be attached to each branch, and this value is a measure of the confidence in this branch. The number of replicates in the bootstrap analysis can be adjusted in the wizard. The default value is 100.

For a more detailed explanation, see "Bioinformatics explained" in section 18.2.

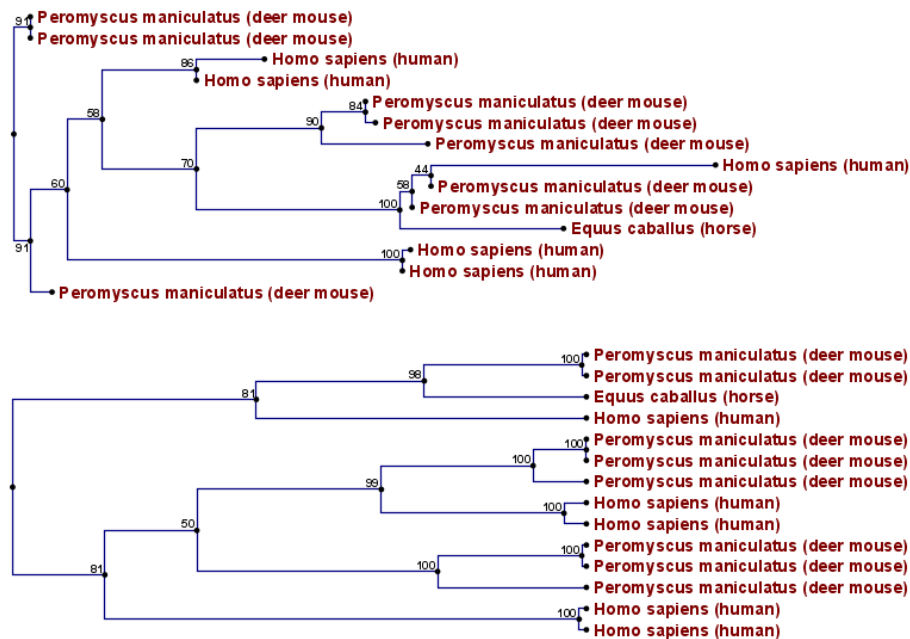


Figure 18.3: *Method choices for phylogenetic inference. The top shows a tree found by neighbor joining, while the bottom shows a tree found by UPGMA. The latter method assumes that the evolution occurs at a constant rate in different lineages.*

18.1.2 Tree View Preferences

The **Tree View** preferences are these:

- **Text format.** Changes the text format for all of the nodes the tree contains.
 - **Text size.** The size of the text representing the nodes can be modified in tiny, small, medium, large or huge.
 - **Font.** Sets the font of the text of all nodes
 - **Bold.** Sets the text bold if enabled.
- **Tree Layout.** Different layouts for the tree.
 - **Node symbol.** Changes the symbol of nodes into box, dot, circle or none if you don't want a node symbol.
 - **Layout.** Displays the tree layout as standard or topology.
 - **Show internal node labels.** This allows you to see labels for the internal nodes. Initially, there are no labels, but right-clicking a node allows you to type a label.
 - **Label color.** Changes the color of the labels on the tree nodes.
 - **Branch label color.** Modifies the color of the labels on the branches.
 - **Node color.** Sets the color of all nodes.
 - **Line color.** Alters the color of all lines in the tree.
- **Annotation Layout.** Specifies the annotation in the tree.

- **Nodes.** Sets the annotation of all nodes either to name or to species.
- **Branches.** Changes the annotation of the branches to bootstrap, length or none if you don't want annotation on branches.

Notice! Dragging in a tree will change it. You are therefore asked if you want to save this tree when the **Tree Viewer** is closed.

You may select part of a **Tree** by clicking on the nodes that you want to select.

Right-click a selected node opens a menu with the following options:

- Set root above node (defines the root of the tree to be just above the selected node).
- Set root at this node (defines the root of the tree to be at the selected node).
- Toggle collapse (collapses or expands the branches below the node).
- Change label (allows you to label or to change the existing label of a node).
- Change branch label (allows you to change the existing label of a branch).

You can also relocate leaves and branches in a tree or change the length.

Notice! To drag branches of a tree, you must first click the node one time, and then click the node again, and this time hold the mouse button.

In order to change the representation:

- Rearrange leaves and branches by
Select a leaf or branch | Move it up and down (Hint: The mouse turns into an arrow pointing up and down)
- Change the length of a branch by
Select a leaf or branch | Press Ctrl | Move left and right (Hint: The mouse turns into an arrow pointing left and right)

Alter the preferences in **Side Panel** for changing the presentation of the tree.

Notice! The preferences will not be saved. Viewing a tree in different viewers gives you the opportunity to change into different preferences in all of the viewers. For example if you select the **Annotation Layout** species for a node then you will only see the change in the specified view. If you now move leaves, the leaves in all views are moved. The options of the right-click pop up menu are changing the tree and therefore they change all views.

Notice! The **Set Root Above** and the **Set Root Here** functions change the tree, and therefore you may save it in order to be able to see it in this format later on.

18.2 Bioinformatics explained: phylogenetics

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.

18.2.1 The phylogenetic tree

The evolutionary hypothesis of a phylogeny can be graphically represented by a phylogenetic tree.

Figure 18.4 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.

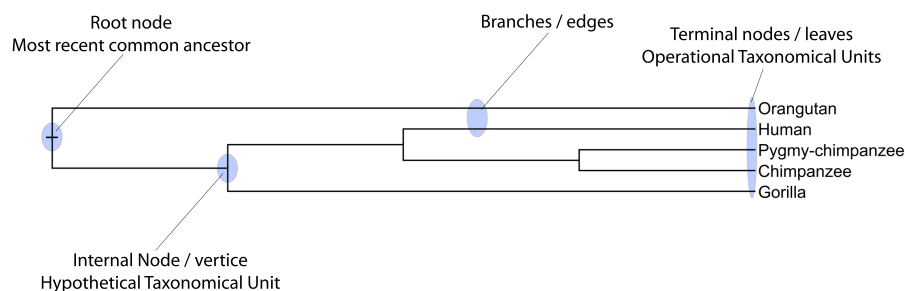


Figure 18.4: A proposed phylogeny of the great apes (Hominidae). Different components of the tree are marked, see text for description.

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 18.4 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 the common ancestor of gorilla, chimpanzee and man existed before the common ancestor of chimpanzee and man. If this information is absent trees can be drawn as unrooted.

18.2.2 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].

18.2.3 Reconstructing phylogenies from molecular data

Traditionally, phylogenies have been constructed from morphological data, but following the growth of genetic information it has become common practice to construct phylogenies based on molecular data, known as *molecular phylogeny*. The data is most commonly represented in the form of DNA or protein sequences, but can also be in the form of e.g. restriction fragment length polymorphism (RFLP).

Methods for constructing molecular phylogenies can be distance based or character based.

Distance based methods

Two common algorithms, both based on pairwise distances, are the UPGMA and the Neighbor Joining algorithms. Thus, the first step in these analyses is to compute a matrix of pairwise distances between OTUs from their sequence differences. To correct for multiple substitutions it is common to use distances corrected by a model of molecular evolution such as the Jukes-Cantor model [Jukes and Cantor, 1969].

UPGMA. A simple but popular clustering algorithm for distance data is Unweighted Pair Group Method using Arithmetic averages (UPGMA). [Michener and Sokal, 1957], [Sneath and Sokal, 1973]. This method works by initially having all sequences in separate clusters and continuously joining these. The tree is constructed by considering all initial clusters as leaf nodes in the tree, and each time two clusters are joined, a node is added to the tree as the parent of the two chosen nodes. The clusters to be joined are chosen as those with minimal pairwise distance. The branch lengths are set corresponding to the distance between clusters, which is calculated as the average distance between pairs of sequences in each cluster.

The algorithm assumes that the distance data has the so-called *molecular clock* property i.e. the divergence of sequences occur at the same constant rate at all parts of the tree. This means that the leaves of UPGMA trees all line up at the extant sequences and that a root is estimated as part of the procedure.

Neighbor Joining. The neighbor joining algorithm, [Saitou and Nei, 1987], on the other hand, builds a tree where the evolutionary rates are free to differ in different lineages, i.e., the tree does not have a particular root. Some programs always draw trees with roots for practical reasons, but for neighbor joining trees, no particular biological hypothesis is postulated by the placement of the root. The method works very much like UPGMA. The main difference is that instead of using pairwise distance, this method subtracts the distance to all other nodes from the pairwise distance. This is done to take care of situations where the two closest nodes are not neighbors in the "real" tree. The neighbor join algorithm is generally considered to be fairly good and is widely used. Algorithms that improves its cubic time performance exist. The improvement is only significant for quite large datasets.

Character based methods

Whereas the distance based methods compress all sequence information into a single number,

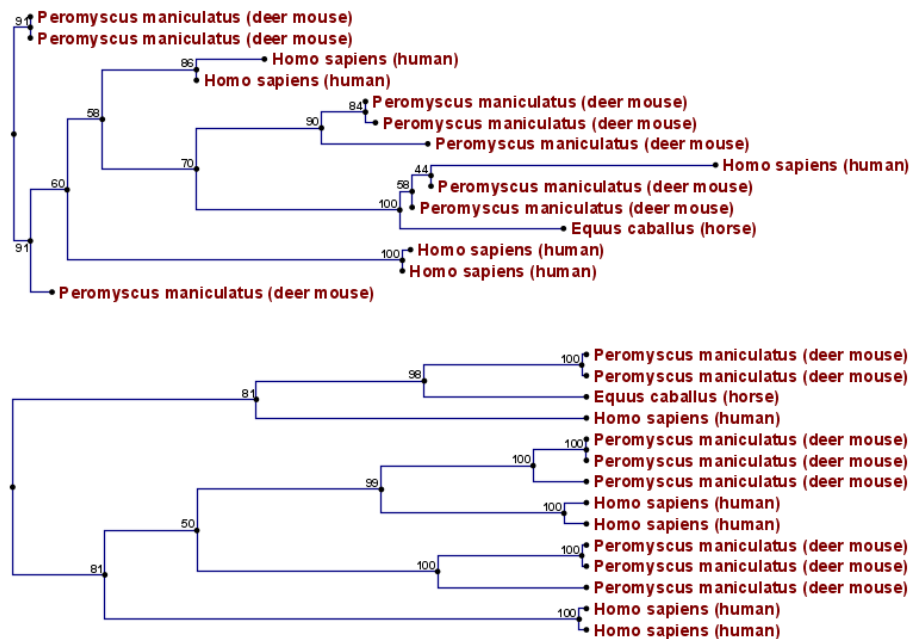


Figure 18.5: Algorithm choices for phylogenetic inference. The top shows a tree found by the neighbor joining algorithm, while the bottom shows a tree found by the UPGMA algorithm. The latter algorithm assumes that the evolution occurs at a constant rate in different lineages.

the character based methods attempt to infer the phylogeny based on all the individual characters (nucleotides or amino acids).

Parsimony. In parsimony based methods a number of sites are defined which are informative about the topology of the tree. Based on these, the best topology is found by minimizing the number of substitutions needed to explain the informative sites. Parsimony methods are not based on explicit evolutionary models.

Maximum Likelihood. Maximum likelihood and Bayesian methods (see below) are probabilistic methods of inference. Both have the pleasing properties of using explicit models of molecular evolution and allowing for rigorous statistical inference. However, both approaches are very computer intensive.

A stochastic model of molecular evolution is used to assign a probability (likelihood) to each phylogeny, given the sequence data of the OTUs. Maximum likelihood inference [Felsenstein, 1981] then consists of finding the tree which assign the highest probability to the data.

Bayesian inference. The objective of Bayesian phylogenetic inference is not to infer a single "correct" phylogeny, but rather to obtain the full posterior probability distribution of all possible phylogenies. This is obtained by combining the likelihood and the prior probability distribution of evolutionary parameters. The vast number of possible trees means that bayesian phylogenetics must be performed by approximative Monte Carlo based methods. [Larget and Simon, 1999], [Yang and Rannala, 1997].

18.2.4 Interpreting phylogenies

Bootstrap values

A popular way of evaluating the reliability of an inferred phylogenetic tree is bootstrap analysis.

The first step in a bootstrap analysis is to re-sample the alignment columns with replacement. I.e., in the re-sampled alignment, a given column in the original alignment may occur two or more times, while some columns may not be represented in the new alignment at all. The re-sampled alignment represents an estimate of how a different set of sequences from the same genes and the same species may have evolved on the same tree.

If a new tree reconstruction on the re-sampled alignment results in a tree similar to the original one, this increases the confidence in the original tree. If, on the other hand, the new tree looks very different, it means that the inferred tree is unreliable. By re-sampling a number of times it is possible to put reliability weights on each internal branch of the inferred tree. If the data was bootstrapped a 100 times, a bootstrap score of 100 means that the corresponding branch occurs in all 100 trees made from re-sampled alignments. Thus, a high bootstrap score is a sign of greater reliability.

Other useful resources

The Tree of Life web-project

<http://tolweb.org>

Joseph Felsensteins list of phylogeny software

<http://evolution.genetics.washington.edu/phylip/software.html>

Creative Commons License

All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in it's original form and "CLC bio" has to be clearly labelled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, or build upon this work.



See <http://creativecommons.org/licenses/by-nc-nd/2.5/> for more about how you may use the contents.

Part IV

Appendix

Appendix A

Comparison of workbenches

Below we list a number of functionalities that differ between CLC Workbenches:

- CLC Free Workbench (■)
- CLC Protein Workbench (■)
- CLC Gene Workbench (■)
- CLC Combined Workbench (■)

Batch processing	Free	Protein	Gene	Combined
Processing of multiple analyses in one single work-step		■	■	■

Database searches	Free	Protein	Gene	Combined
GenBank Entrez searches	■	■	■	■
UniProt searches (Swiss-Prot/TrEMBL)		■		■
Web-based sequence search using BLAST		■	■	■
PubMed searches		■	■	■
Web-based lookup of sequence data		■	■	■

General sequence analyses	Free	Protein	Gene	Combined
Linear sequence view	■	■	■	■
Circular sequence view	■	■	■	■
Text based sequence view	■	■	■	■
Editing sequences		■	■	■
Adding and editing sequence annotations		■	■	■
Sequence statistics	■	■	■	■
Shuffle sequence	■	■	■	■
Local complexity region analyses		■	■	■
Advanced protein statistics		■		■
Comprehensive protein characteristics report		■		■

For a more detailed comparison, we refer to <http://www.clcbio.com>.

Nucleotide analyses	Free	Protein	Gene	Combined
Basic gene finding	■	■	■	■
Reverse complement without loss of annotation	■	■	■	■
Restriction site analysis	■	■	■	■
Advanced interactive restriction site analysis			■	■
Translation of sequences from DNA to proteins	■	■	■	■
Interactive translations of sequences and alignments		■	■	■
G/C content analyses and graphs		■	■	■
Annotate with known SNP's in dbSNP database			■	■
Protein analyses	Free	Protein	Gene	Combined
3D molecule view		■		■
Hydrophobicity analyses		■	■	■
Antigenicity analysis		■		■
Protein charge analysis		■		■
Reverse translation from protein to DNA		■	■	■
Proteolytic cleavage detection		■		■
Prediction of signal peptides (SignalP)		■		■
Transmembrane helix prediction (TMHMM)		■		■
Secondary protein structure prediction		■		■
PFAM domain search		■		■
Sequence alignment	Free	Protein	Gene	Combined
Multiple sequence alignments (Two algorithms)	■	■	■	■
Advanced re-alignment and fix-point alignment options		■	■	■
Advanced alignment editing options		■	■	■
Consensus sequence determination and management	■	■	■	■
Conservation score along sequences	■	■	■	■
Sequence logo graphs along alignments		■	■	■
Gap fraction graphs		■	■	■
Dot plots	Free	Protein	Gene	Combined
Dot plot based analyses		■	■	■
Phylogenetic trees	Free	Protein	Gene	Combined
Neighbor-joining and UPGMA phylogenies	■	■	■	■
Pattern discovery	Free	Protein	Gene	Combined
Search for sequence match		■	■	■
Motif search		■	■	■
Pattern discovery		■	■	■

Primer design	Free	Protein	Gene	Combined
Advanced primer design tools			■	■
Detailed primer and probe parameters			■	■
Graphical display of primers			■	■
Generation of primer design output			■	■
Support for Standard PCR			■	■
Support for Nested PCR			■	■
Support for TaqMan PCR			■	■
Support for Sequencing primers			■	■
Match primer with sequence			■	■
Ordering of primers			■	■

Assembly of sequencing data	Free	Protein	Gene	Combined
Advanced contig assembly			■	■
Importing and viewing trace data			■	■
Trim sequences			■	■
Assemble without use of reference sequence			■	■
Assemble to reference sequence			■	■
Viewing and edit contigs			■	■

Molecular cloning	Free	Protein	Gene	Combined
Advanced molecular cloning			■	■
Graphical display of in silico cloning			■	■
Advanced sequence manipulation			■	■

Appendix B

BLAST databases

Several databases are available at NCBI, which can be selected to narrow down the possible BLAST hits.

B.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..

B.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 Sagarso Sea project. This does overlap with nucleotide nr.

Appendix C

Proteolytic cleavage enzymes

Most proteolytic enzymes cleave at a distinct pattern. We have compiled a list of enzymes which are used in CLC Protein Workbench

Name	P4	P3	P2	P1	P1'	P2'
Cyanogen bromide (CNBr)	-	-	-	M	-	-
Asp-N endopeptidase	-	-	-	-	D	-
Arg-C	-	-	-	R	-	-
Lys-C	-	-	-	K	-	-
Trypsin	-	-	-	K, R	not P	-
Trypsin	-	-	W	K	P	-
Trypsin	-	-	M	R	P	-
Trypsin*	-	-	C, D	K	D	-
Trypsin*	-	-	C	K	H, Y	-
Trypsin*	-	-	C	R	K	-
Trypsin*	-	-	R	R	H,R	-
Chymotrypsin-high spec.	-	-	-	F, Y	not P	-
Chymotrypsin-high spec.	-	-	-	W	not M, P	-
Chymotrypsin-low spec.	-	-	-	F, L, Y	not P	-
Chymotrypsin-low spec.	-	-	-	W	not M, P	-
Chymotrypsin-low spec.	-	-	-	M	not P, Y	-
Chymotrypsin-low spec.	-	-	-	H	not D, M, P, W	-
o-Iodosobenzoate	-	-	-	W	-	-
Thermolysin	-	-	-	not D, E	A, F, I, L, M or V	-
Post-Pro	-	-	H, K, R	P	not P	-
Glu-C	-	-	-	E	-	-
Asp-N	-	-	-	-	D	-
Proteinase K	-	-	-	A, E, F, I, L, T, V, W, Y	-	-
Factor Xa	A, F, G, I, L, T, V, M	D,E	G	R	-	-
Granzyme B	I	E	P	D	-	-
Thrombin	-	-	G	R	G	-
Thrombin	A, F, G, I, L, T, V, M	A, F, G, I, L, T, V, W, A	P	R	not D, E	not D, E

Table C.1: **Proteolytic cleavage**.Enzymes and chemicals. * exceptions for trypsin where no cleavage occurs.

Appendix D

Formats for import and export

D.1 List of bioinformatic data formats

Below is a list of bioinformatic data formats, i.e. formats for importing and exporting sequences, alignments and trees.

File type	Suffix	File format used for
Phylip Alignment	.phy	alignments
GCG Alignment	.msf	alignments
Clustal Alignment	.aln	alignments
Newick	.nwk	trees
FASTA	.fsa/.fasta	sequences
GenBank	.gbk/.gb/.gp	sequences
GCG sequence	.gcg	sequences (only import)
PIR (NBRF)	.pir	sequences (only import)
Staden	.sdn	sequences (only import)
VectorNTI		sequences (only import)
DNAstrider	.str/.strider	sequences
Swiss-Prot	.swp	protein sequences
Lasergene sequence	.pro	protein sequence (only import)
Lasergene sequence	.seq	nucleotide sequence (only import)
Embl	.embl	nucleotide sequences
Nexus	.nxs/.nexus	sequences, trees, alignments, and sequence lists
CLC	.clc	sequences, trees, alignments, reports, etc.
Text	.txt	all data in a textual format
ABI		Trace files (only import)
AB1		Trace files (only import)
SCF2		Trace files (only import)
SCF3		Trace files (only import)
Phred		Trace files (only import)
mmCIF	.cif	structure (only import)
PDB	.pdb	structure (only import)
Preferences	.cpf	CLC workbench preferences

Notice that *CLC Protein Workbench* can import 'external' files, too. This means that *CLC Protein Workbench* can import all files and display them in the **Navigation Area**, while the above

mentioned formats are the types which can be read by *CLC Protein Workbench*.

D.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 6.3 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

- [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. 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., Kierner, 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.
- [Blobel, 2000] Blobel, G. (2000). Protein targeting (Nobel lecture). *Chembiochem.*, 1:86–102.
- [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.
- [Dayhoff and Schwartz, 1978] Dayhoff, M. O. and Schwartz, R. M. (1978). *Atlas of Protein Sequence and Structure*, volume 3 of 5 suppl., chapter Atlas of Protein Sequence and Structure, 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.
- [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.

- [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 Chem*, 15:321–353.
- [Felsenstein, 1981] Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 17(6):368–376.
- [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., Wüning, 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.
- [Hein, 2001] Hein, J. (2001). An algorithm for statistical alignment of sequences related by a binary tree. *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.
- [Jukes and Cantor, 1969] Jukes, T. and Cantor, C. (1969). *Mammalian Protein Metabolism* (ed. HN Munro), chapter Evolution of protein molecules, pages 21–32. New York: Academic Press.
- [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.

- [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.
- [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.
- [Larget and Simon, 1999] Larget, B. and Simon, D. (1999). Markov chain monte carlo algorithms for the bayesian analysis of phylogenetic trees. *Mol Biol Evol*, 16:750-759.
- [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.
- [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.
- [Michener and Sokal, 1957] Michener, C. and Sokal, R. (1957). A quantitative approach to a problem in classification. *Evolution*, 11:130-162.
- [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.
- [Schechter and Berger, 1967] Schechter, I. and Berger, A. (1967). On the size of the active site in proteases. I. Papain. *Biochem Biophys Res Commun*, 27(2):157-162.
- [Schechter and Berger, 1968] Schechter, I. and Berger, A. (1968). On the active site of proteases. 3. Mapping the active site of papain; specific peptide inhibitors of papain. *Biochem Biophys Res Commun*, 32(5):898-902.

- [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.
- [Sneath and Sokal, 1973] Sneath, P. and Sokal, R. (1973). *Numerical Taxonomy*. Freeman, San Francisco.
- [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.
- [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 iin Chemistry*, 17:149–163.
- [Yang and Rannala, 1997] Yang, Z. and Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Mol Biol Evol*, 14(7):717–724.

Part V

Index

Index

- 3D molecule view, [131](#)
 - export graphics, [136](#)
 - navigate, [132](#)
 - output, [136](#)
 - rotate, [132](#)
 - zoom, [132](#)
- AB1, file format, [28, 80, 248](#)
- ABI, file format, [28, 80, 248](#)
- About CLC Workbenches, [18](#)
- Accession number, display, [55](#)
- Activate license
 - commercial, [17](#)
 - demo, [16](#)
- Add
 - annotations, [120, 241](#)
 - enzymes cutting selection, [116](#)
 - sequences to alignment, [227](#)
- Advanced preferences, [72](#)
- Algorithm
 - alignment, [216](#)
 - neighbor joining, [237](#)
 - UPGMA, [237](#)
- Align
 - alignments, [219](#)
 - protein sequences, tutorial, [32](#)
 - sequences, [241](#)
- Alignments, [216, 241](#)
 - add sequences to, [227](#)
 - create, [217](#)
 - edit, [225](#)
 - fast algorithm, [218](#)
 - join, [227](#)
 - multiple, Bioinformatics explained, [229](#)
 - remove sequences from, [226](#)
 - view, [222](#)
- Aliphatic index, [155](#)
- .aln, file format, [80](#)
- Ambiguities, reverse translation, [200](#)
- Amino acid composition, [157](#)
- Annotate with SNP's, [241](#)
- Annotation
 - add, [120](#)
 - copy to other sequences, [226](#)
 - edit, [120](#)
 - in alignments, [226](#)
 - layout, [115](#)
 - map, [124](#)
 - overview, [124](#)
 - types, [116](#)
- Antigenicity, [183, 241](#)
- Append wildcard, search, [96, 99](#)
- Arrange
 - layout of sequence, [29](#)
 - views in View Area, [61](#)
- Assembly, [241](#)
- Atomic composition, [156](#)
- Automatic parsing, [81](#)
- Back up, [84](#)
- Basic concepts of use, [20](#)
- Batch processing, [91, 241](#)
 - log of, [92](#)
- Bibliography, [250](#)
- Bioinformatic data
 - export, [82](#)
 - formats, [79, 248](#)
- BLAST, [241](#)
 - against local Database, [109](#)
 - against NCBI, [103](#)
 - create database from file system, [110](#)
 - create database from Navigation Area, [110](#)
 - create local database, [110](#)
 - graphics output, [107](#)
 - list of databases, [244](#)
 - parameters, [105](#)
 - search, [103](#)
 - table output, [108](#)
 - tutorial, [38](#)
- BLAST DNA sequence
 - BLASTn, [104](#)
 - BLASTx, [104](#)

- tBLASTx, 104
- BLAST Protein sequence
 - BLASTp, 104
 - tBLASTn, 104
- BLOSUM, scoring matrices, 144
- Bootstrap values, 238
- Bug reporting, 19
- C/G content, 118
- CDS, translate to protein, 120
- Cheap end gaps, 218
 - .cif, file format, 80, 131
- Circular view of sequence, 128, 241
 - .clc, file format, 80, 83
- CLC Standard Settings, 72, 73
- CLC Workbenches, 18
- CLC, file format, 28, 80, 248
- Cleavage, 201
 - the Peptidase Database, 204
- Cloning, 241
- Close View, 59
- Clustal, file format, 28, 80, 248
- Coding sequence, translate to protein, 120
- Codon
 - frequency tables, reverse translation, 198
 - usage, 200
- Color residues, 223
- Compare workbenches, 241
- Complexity plot, 149
- Configure network, 22
- Consensus sequence, 222, 241
 - open, 222
- Conservation, 222
 - graphs, 241
- Contact information, 11
- Contig, 241
- Convert old data, 81
- Copy, 87
 - annotations in alignments, 226
 - elements in Navigation Area, 54
 - into sequence, 120
 - search results, GenBank, 98
 - search results, UniProt, 101
 - sequence, 124, 126
 - sequence selection, 167
 - text selection, 124
 - .cpf, file format, 72
- Create
 - a project, tutorial, 26
- alignment, 217
- dot plots, 139
- enzyme list, 210
- local BLAST database, 110
- new folder, 53
- new project, 53
- workspace, 67
- Data formats
 - bioinformatic, 248
 - graphics, 249
- Data structure, 52
- Database
 - GenBank, 95
 - local, 52
 - nucleotide, 244
 - peptide, 244
 - UniProt, 98
- Delete
 - element, 56
 - residues and gaps in alignment, 226
 - workspace, 67
- Demo license, 15
- Dipeptide distribution, 157
- DNA translation, 168
- DNAstrider, file format, 28, 80, 248
- Dot plots, 241
 - Bioinformatics explained, 141
 - create, 139
 - print, 140
- Double stranded DNA, 114
- Download and open
 - search results, GenBank, 98
 - search results, UniProt, 101
- Download and save
 - search results, GenBank, 98
 - search results, UniProt, 101
- Download of *CLC Protein Workbench*, 11
- Drag and drop, 42
 - Navigation Area, 54
 - search results, GenBank, 98
 - search results, UniProt, 101
- Edit
 - alignments, 225, 241
 - annotations, 120, 241
 - enzymes, 116
 - sequence, 120
 - sequences, 241

- Element, 52
 - delete, 56
 - rename, 56
- .embl, file format, 80
- Embl, file format, 28, 80, 248
- Encapsulated PostScript, export, 86
- End gap cost, 218
- End gap costs
 - cheap end caps, 218
 - free end gaps, 218
- Enzyme list
 - create, 210
 - modify, 211
- .eps-format, export, 86
- Error reports, 19
- Evolutionary relationship, 232
- Example data, import, 22
- Expect, BLAST search, 107
- Export
 - bioinformatic data, 82
 - dependent objects, 83
 - folder, 82
 - graphics, 85
 - history, 83
 - list of formats, 248
 - multiple files, 82
 - preferences, 72
 - project, 82
- External files, import and export, 84
- Extinction coefficient, 155
- Extract sequences, 128
- FASTA, file format, 28, 80, 248
- Feature request, 19
- Feature table, 157
- Features, see Annotation
- File system, local BLAST database, 110
- Find open reading frames, 169
- Fit Width, 65
- Fixpoints, for alignments, 220
- Floating Side Panel, 73
- Format, of the manual, 24
- FormatDB, 110
- Fragments, separate on gel, 213
- Free end gaps, 218
 - .fsa, file format, 80
- G/C content, 118, 241
- Gap
 - delete, 226
 - extension cost, 218
 - fraction, 223, 241
 - insert, 225
 - open cost, 218
- .gbk, file format, 80
- GCG Alignment, file format, 28, 80, 248
- GCG Sequence, file format, 28, 80, 248
- Gel electrophoresis, 212
 - marker, 214
 - view, 213
 - view preferences, 213
 - when finding restriction sites, 208
- GenBank
 - file format, 28, 80, 248
 - search, 95, 241
 - search sequence in, 102
 - tutorial, 30
- Gene finding, 169
- General preferences, 71
- General Sequence Analyses, 138
- Genetic code, reverse translation, 199
- Getting started, 20
- Google sequence, 102
- Graphics
 - data formats, 249
 - export, 85
- Half-life, 155
- Handling of results, 91
- Help, 20
- Hide/show Toolbox, 66
- History, 89
 - export, 83
 - preserve when exporting, 90
 - source elements, 90
- Hydrophobicity, 185, 241
 - Bioinformatics explained, 188
 - Cornette, 189
 - Eisenberg, 189
 - Engelman (GES), 189
 - Hopp-Woods, 189
 - Janin, 189
 - Kyte-Doolittle, 189
 - Rose, 189
- Import
 - bioinformatic data, 80
 - data from older versions, 81

- existing data, 27
- external files, 84
- FASTA-data, 27
- list of formats, 248
- preferences, 72
- Vector NTI data, 81
- Infer Phylogenetic Tree, 232
- Insert
 - gaps, 225
 - restriction site, 116
- Installation, 11
- Isoelectric point, 155
- Join
 - alignments, 227
 - sequences, 158
 - .jpg-format, export, 86
- Lasergene sequence
 - protein file format, 28, 80, 248
 - sequence file format, 28, 80, 248
- License, 15
- Linux
 - installation, 13
 - installation with RPM-package, 14
- List of sequences, 126
- Load enzymes, 116
- Local BLAST Database, 110
- Local complexity plot, 149, 241
- Local Database, BLAST, 109
- Locale setting, 71
- Location
 - of selection on sequence, 65
 - Side Panel, 71
- Log of batch processing, 92
- Logo, sequence, 222, 241
- Mac OS X installation, 13
- Manipulate sequences, 241
- Manual format, 24
- Marker, in gel view, 214
- Max Sequence length for BLAST, 103
- Maximize size of view, 61
- Maximum memory, adjusting, 22
- Memory, adjust maximum amount, 22
- Menu Bar, illustration, 52
- mmCIF, file format, 28, 80, 248
- Mode toolbar, 63
- Modify enzyme list, 211
- Molecular weight, 154
- Motif search, 159, 241
- Mouse modes, 63
- Move
 - content of a view, 65
 - elements in Navigation Area, 54
 - sequences in alignment, 226
- .msf, file format, 80
- Multiple alignments, 229, 241
- Multiselecting, 54
- Navigate, 3D structure, 132
- Navigation Area, 52
 - create local BLAST database, 110
 - illustration, 52
- NCBI, 95
 - search sequence in, 102
 - search, tutorial, 30
- Negatively charged residues, 156
- Neighbor Joining algorithm, 237
- Neighbor-joining, 241
- Nested PCR primers, 241
- Network configuration, 22
- New
 - feature request, 19
 - folder, 27, 53
 - project, 27, 53
 - sequence, 125
- Newick, file format, 28, 80, 248
- .nexus, file format, 80
- Nexus, file format, 28, 80, 248
- Non-standard residues, 117
- nr, BLAST databases, 105
- Nucleotide
 - info, 117
 - sequence databases, 244
- Numbers on sequence, 114
- .nwk, file format, 80
- .nxs, file format, 80
- Old data, import, 81
- Online check, of demo license key, 15
- Open
 - consensus sequence, 222
 - files, 20
- Open reading frame determination, 169
- Open-ended sequence, 170
- Order primers, 241
- ORF, 169

- Origins from, 90
- Page setup, 77
- PAM, scoring matrices, 144
- Parameters
 - search, 96, 99
- Parsing, automatic, 81
- Paste/copy, 87
- Pattern Discovery, 162
- Pattern discovery, 241
- Pattern Search, 159
- PCR primers, 241
 - .pdb, file format, 80, 131
 - .seq, file format, 80
- PDB, file format, 28, 80, 248
 - .pdf-format, export, 86
- Peptidase, 201
- Peptide sequence databases, 244
- Personal information, 19
- Pfam domain search, 190, 241
- Phred, file format, 28, 80, 248
 - .phy, file format, 80
- Phylip, file format, 28, 80, 248
- Phylogenetic tree, 232, 241
 - tutorial, 34
- Phylogenetics, Bioinformatics explained, 235
 - .pir, file format, 80
- PIR (NBRF), file format, 28, 80, 248
- Plot
 - dot plot, 139
 - local complexity, 149
 - .png-format, export, 86
- Polarity colors, 117
- Positively charged residues, 157
- PostScript, export, 86
- Preferences, 70
 - advanced, 72
 - export, 72
 - General, 71
 - import, 72
 - style sheet, 72
 - toolbar, 71
 - View, 71
 - view, 62
- Primer
 - design, 241
- Print, 76
 - 3D molecule view, 136
 - dot plots, 140
 - preview, 77
 - visible area, 76
 - whole view, 76
- .pro, file format, 80
- Problems when starting up, 19
- Processes, 66
- Project, create new, 27
- Protease, cleavage, 201
- Protein
 - charge, 179, 241
 - cleavage, 201
 - hydrophobicity, 188
 - Isoelectric point, 155
 - report, 194, 241
 - report, output, 196
 - signal peptide, 173
 - statistics, 154
 - structure prediction, 193
 - translation, 197
- Proteolytic cleavage, 201, 241
 - Bioinformatics explained, 203
 - tutorial, 40
- Proxy server, 22
- Proxy settings
 - and license activation, 15
- .ps-format, export, 86
- PubMed references, search, 102
- PubMed references,search, 241
- Quick start, 21
- Rasmol colors, 117
- Reading frame, 169
- Realign alignment, 241
- Rebase, restriction enzyme database, 210
- Recycle Bin, 56
- Redo alignment, 219
- Redo/Undo, 60
- Reference sequence, 241
- References, 250
- Region
 - syntax, 121
 - types, 122
- Remove
 - annotations, 122
 - sequences from alignment, 226
 - terminated processes, 66
- Rename element, 56
- Replace file, 85

- Report program errors, 19
- Report, protein, 241
- Request new feature, 19
- Reset license, 17, 18
- Residue coloring, 117
- Restore
 - deleted elements, 56
 - size of view, 62
- Restriction enzymes, 206
 - separate on gel, 213
- Restriction sites, 206, 241
 - enzyme database Rebase, 210
 - on sequence, 116
 - parameters, 206
 - tutorial, 35
- Results handling, 91
- Reverse complement, 167, 241
- Reverse translation, 197, 241
 - Bioinformatics explained, 198
- RNA translation, 168
- Rotate, 3D structure, 132
- Safe mode, 19
- Save
 - changes in a view, 60
 - search, 31
 - sequence, 32
 - style sheet, 72
 - view preferences, 72
 - workspace, 67
- SCF2, file format, 28, 80, 248
- SCF3, file format, 28, 80, 248
- Score, BLAST search, 107
- Scoring matrices
 - Bioinformatics explained, 144
 - BLOSUM, 144
 - PAM, 144
- Search
 - BLAST, 103
 - GenBank, 95
 - handle results from GenBank, 97
 - handle results from UniProt, 100
 - hits, number of, 71
 - in a sequence, 118
 - in annotations, 118
 - Local BLAST, 109
 - options, GenBank, 95
 - options, UniProt, 99
 - parameters, 96, 99
 - patterns, 159, 162
 - Pfam domains, 190
 - PubMed references, 102
 - sequence in UniProt, 102
 - sequence on Google, 102
 - sequence on NCBI, 102
 - sequence on web, 101
 - TrEMBL, 98
 - UniProt, 98
- Secondary structure prediction, 193, 241
- Select
 - exact positions, 118
 - in sequence, 119
 - parts of a sequence, 119
 - workspace, 67
- Selection mode in the toolbar, 65
- Selection, location on sequence, 65
- Separate sequences on gel, 212
 - using restriction enzymes, 213
- Sequence
 - alignment, 216
 - analysis, 138
 - display different information, 55
 - extract from sequence list, 128
 - information, 123
 - information, tutorial, 36
 - join, 158
 - layout, 114
 - lists, 126
 - logo, 241
 - new, 125
 - region types, 122
 - search, 118
 - select, 119
 - shuffle, 148
 - statistics, 151
 - view, 113
 - view as text, 124
 - view circular, 128
 - view format, 55
 - web info, 101
- Sequence logo, 222, 223
- Sequencing data, 241
- Sequencing primers, 241
- Shortcuts, 68
- Show/hide Toolbox, 66
- Shuffle sequence, 148, 241
- Side Panel, location of, 71

- Signal peptide, 173, 174, 241
- SignalP, 173
 - Bioinformatics explained, 174
- SNP
 - annotation, 241
- Sort
 - sequences, 128
 - sequences alphabetically, 226
 - sequences by similarity, 226
- Source element, 90
- Species, display sequence species, 55
- Staden, file format, 28, 80, 248
- Standard layout, trees, 235
- Standard Settings, CLC, 73
- Start Codon, 170
- Start-up problems, 19
- Statistics
 - about sequence, 241
 - protein, 154
 - sequence, 151
- Status Bar, 66, 67
 - illustration, 52
- .str, file format, 80
- Structure, prediction, 193
- Style sheet, preferences, 72
- Support mail, 11
 - .svg-format, export, 86
- Swiss-Prot, 98
 - search, see UniProt
- Swiss-Prot, file format, 28, 80, 248
- Swiss-Prot/TrEMBL, 241
 - .swp, file format, 80
- System requirements, 14
- Tabs, use of, 58
- TaqMan primers, 241
- tBLASTn, 104
- tBLASTx, 104
- Terminated processes, 66
- Text format, 119
 - user manual, 24
 - view sequence, 124
- Text, file format, 28, 80, 248
 - .tif-format, export, 86
- Tips and tricks, tutorial, 41
- TMHMM, 181
- Toolbar
 - illustration, 52
 - preferences, 71
- Toolbox, 66
 - illustration, 52
 - show/hide, 66
- Topology layout, trees, 235
- Trace data, 241
- Translate
 - a selection, 117
 - along DNA sequence, 117
 - annotation to protein, 120
 - CDS, 169
 - coding regions, 169
 - DNA to RNA, 165
 - nucleotide sequence, 168
 - ORF, 169
 - protein, 197
 - RNA to DNA, 166
 - to DNA, 241
 - to protein, 168, 241
- Translation
 - of a selection, 117
 - show together with DNA sequence, 117
 - tables, 168
- Transmembrane helix prediction, 181, 241
- TrEMBL, search, 98
- Trim, 241
 - .txt, file format, 80
- Undo limit, 71
- Undo/Redo, 60
- UniProt, 98
 - search, 98, 241
 - search sequence in, 102
- UPGMA algorithm, 237, 241
- Upgrade license, 18
- Urls, Navigation Area, 84
- User defined view settings, 72
- User interface, 52
- Vector graphics, export, 86
- VectorNTI
 - file format, 28, 80, 248
 - import data from, 81
- View, 58
 - alignment, 222
 - dot plots, 140
 - preferences, 62
 - save changes, 60
 - sequence, 113
 - sequence as text, 124

- View Area, 58
 - illustration, 52
- View preferences, 71
 - show automatically, 71
 - style sheet, 72
- View settings
 - user defined, 72
- Wildcard, append to search, 96, 99
- Windows installation, 12
- Workspace, 67
 - create, 67
 - delete, 67
 - save, 67
 - select, 67
- Wrap sequences, 114
- Zoom, 63
 - tutorial, 29
- Zoom In, 63
- Zoom Out, 65
- Zoom to 100% , 65
- Zoom, 3D structure, 132