



# CLC **Free** Workbench

User manual

User manual for  
*CLC Free Workbench 3.0*  
Windows, Mac OS X and Linux

July 6, 2006

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



# Contents

<b>I</b>	<b>Introduction</b>	<b>7</b>
<b>1</b>	<b>Introduction to CLC Free Workbench</b>	<b>8</b>
1.1	Contact information . . . . .	9
1.2	Download and installation . . . . .	9
1.3	System requirements . . . . .	12
1.4	About CLC Workbenches . . . . .	13
1.5	When the program is installed: Getting started . . . . .	14
1.6	Network configuration . . . . .	16
1.7	Adjusting the maximum amount of memory . . . . .	17
1.8	The format of the user manual . . . . .	18
<b>2</b>	<b>Tutorials</b>	<b>20</b>
2.1	Tutorial: Starting up the program . . . . .	21
2.2	Tutorial: View sequence . . . . .	24
2.3	Tutorial: GenBank search and download . . . . .	25
2.4	Tutorial: Align protein sequences . . . . .	27
2.5	Tutorial: Create and modify a phylogenetic tree . . . . .	28
2.6	Tutorial: Detect restriction sites . . . . .	30
2.7	Tutorial: Sequence information . . . . .	31
2.8	Tips and tricks for the experienced user . . . . .	33
<b>II</b>	<b>Basic Program Functionalities</b>	<b>40</b>
<b>3</b>	<b>User Interface</b>	<b>41</b>
3.1	Navigation Area . . . . .	42

3.2 View Area . . . . .	48
3.3 Zoom and selection in View Area . . . . .	53
3.4 Toolbox and Status Bar . . . . .	56
3.5 Workspace . . . . .	57
3.6 List of shortcuts . . . . .	58
<b>4 User preferences</b>	<b>60</b>
4.1 General preferences . . . . .	61
4.2 Default View preferences . . . . .	61
4.3 Advanced preferences . . . . .	62
4.4 Export/import of preferences . . . . .	62
4.5 View preference style sheet . . . . .	62
<b>5 Printing</b>	<b>65</b>
5.1 Selecting which part of the view to print . . . . .	65
5.2 Page setup . . . . .	66
5.3 Print preview . . . . .	66
<b>6 Import/export of data and graphics</b>	<b>68</b>
6.1 Bioinformatic data formats . . . . .	68
6.2 External files . . . . .	73
6.3 Export graphics to files . . . . .	74
6.4 Copy/paste view output . . . . .	76
<b>7 History</b>	<b>78</b>
7.1 Element history . . . . .	78
<b>8 Handling of results</b>	<b>80</b>
8.1 How to handle results of analyses . . . . .	80
<b>III Bioinformatics</b>	<b>83</b>
<b>9 Database search</b>	<b>84</b>
9.1 GenBank search . . . . .	84

<b>10 Viewing and editing sequences</b>	<b>88</b>
10.1 View sequence . . . . .	88
10.2 Sequence information . . . . .	94
10.3 View as text . . . . .	96
10.4 Creating a new sequence . . . . .	96
10.5 Sequence Lists . . . . .	98
10.6 Circular DNA . . . . .	100
<b>11 General sequence analyses</b>	<b>102</b>
11.1 Sequence statistics . . . . .	102
11.2 Shuffle sequence . . . . .	105
11.3 Join sequences . . . . .	105
<b>12 Nucleotide analyses</b>	<b>108</b>
12.1 Convert DNA to RNA . . . . .	108
12.2 Convert RNA to DNA . . . . .	109
12.3 Reverse complements of sequences . . . . .	110
12.4 Translation of DNA or RNA to protein . . . . .	111
12.5 Find open reading frames . . . . .	111
<b>13 Restriction site analyses</b>	<b>115</b>
13.1 Restriction sites and enzyme lists . . . . .	115
13.2 Restriction site analysis . . . . .	115
13.3 Restriction enzyme lists . . . . .	118
<b>14 Sequence alignment</b>	<b>120</b>
14.1 Create an alignment . . . . .	120
14.2 View alignments . . . . .	123
14.3 Edit alignments . . . . .	124
14.4 Bioinformatics explained: Multiple alignments . . . . .	126
<b>15 Phylogenetic trees</b>	<b>129</b>
15.1 Inferring phylogenetic trees . . . . .	129
15.2 Bioinformatics explained: phylogenetics . . . . .	132

---

<b>IV</b>	<b>Appendix</b>	<b>137</b>
<b>A</b>	<b>Comparison of workbenches</b>	<b>138</b>
<b>B</b>	<b>Formats for import and export</b>	<b>141</b>
B.1	List of bioinformatic data formats . . . . .	141
B.2	List of graphics data formats . . . . .	142
	<b>Bibliography</b>	<b>143</b>
<b>V</b>	<b>Index</b>	<b>145</b>

## **Part I**

# **Introduction**

# Chapter 1

## Introduction to *CLC Free Workbench*

### Contents

---

<b>1.1</b>	<b>Contact information</b>	<b>9</b>
<b>1.2</b>	<b>Download and installation</b>	<b>9</b>
1.2.1	Program download	9
1.2.2	Installation on Microsoft Windows	10
1.2.3	Installation on Mac OS X	11
1.2.4	Installation on Linux with an installer	11
1.2.5	Installation on Linux with an RPM-package	12
<b>1.3</b>	<b>System requirements</b>	<b>12</b>
<b>1.4</b>	<b>About CLC Workbenches</b>	<b>13</b>
1.4.1	New program feature request	13
1.4.2	Report program errors	13
1.4.3	Free vs. commercial workbenches	14
<b>1.5</b>	<b>When the program is installed: Getting started</b>	<b>14</b>
1.5.1	Basic concepts of using CLC Workbenches	14
1.5.2	Quick start	15
1.5.3	Import of example data	16
<b>1.6</b>	<b>Network configuration</b>	<b>16</b>
<b>1.7</b>	<b>Adjusting the maximum amount of memory</b>	<b>17</b>
1.7.1	Microsoft Windows	17
1.7.2	Mac OS X	17
1.7.3	Linux	18
<b>1.8</b>	<b>The format of the user manual</b>	<b>18</b>
1.8.1	Text formats	19

---

Welcome to *CLC Free Workbench 3.0* — 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 Free Workbench 3.0* 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](mailto: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](mailto:support@clcbio.com)

## 1.2 Download and installation

The *CLC Free 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](mailto: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.

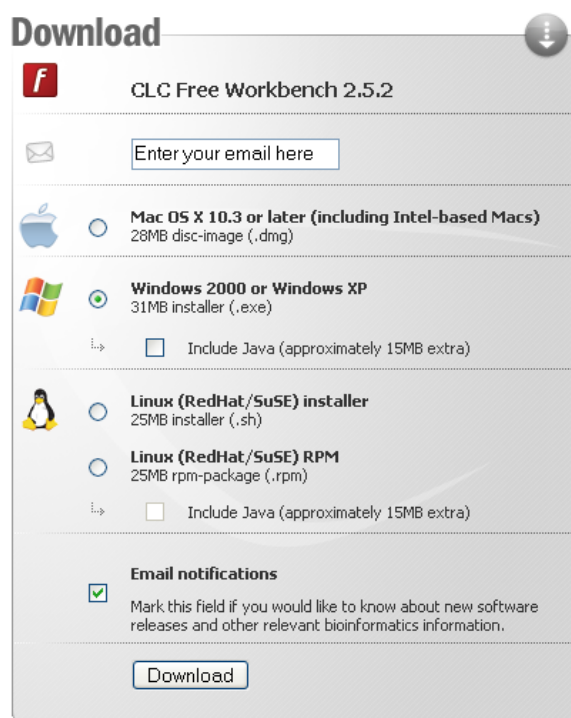


Figure 1.1: Download dialog.

### 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 Free Workbench" from the menu displayed.

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

Installing the program is done in the following steps:

- On the welcome screen, click Next.
- Read and accept the License agreement and click Next.
- Choose where you would like to install the application and click Next.
- Choose a name for the Start Menu folder used to launch *CLC Free Workbench* and click Next.
- Choose where you would like to create shortcuts for launching *CLC Free Workbench* and click Next.

- Wait for the installation process to complete, choose whether you would like to launch *CLC Free 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 Free 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 Free Workbench* and click Next.
- Wait for the installation process to complete, choose whether you would like to launch *CLC Free 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 CLCFreeWorkbench_2_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:

```
# clcfreewb2
```

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

```
# ./clcfreewb2
```

### 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 CLCFreeWorkbench_2_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:

```
# clcfreewb2
```

## 1.3 System requirements

The system requirements of *CLC Free Workbench 3.0* 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 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.

### 1.4.1 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](mailto:support@clcbio.com).

### 1.4.2 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](mailto: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.

### 1.4.3 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 1.2.

## 1.5 When the program is installed: Getting started

*CLC Free Workbench 3.0* 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.

### 1.5.1 Basic concepts of using CLC Workbenches

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

- All data for use in the *CLC Free 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 1.3 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()).

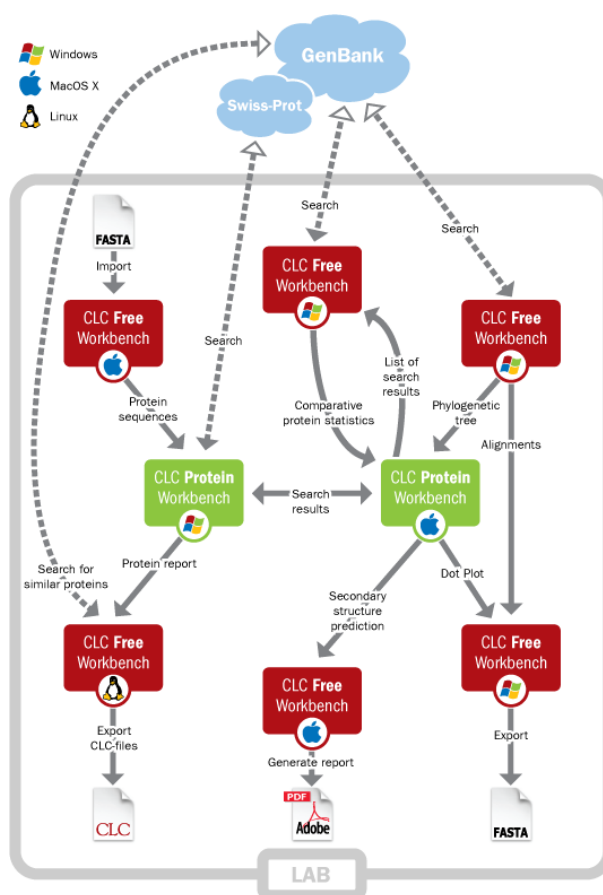


Figure 1.2: An example of how research can be organized and how data can flow between users of different workbenches, working on different platforms.

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**

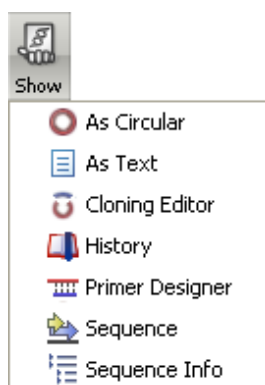


Figure 1.3: The different ways of viewing DNA sequences.

### 1.5.2 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 1.4.



Figure 1.4: 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 Free Workbench 3.0* includes an example data set, which can be found on our web page, or downloaded from the program (Also found in the **Help** menu).

### 1.5.3 Import of example data

When downloading *CLC Free Workbench 3.0* 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.

## 1.6 Network configuration

If you use a proxy server to access the Internet you must configure *CLC Free Workbench 3.0* to use this. Otherwise you will not be able to perform any on-line activities (e.g. searching GenBank). *CLC Free Workbench 3.0* supports the use of a HTTP-proxy and an anonymous SOCKS-proxy.

To configure your proxy settings, open *CLC Free Workbench 3.0*, and go to the **Advanced**-tab of the **Preferences** dialog (figure 1.5) and enter the appropriate information.

You have the choice between a HTTP-proxy and a SOCKS-proxy. *CLC Free Workbench 3.0* 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.



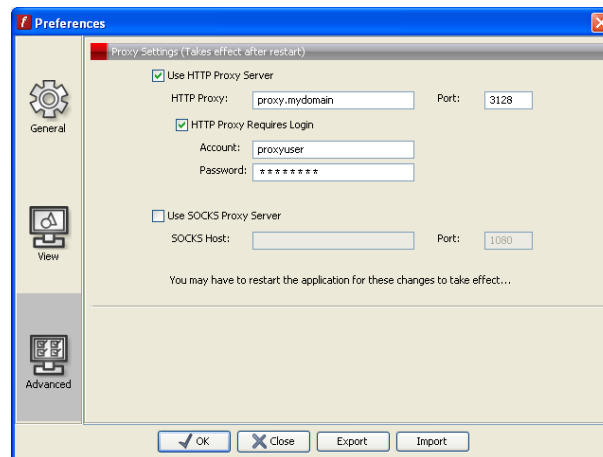


Figure 1.5: Adjusting proxy preferences.

## 1.7 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 Free Workbench 3.0* to a newer version.

### 1.7.1 Microsoft Windows

- Locate the *CLC Free Workbench 3.0* directory inside your Program Files directory and open it
- Create a new, empty text-file called `clwb.voptions` (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.

### 1.7.2 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.

### 1.7.3 Linux

- Locate the directory where you installed *CLC Free Workbench 3.0* 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.

## 1.8 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 Free Workbench 3.0*.
- 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 Free Workbench 3.0* and provide more general knowledge of bioinformatic concepts.
- The **fourth part** is the Appendix and Index.

Each chapter includes a short table of contents.

### 1.8.1 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 Free Workbench 3.0***
- The captions of displayed screenshots are in *italic*.

## Chapter 2

# Tutorials

### Contents

---

<b>2.1 Tutorial: Starting up the program</b>	<b>21</b>
2.1.1 Creating a project and a folder	22
2.1.2 Import data	22
2.1.3 Supported data formats	22
<b>2.2 Tutorial: View sequence</b>	<b>24</b>
<b>2.3 Tutorial: GenBank search and download</b>	<b>25</b>
2.3.1 Saving the search	26
2.3.2 Searching for matching objects	26
2.3.3 Saving the sequence	27
<b>2.4 Tutorial: Align protein sequences</b>	<b>27</b>
2.4.1 Alignment dialog	27
<b>2.5 Tutorial: Create and modify a phylogenetic tree</b>	<b>28</b>
2.5.1 Tree layout	29
<b>2.6 Tutorial: Detect restriction sites</b>	<b>30</b>
2.6.1 View restriction site	30
<b>2.7 Tutorial: Sequence information</b>	<b>31</b>
<b>2.8 Tips and tricks for the experienced user</b>	<b>33</b>
2.8.1 Open and arrange views using drag and drop	34
2.8.2 Find element in the Navigation Area	34
2.8.3 Find specific annotations on a sequence	34
2.8.4 Split sequences into several lines	35
2.8.5 Make a new sequence of a coding region	35
2.8.6 Get overview and detail of a sequence at the same time	36
2.8.7 Smart selecting in sequences and alignments	36
2.8.8 Quickly import sequences using copy-paste	37
2.8.9 Perform analyses on many elements	38
2.8.10 Drag elements to the Toolbox	38
2.8.11 Export elements while preserving history	38
2.8.12 Avoid the mouse trap - use keyboard shortcuts	39

---

This chapter contains tutorials representing some of the features of *CLC Free Workbench 3.0*. 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 Free Workbench 3.0*.

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 Free 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 Free Workbench* for the first time, the user interface looks like figure 2.1.

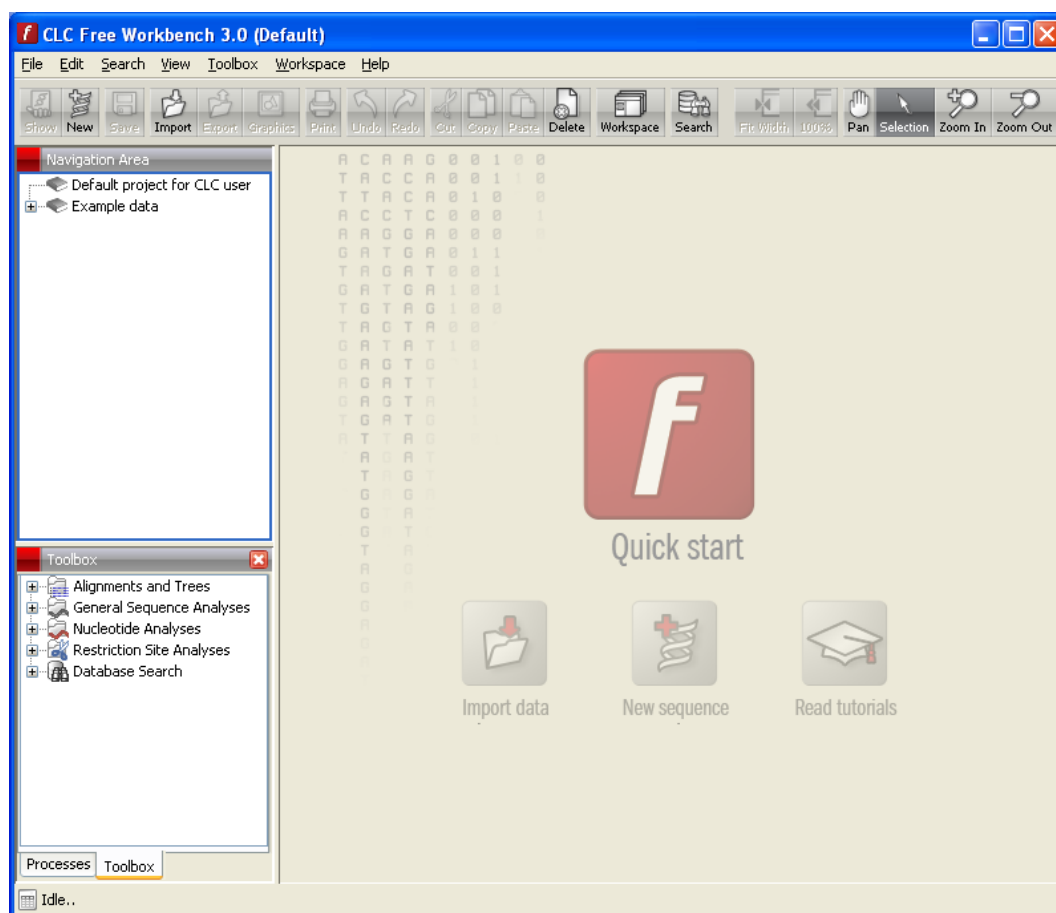


Figure 2.1: The user interface as it looks when you start the program for the first time. (Windows version of **CLC Free Workbench**. The interface is similar for Mac and Linux.)

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 Free 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 Free 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 (⌘ + 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 (⌘ + 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 Free 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 Free Workbench* can import and export the following formats:

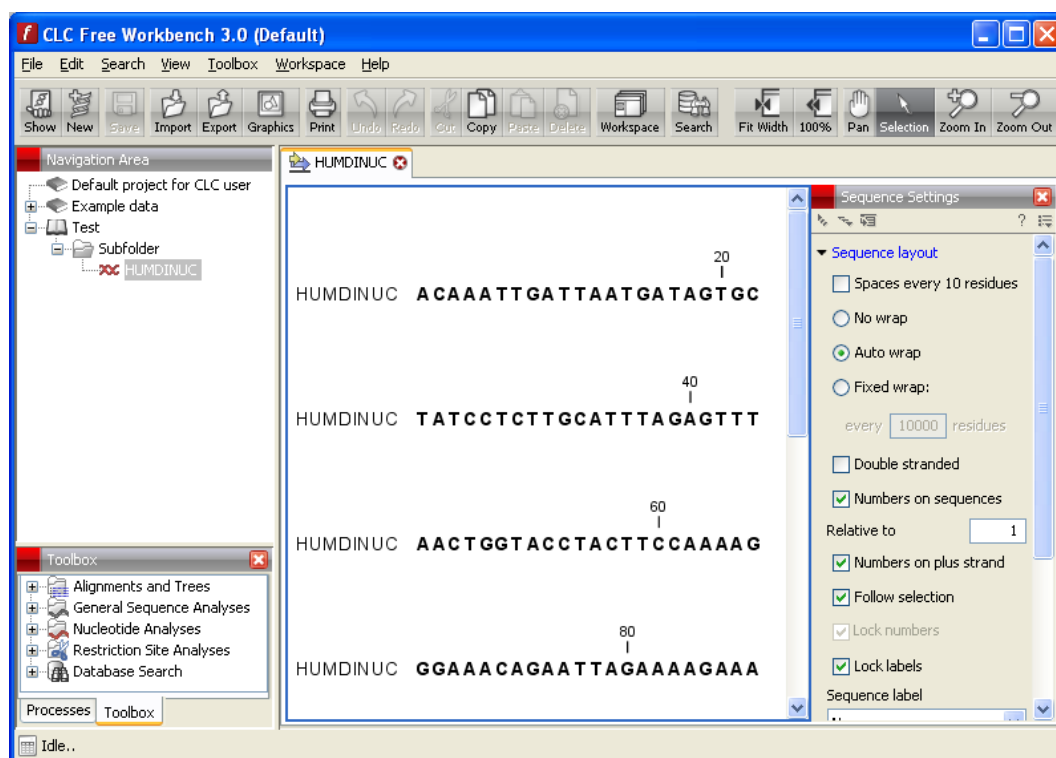


Figure 2.2: The HUMDINUC file is imported and opened.

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 Free Workbench* can import 'external' files, too. This means that *CLC Free 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 Free 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 (✖) at the top right corner of the **Side Panel** (in the right side of the **View Area**). (See figure 2.3).

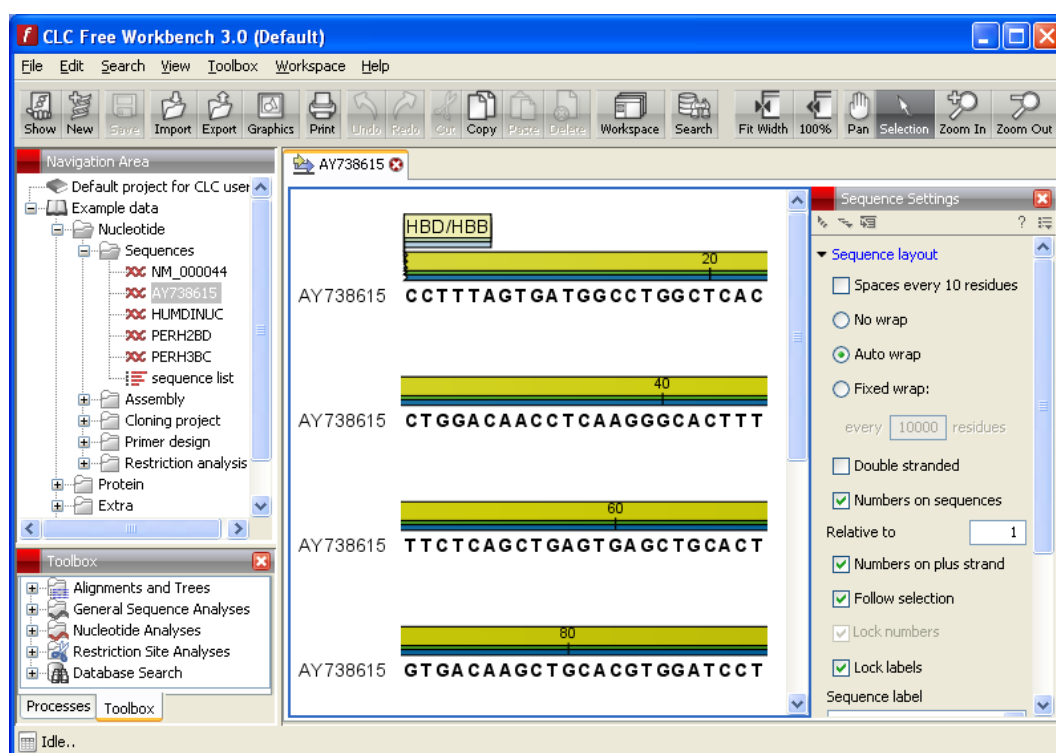


Figure 2.3: DNA sequence 'AY738615' opened in a view. The view preferences has been hidden to provide more space for the view.

As default, *CLC Free 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 (🔍) 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).



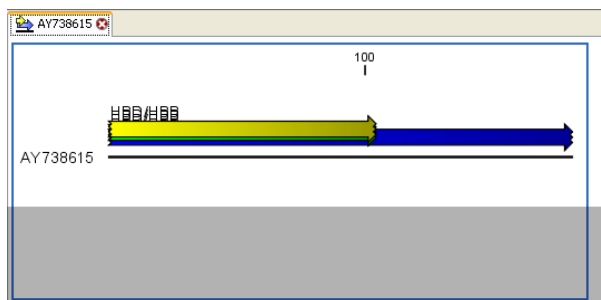


Figure 2.4: Dragging the tab down to the bottom of the view will display a gray area indicating that the tab can be "dropped" here and split the view.

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

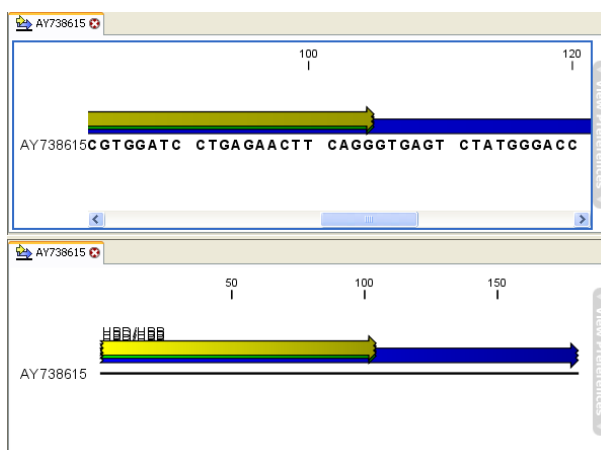



Figure 2.5: The resulting two views which are split horizontally.

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 Free 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 Free 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 'Add search parameters' and 'Start search' buttons. Below the search criteria, there's a checkbox for '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 10 results, including human and mouse hemoglobin sequences. At the bottom of the table, 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 Free 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).

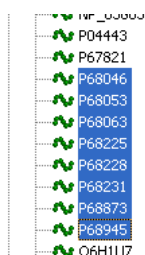


Figure 2.7: Eight protein sequences in a Protein project in the Navigation Area.

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**

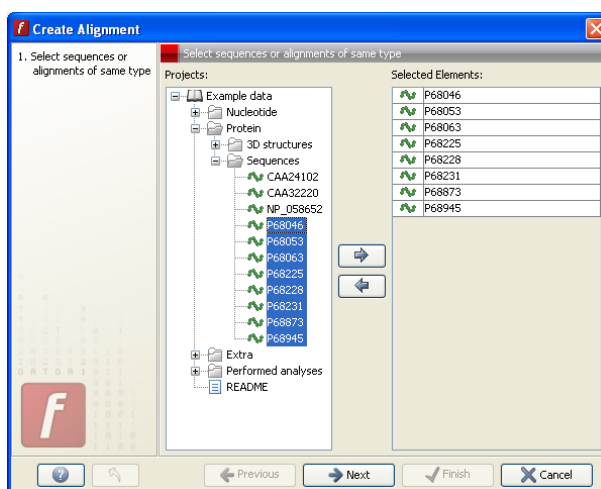


Figure 2.8: The alignment dialog displaying the 8 chosen protein sequences.

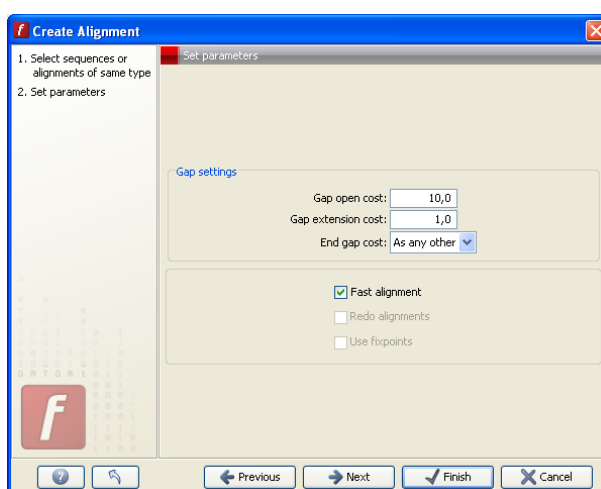


Figure 2.9: The alignment dialog displaying the available parameters which can be adjusted.

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

**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



Figure 2.10: The resulting alignment.

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).

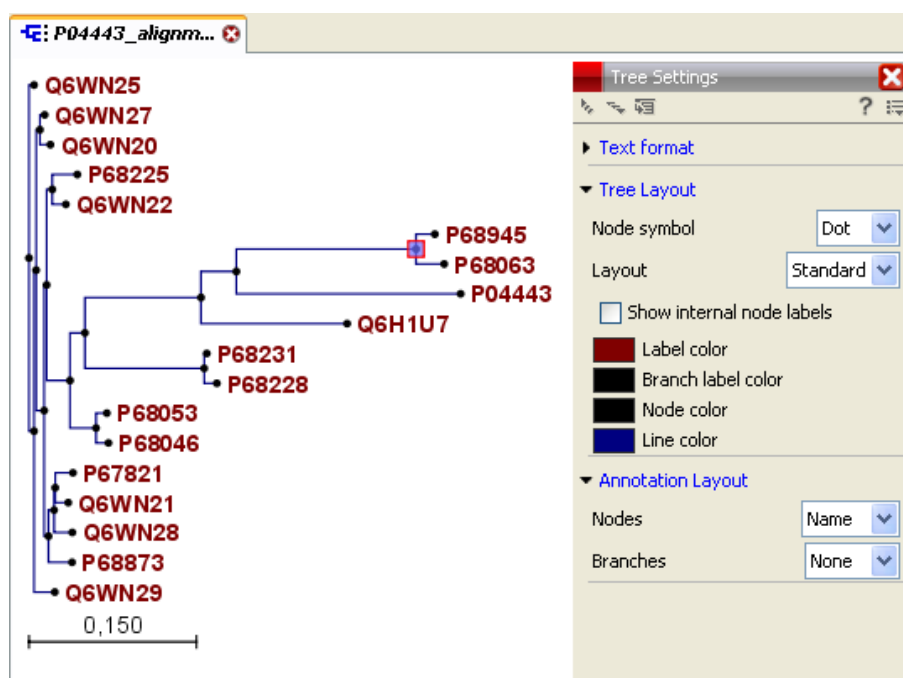


Figure 2.11: After choosing which algorithm should be used, the tree appears in the View Area. The Side panel in the right side of the view allows you to adjust the way the tree is displayed.

### 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.

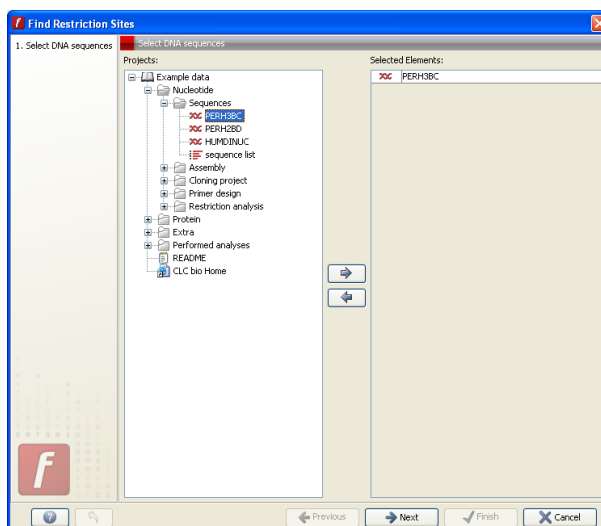


Figure 2.12: Choosing sequence PERH3BC.

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.

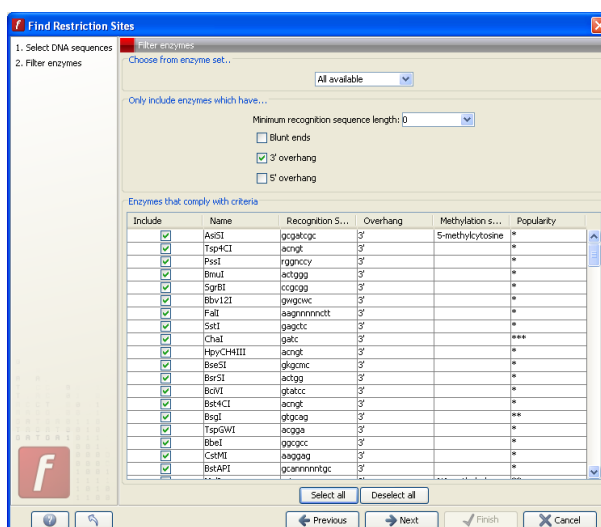


Figure 2.13: Setting parameters for restriction site detection.

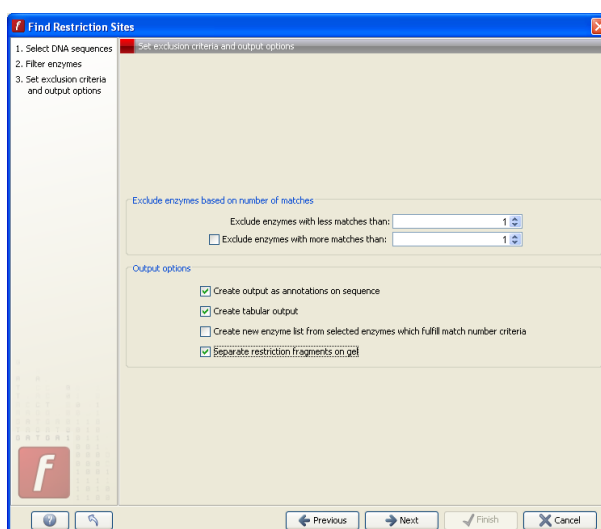


Figure 2.14: Selecting enzymes.

**Notice!** The results are not automatically saved.

To save the result:

**Right-click the tab | File | Save** (💾)

## 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 Free Workbench*:

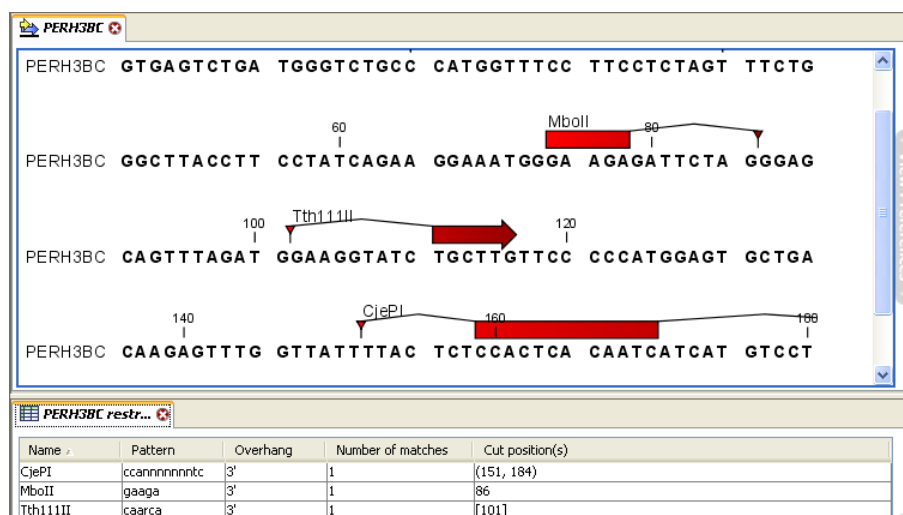


Figure 2.15: The result of the restriction site detection is displayed as text, and in this tutorial the View shares the View Area with a View of the PERH3BC sequence displaying the restriction sites (split-screen-view).

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

This opens a new view shown in figure 2.16.



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

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 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).

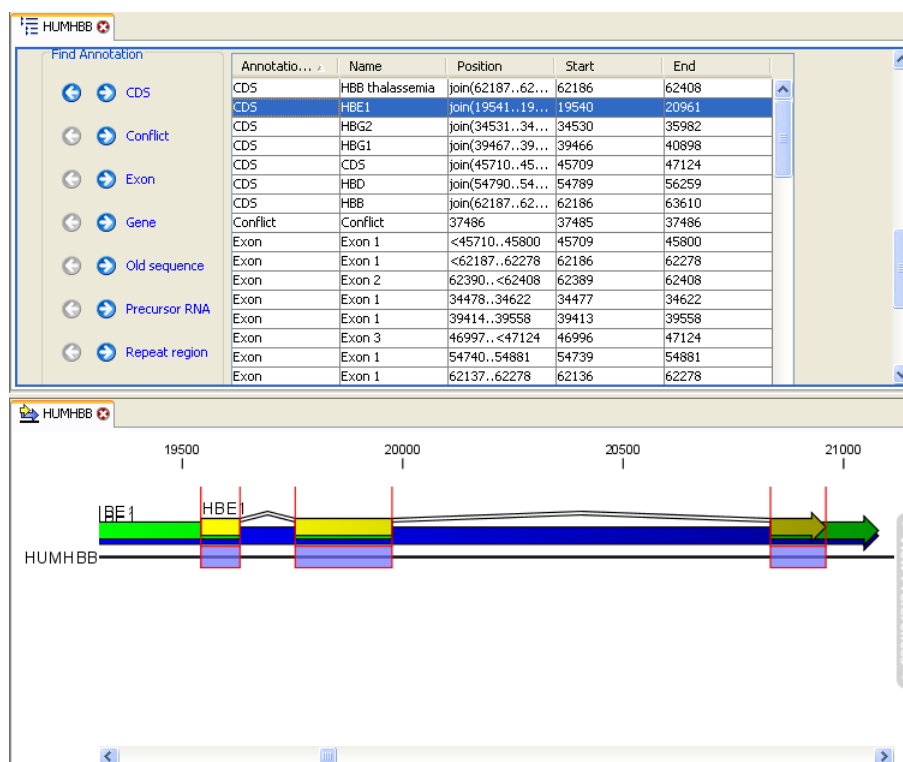


Figure 2.17: Two views of the HUMHBB sequence. The upper view shows the coding sequences (CDS), and the bottom view shows a selection corresponding to the CDS chosen in the upper view.

## 2.8 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 Free 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 explained.

### 2.8.1 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.

### 2.8.2 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 2.18).

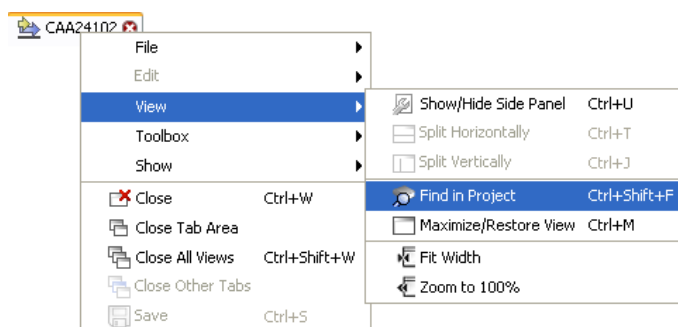


Figure 2.18: This will select the sequence in the Navigation Area.

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

### 2.8.3 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 2.19).

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

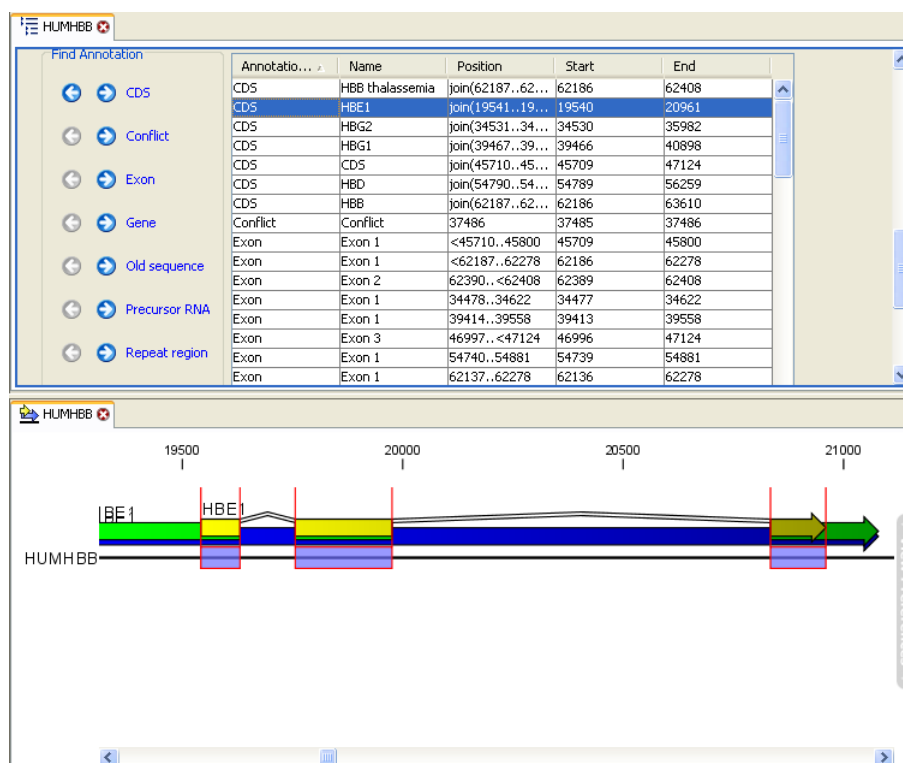


Figure 2.19: Clicking the HBE1 coding region in the top view selects the annotation on the sequence in the bottom view.

#### 2.8.4 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 2.20).

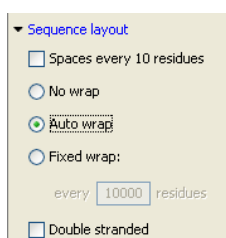


Figure 2.20: Wrapping the sequence automatically.

#### 2.8.5 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 2.21):

**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.

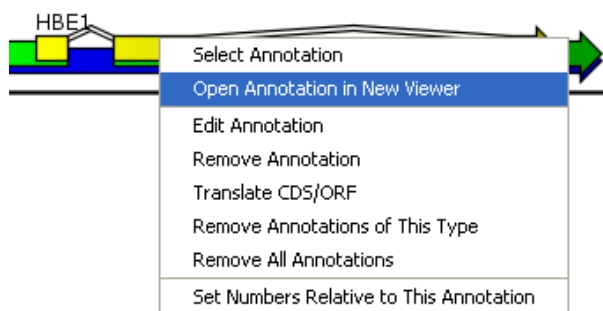


Figure 2.21: Opening the coding region in a new view.

### 2.8.6 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 2.22), 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.

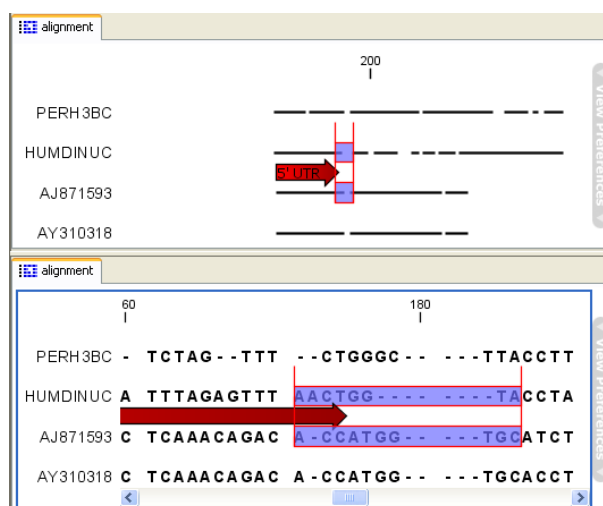


Figure 2.22: Using the split views and follow selection functionalities.

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**).

### 2.8.7 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 2.21).

**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 2.23). Remember to separate the start and end numbers with two punctuation marks (..).

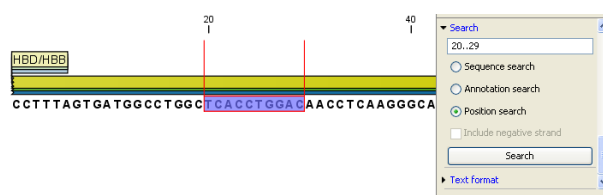


Figure 2.23: Making a selection from position 20 to 29 (both included) using the Search function.

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**.

### 2.8.8 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 2.24).

Figure 2.24: 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.

### 2.8.9 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 2.25), 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.

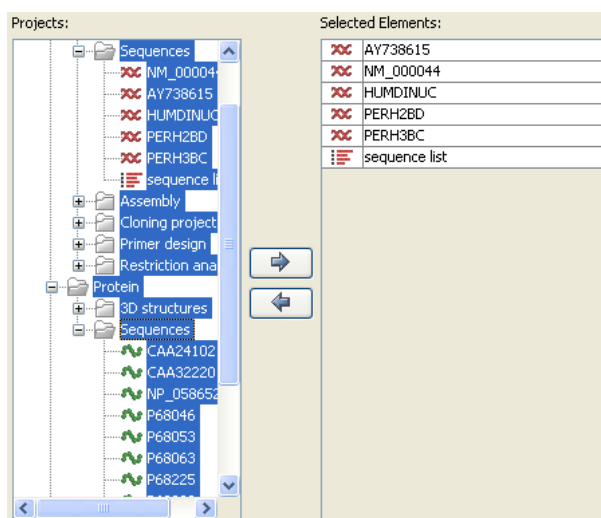


Figure 2.25: Selecting protein and dna sequences, but the dialog automatically filters out the protein sequences.

### 2.8.10 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,

### 2.8.11 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 2.26):

**Select the alignment | File | Export with dependent elements**

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.

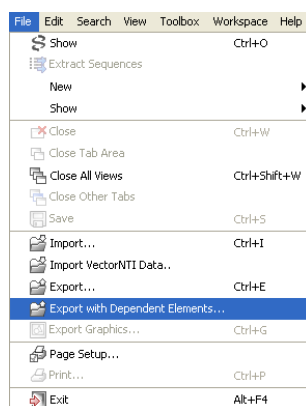


Figure 2.26: *Export with dependent elements in order to preserve the detailed history of an element.*

### 2.8.12 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 Free 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

---

<b>3.1 Navigation Area</b>	<b>42</b>
3.1.1 Data structure	42
3.1.2 Create new projects and folders	43
3.1.3 Multiselecting elements	44
3.1.4 Moving and copying elements	44
3.1.5 Change element names	45
3.1.6 Delete elements	46
3.1.7 Show folder elements in View	47
3.1.8 Sequence properties	48
<b>3.2 View Area</b>	<b>48</b>
3.2.1 Open View	48
3.2.2 Close Views	49
3.2.3 Save changes in a View	50
3.2.4 Undo/Redo	50
3.2.5 Arrange Views in View Area	51
3.2.6 Side Panel	52
<b>3.3 Zoom and selection in View Area</b>	<b>53</b>
3.3.1 Zoom In	53
3.3.2 Zoom Out	55
3.3.3 Fit Width	55
3.3.4 Zoom to 100%	55
3.3.5 Move	55
3.3.6 Selection	55
<b>3.4 Toolbox and Status Bar</b>	<b>56</b>
3.4.1 Processes	56
3.4.2 Toolbox	56
3.4.3 Status Bar	57
<b>3.5 Workspace</b>	<b>57</b>
3.5.1 Create Workspace	57
3.5.2 Select Workspace	57

3.5.3 Delete Workspace . . . . .	57
3.6 List of shortcuts . . . . .	58

This chapter provides an overview of the different areas in the user interface of *CLC Free Workbench 3.0*. As can be seen from figure 3.1 this includes a **Navigation Area**, **View Area**, **Menu Bar**, **Toolbar**, **Status Bar** and **Toolbox**.

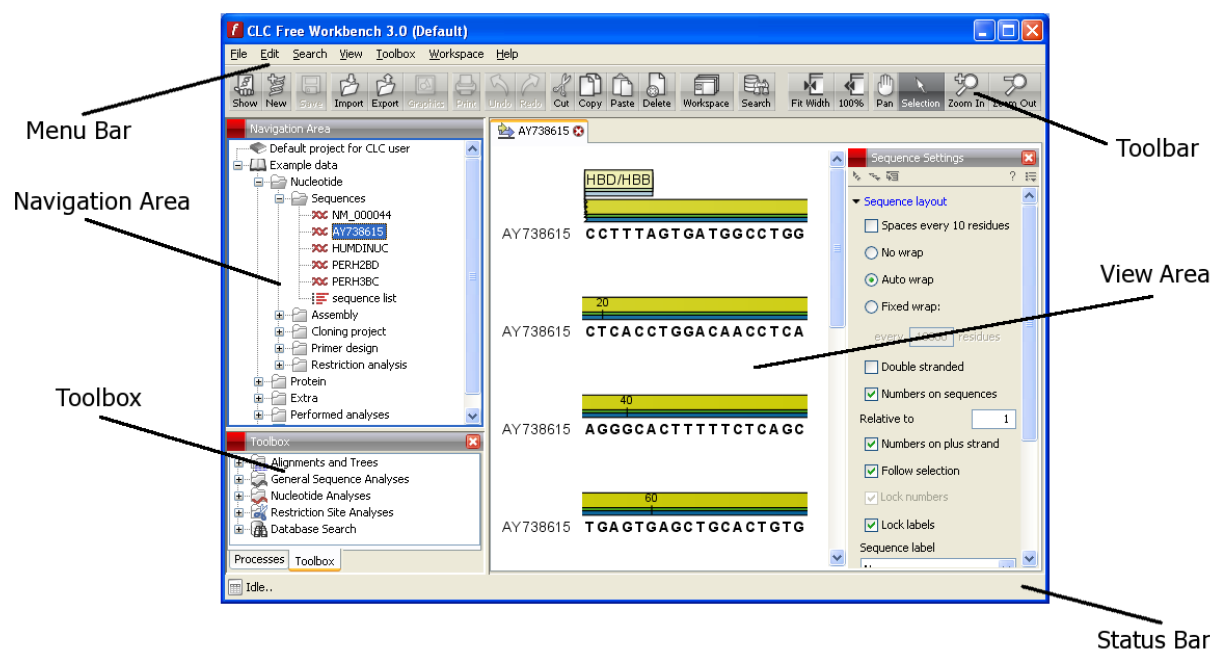


Figure 3.1: The user interface consists of the Menu Bar, Toolbar, Status Bar, Navigation Area, Toolbox, and View Area.

## 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 Free Workbench 3.0* 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:

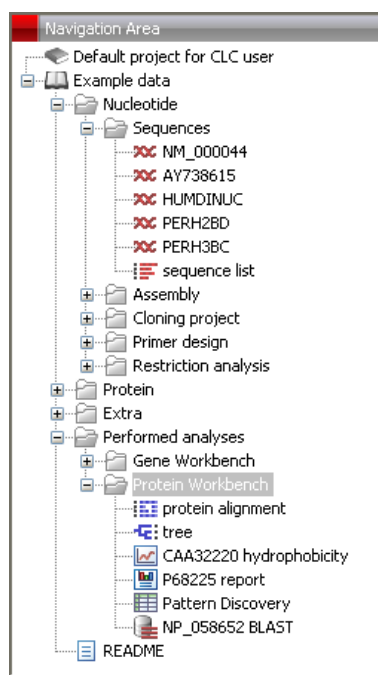


Figure 3.2: The Navigation Area.

**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.

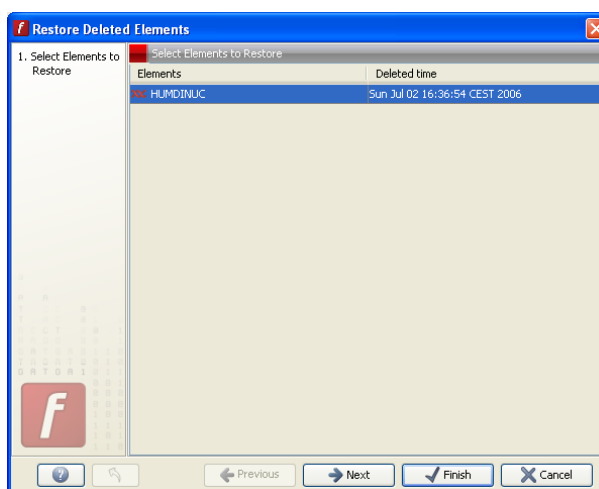


Figure 3.3: The Restore Deleted Elements dialog.

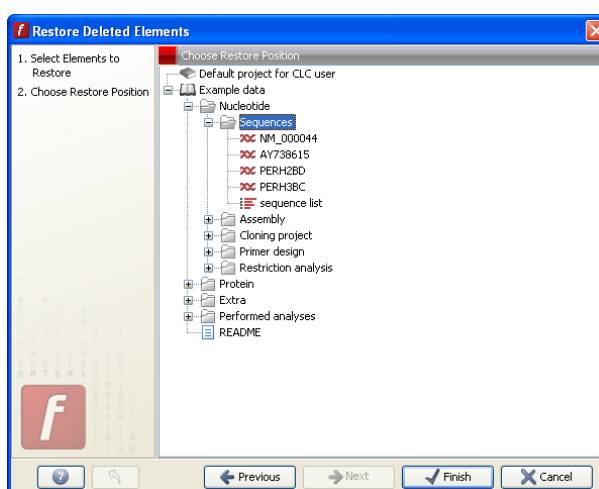


Figure 3.4: The Restore Deleted Elements dialog.

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 10.2.

## 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 Free Workbench 3.0*, 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.



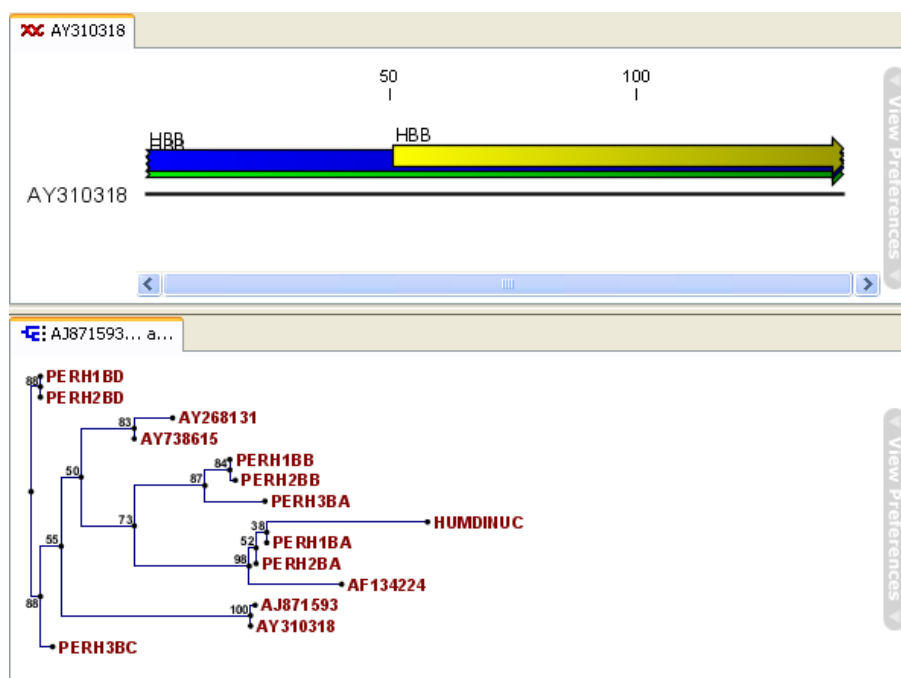


Figure 3.6: A View Area can enclose several Views, each View is indicated with a tab (see top left View, which shows protein P12675). Furthermore, several Views can be shown at the same time (in this example, three views are displayed).

**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.

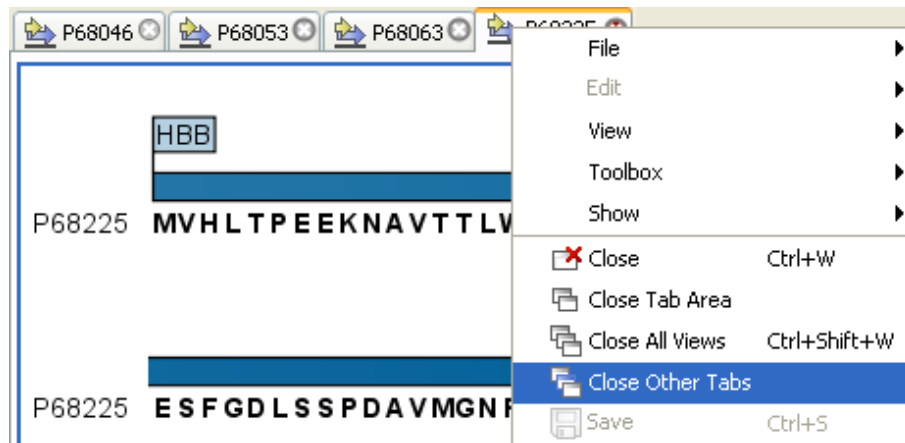


Figure 3.7: By right-clicking a tab, several close options are available.

### 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).

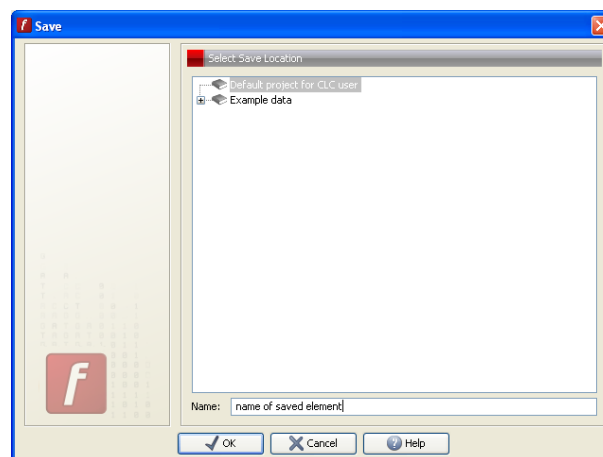


Figure 3.8: Save dialog.

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**

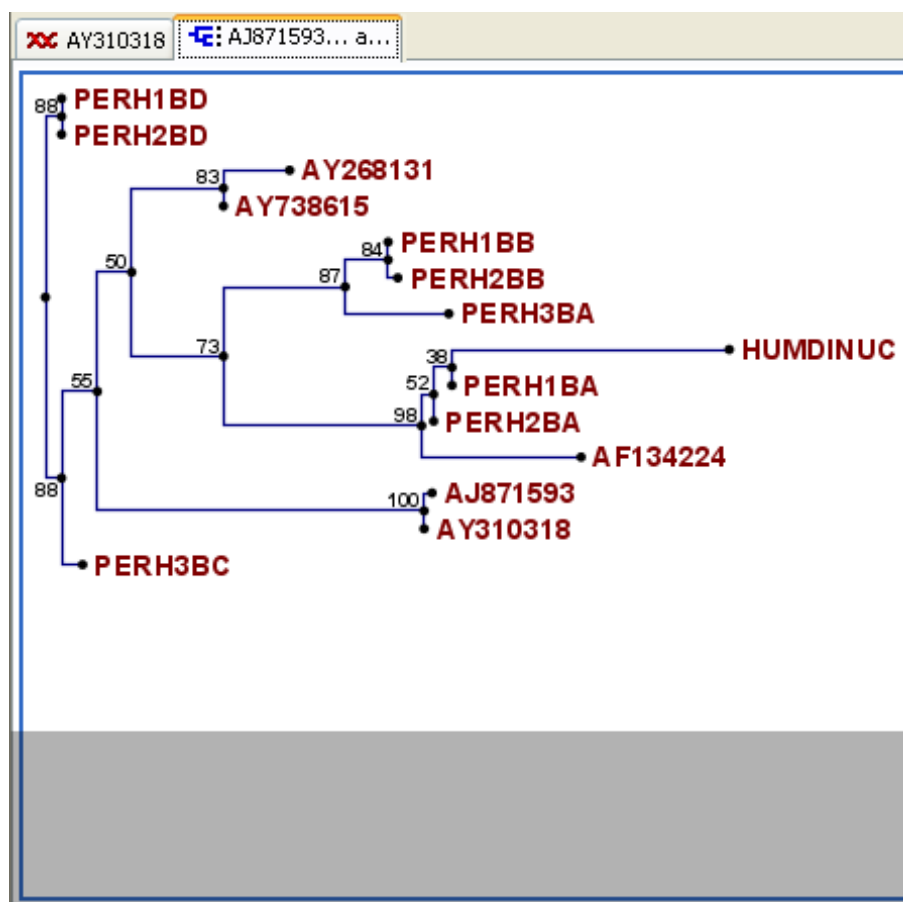


Figure 3.9: When dragging a View, a gray area indicates where the View will be shown.

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

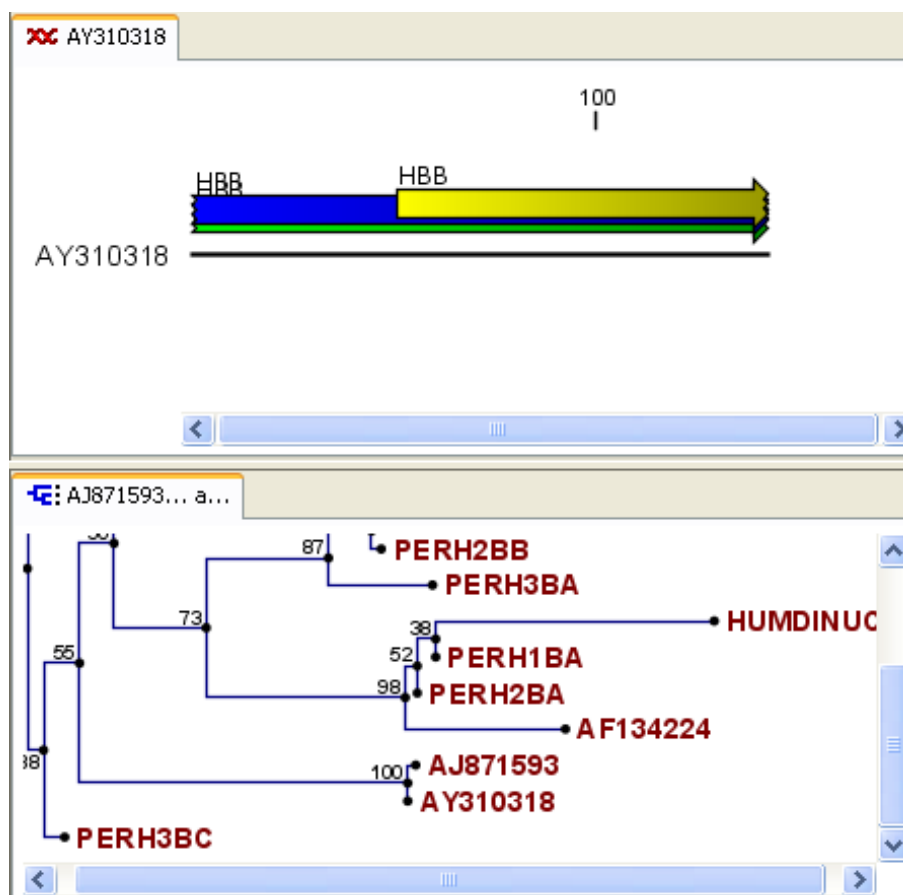




Figure 3.10: A horizontal split-screen. The two Views split the View Area.

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**

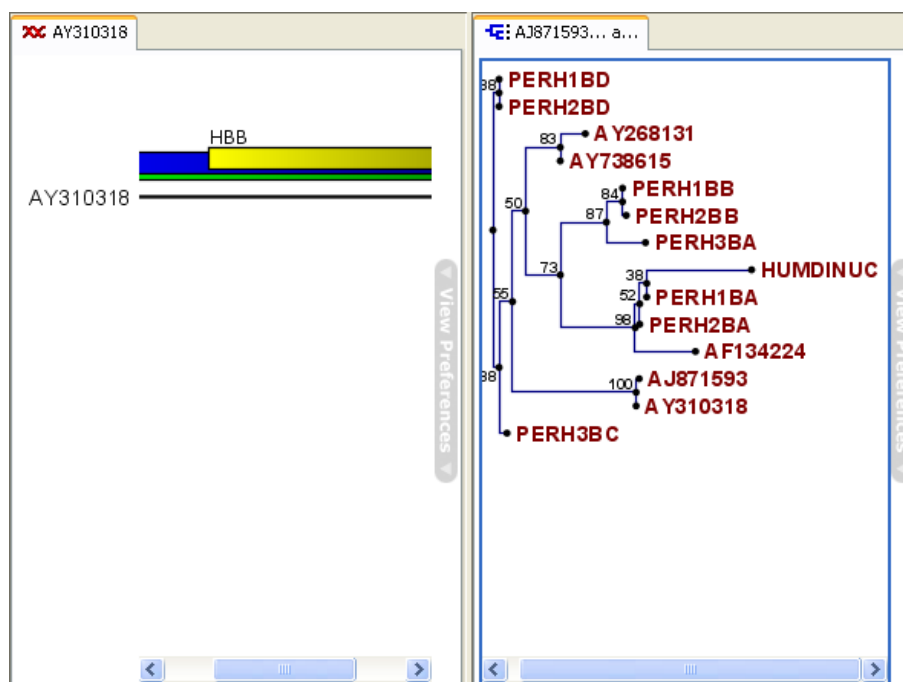


Figure 3.11: A vertical split-screen.

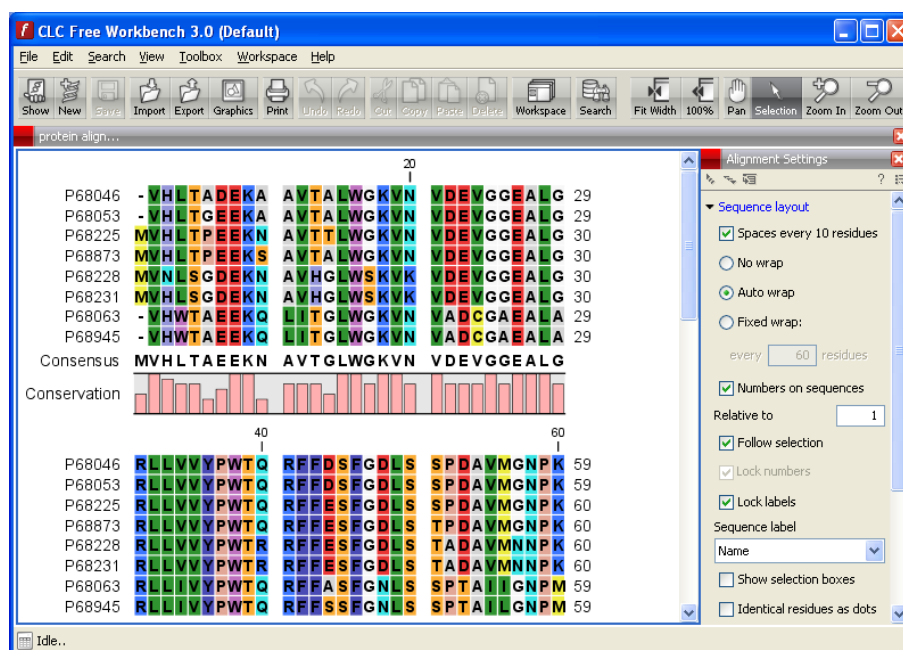


Figure 3.12: A maximized View. The function hides the Navigation Area and the Toolbox.

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 Free Workbench 3.0* 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.

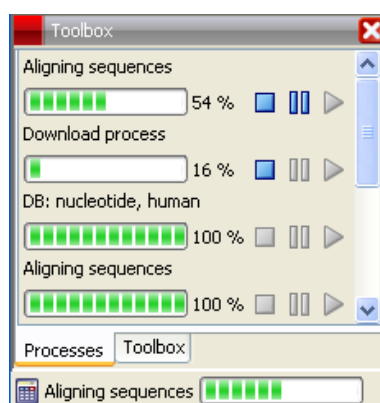


Figure 3.14: Two running, and a number of terminated processes in the Toolbox.

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 Free Workbench 3.0*. 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 Free Workbench 3.0* 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 Free 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 Free Workbench 3.0*" 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:

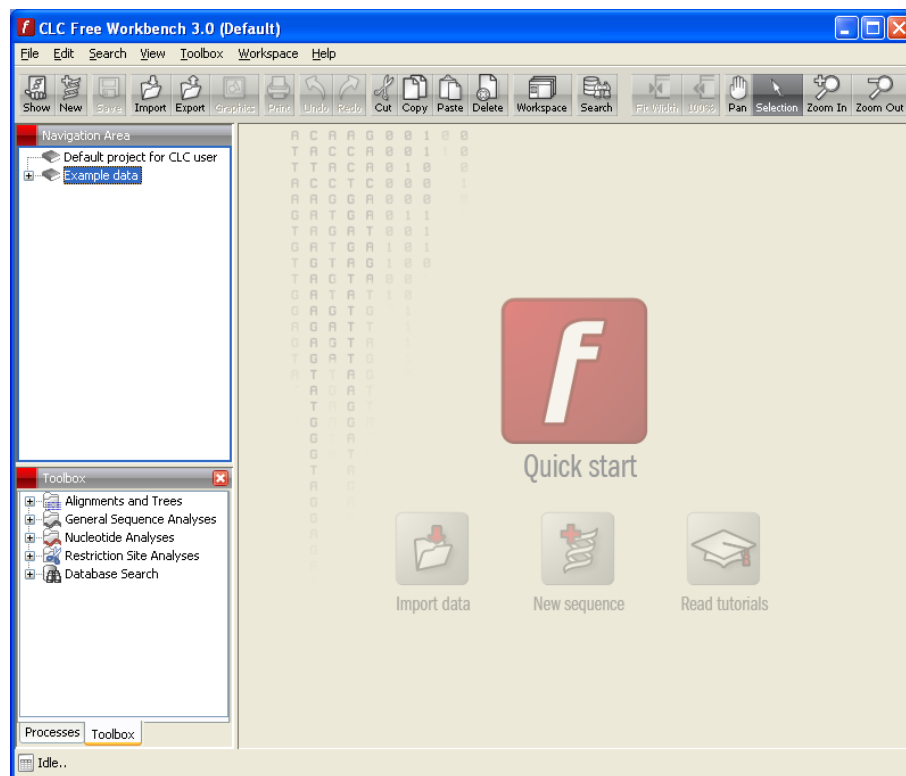


Figure 3.15: An empty Workspace.

**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 Free Workbench 3.0* 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

<b>4.1 General preferences</b>	<b>61</b>
<b>4.2 Default View preferences</b>	<b>61</b>
<b>4.3 Advanced preferences</b>	<b>62</b>
<b>4.4 Export/import of preferences</b>	<b>62</b>
<b>4.5 View preference style sheet</b>	<b>62</b>
4.5.1 Floating Side Panel	63

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)

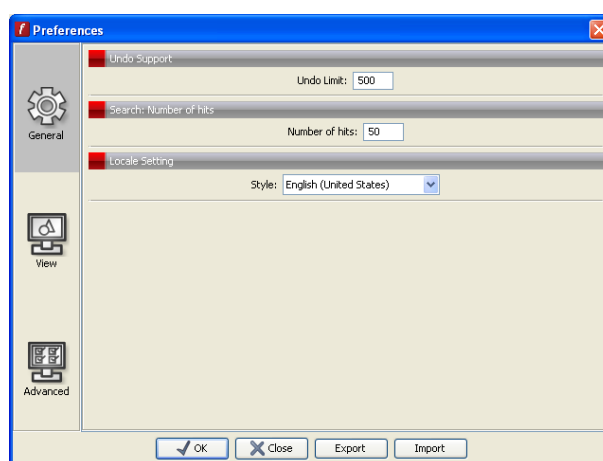


Figure 4.1: Preferences include General preferences, View preferences, Colors preferences, and Advanced settings.

## 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 Free Workbench 3.0*, 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 1.6 .

### 4.4 Export/import of preferences

The user preferences of the *CLC Free Workbench 3.0* 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.

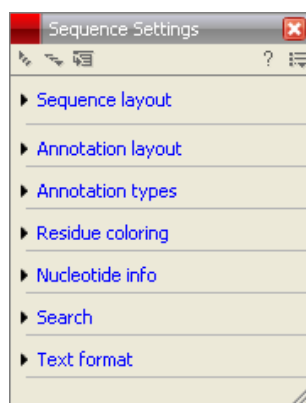


Figure 4.2: View preferences for a view of a sequence include several preference groups. In this case the groups are: Sequence layout, Annotation types, Annotation layout, etc. Several of these preference groups are present in more views. E.g. Sequence layout is also present when an alignment is viewed.

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 10.1.1 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).

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.

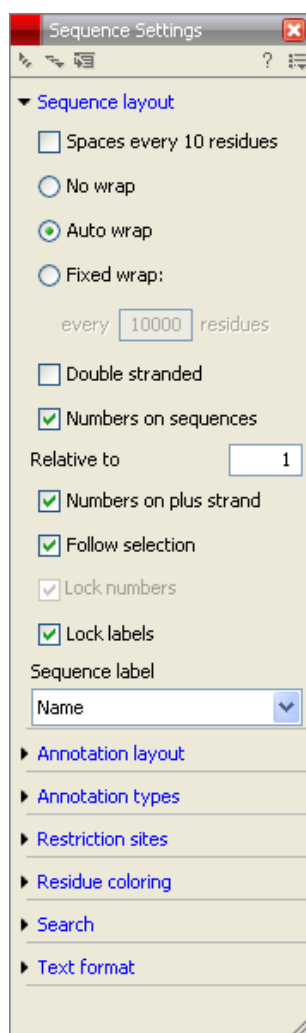


Figure 4.3: The many preferences for each view are stored in preference groups which can be opened and closed.



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

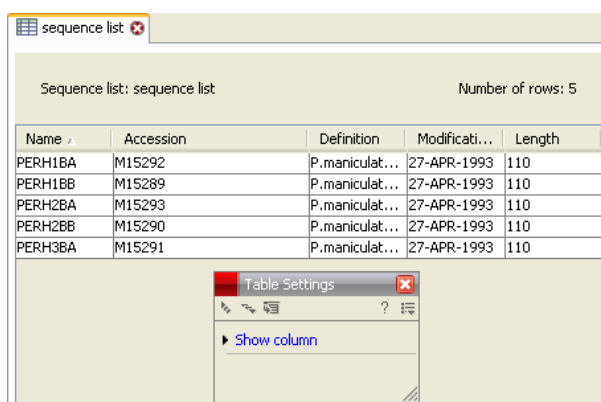


Figure 4.5: The floating Side Panel can be moved out of the way, e.g. to allow for a wider view of a table.



# Chapter 5

## Printing

### Contents

<b>5.1 Selecting which part of the view to print</b>	<b>65</b>
<b>5.2 Page setup</b>	<b>66</b>
<b>5.3 Print preview</b>	<b>66</b>

CLC Free Workbench 3.0 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).

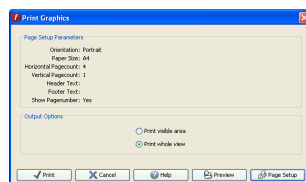


Figure 5.1: When printing graphics you get the options of printing the visible area or printing the whole view.

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

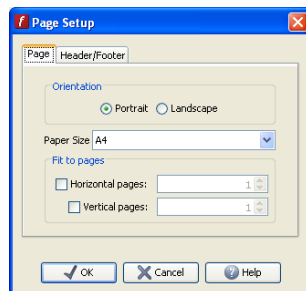



Figure 5.2: In this dialog the default settings Portrait and A4 apply to print of an alignment. By checking Fit to pages it is possible to adjust Horizontal pages to 2. This is done allow a long sequence to stretch the width of two A4 pages. This is illustrated in the Page Layout field.

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

<b>6.1 Bioinformatic data formats</b>	<b>68</b>
6.1.1 Import of bioinformatic data	69
6.1.2 Export of bioinformatic data	71
<b>6.2 External files</b>	<b>73</b>
6.2.1 Import external files	73
6.2.2 Export external files	73
6.2.3 Technical details	74
<b>6.3 Export graphics to files</b>	<b>74</b>
6.3.1 Exporting protein reports	76
<b>6.4 Copy/paste view output</b>	<b>76</b>

*CLC Free Workbench 3.0* 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 Free Workbench 3.0*, 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 Free Workbench 3.0* 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 Free Workbench* can import 'external' files, too. This means that *CLC Free 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 Free Workbench*.

The *CLC Free Workbench 3.0* 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 Free Workbench*. If *CLC Free 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):

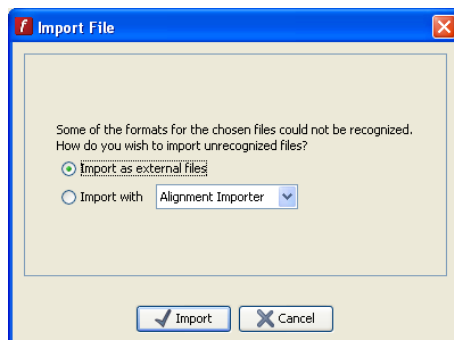


Figure 6.1: If the dragged file is not recognized by **CLC Free Workbench** the dialog allows you to "force" the import in a certain format.

**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 10.5 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 Free Workbench 3.0* 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 Free Workbench 3.0* 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 Free Workbench 3.0*.

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.

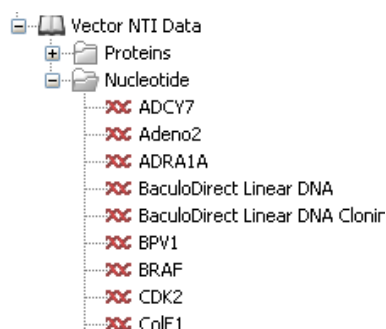


Figure 6.2: Project Vector NTI Data containing all imported sequences of the Vector NTI Database.

### 6.1.2 Export of bioinformatic data

*CLC Free Workbench 3.0* 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 Free Workbench 3.0* 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 Free 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 Free 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 Free Workbench 3.0* 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 Free Workbench 3.0*. Importing an external file creates a copy of the file which is saved in a project in *CLC Free Workbench 3.0*. 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 Free 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 Free Workbench 3.0*.

### 6.2.2 Export external files

If you export an entire project or folder from *CLC Free Workbench 3.0*, 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 Free Workbench 3.0*:

**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 Free Workbench 3.0* 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 Free Workbench 3.0*, 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 Free Workbench 3.0* is closed.

If you have made changes to a file after the *CLC Free Workbench 3.0* 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).

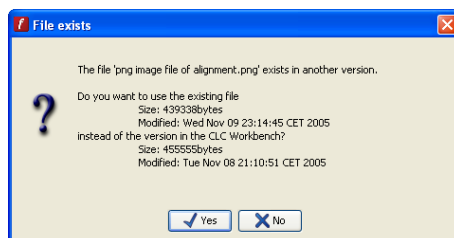


Figure 6.3: A dialog asking which version of the file you want to keep.

## 6.3 Export graphics to files

*CLC Free Workbench 3.0* 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 Free Workbench 3.0* 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.

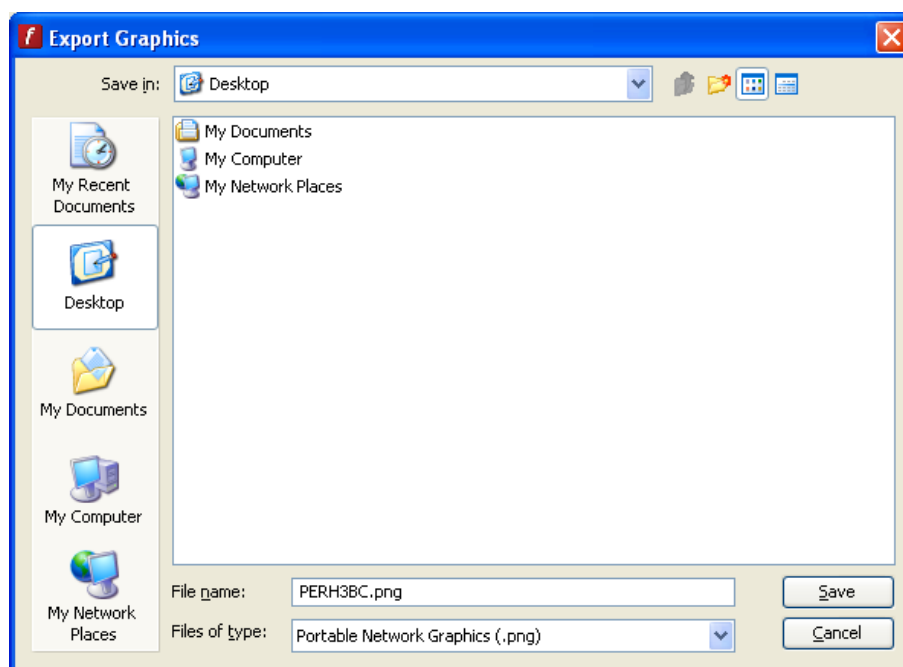


Figure 6.4: Exporting a phylogenetic tree from **CLC Free Workbench 3.0**.

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.

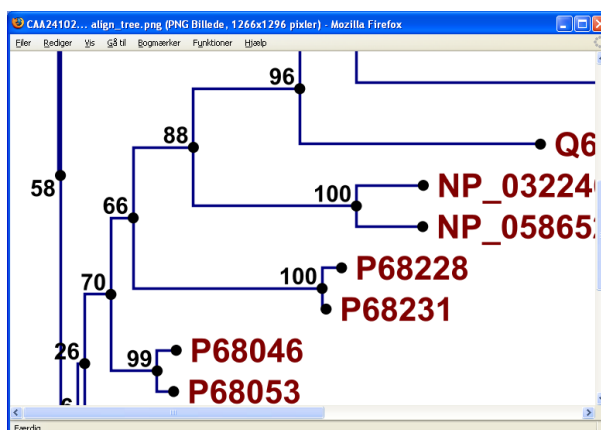


Figure 6.5: The exported .png-file opened in a browser. (Due to high resolution of the exported graphics, it is not possible to see the entire file in the browser window.)

The following file types are available for exporting graphics in *CLC Free Workbench 3.0*:

### 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 Free Workbench 3.0*. See section 6.2.1 for more about importing external files into *CLC Free Workbench 3.0*.

### 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.

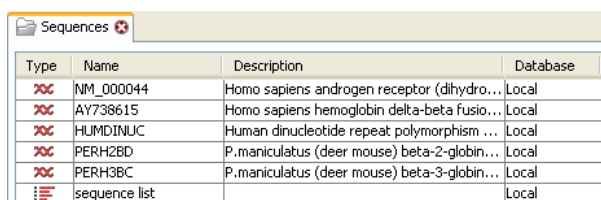
## 6.4 Copy/paste view output

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 Free Workbench 3.0* 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 Free Workbench 3.0* can be produced. (Except the icons which are replaced by file references in Excel.)

# Chapter 7

## History

### Contents

<b>7.1 Element history</b> . . . . .	<b>78</b>
7.1.1 Sharing data with history . . . . .	79

*CLC Free Workbench 3.0* 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 Free Workbench 3.0*.

### 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 Free Workbench 3.0*. 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.

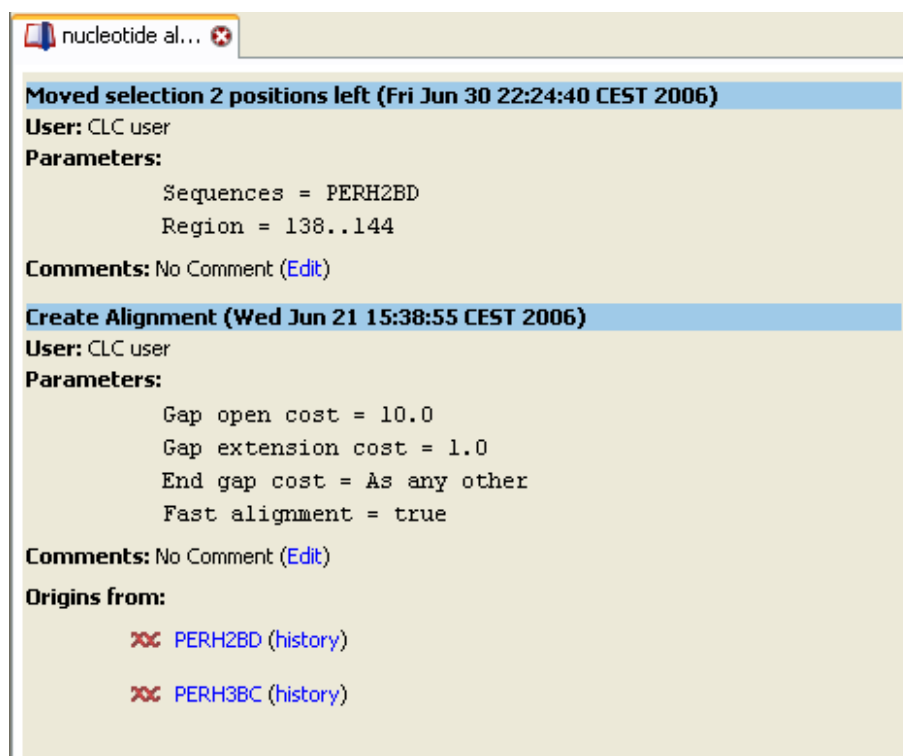


Figure 7.1: An element's history.

- **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

<b>8.1 How to handle results of analyses</b> . . . . .	<b>80</b>
8.1.1 When the analysis does not create new elements . . . . .	80
8.1.2 Batch log . . . . .	81

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):



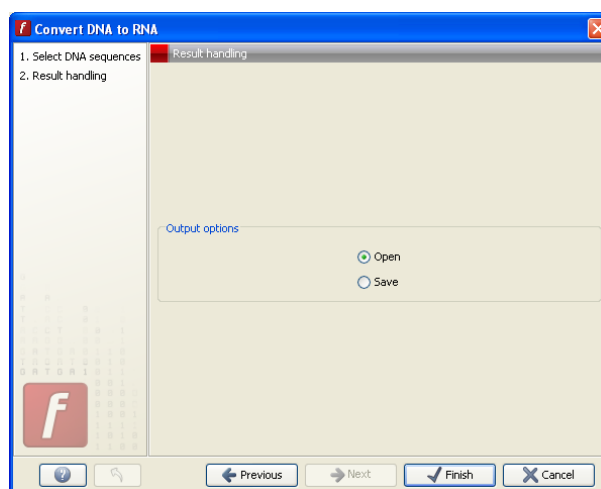


Figure 8.1: The last step of the analyses exemplified by Translate DNA to RNA.

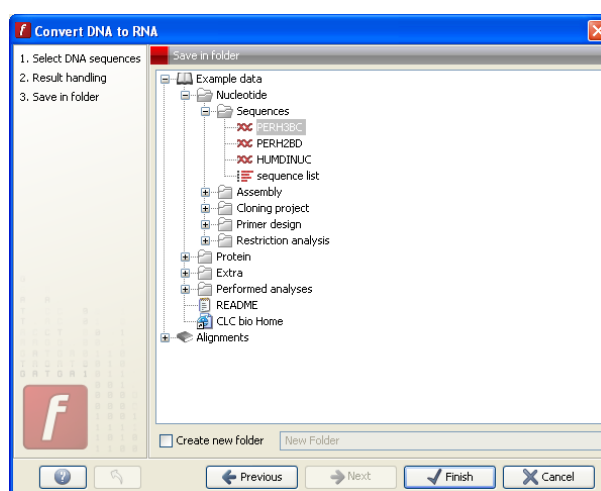


Figure 8.2: Specify a folder for the results of the analysis.

- **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.

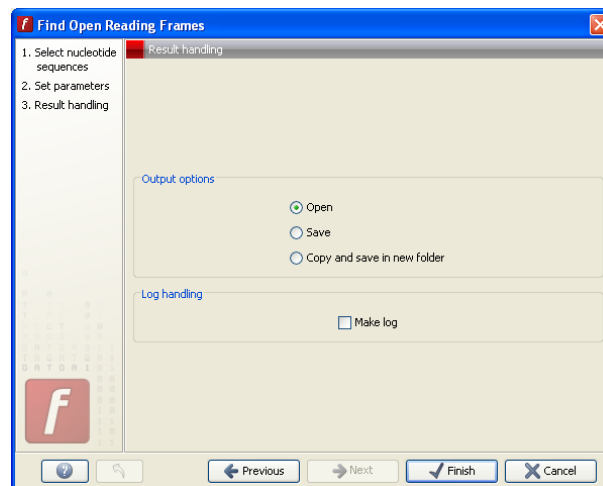


Figure 8.3: The final step when the analysis does not create new elements but add annotations to existing elements.

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
HJMDNUC	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

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

**Part III**

**Bioinformatics**

# Chapter 9

## Database search

### Contents


<b>9.1 GenBank search</b> . . . . .	<b>84</b>
9.1.1 GenBank search options . . . . .	84
9.1.2 Handling of GenBank search results . . . . .	86

CLC Free Workbench 3.0 allows you to search the for sequences on the Internet. You must be online when initiating and performing searches in NCBI.

### 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 Free Workbench 3.0* corresponds to conducting the search on NCBI's website. When conducting the search from *CLC Free Workbench 3.0*, 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.

As default, *CLC Free 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".

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.

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](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.

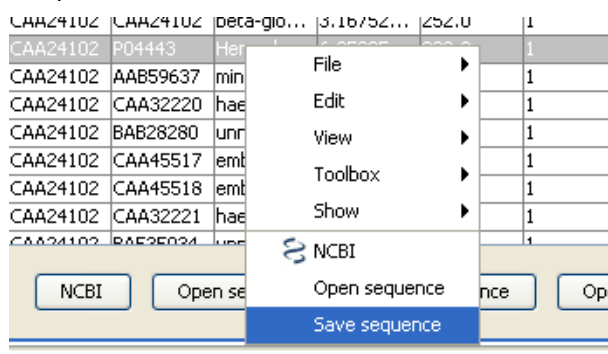


Figure 9.2: By right-clicking a search result, it is possible to choose how to handle the relevant sequence.

### 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.

# Chapter 10

## Viewing and editing sequences

### Contents

<b>10.1 View sequence</b>	<b>88</b>
10.1.1 Sequence Layout in Side Panel	89
10.1.2 Selecting parts of the sequence	93
10.1.3 Editing the sequence	94
10.1.4 Removing annotations	94
10.1.5 Sequence region types	94
<b>10.2 Sequence information</b>	<b>94</b>
10.2.1 Annotation map	95
<b>10.3 View as text</b>	<b>96</b>
<b>10.4 Creating a new sequence</b>	<b>96</b>
<b>10.5 Sequence Lists</b>	<b>98</b>
10.5.1 Graphical view of sequence lists	99
10.5.2 Sequence list table	99
10.5.3 Extract sequences	100
<b>10.6 Circular DNA</b>	<b>100</b>
10.6.1 Using split views to see details of the circular molecule	101
10.6.2 Mark molecule as circular and specify starting point	101

*CLC Free Workbench 3.0* 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.

### 10.1 View sequence

When you double-click a sequence in the **Navigation Area**, the sequence will open automatically, and you will see the nucleotides or amino acids. The zoom options described in section 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 14.2).



### 10.1.1 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 (X) 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** analysis, the cut sites can be displayed as annotations on the sequence. Other analyses also attach annotations on the sequence. See section [10.1.5](#) 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.

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 10.1).



Figure 10.1: 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 two 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 13.2.

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.

## 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.

### 10.1.2 Selecting parts of the sequence

You can select parts of a sequence:

**Click Selection (  ) in Toolbar | Press and hold down the mouse button on the sequence where you want the selection to start | move the mouse to the end of the selection while holding the button | release the mouse button**

Alternatively, you can search for a specific interval using the 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.

### 10.1.3 Editing the sequence

When you make a selection, it can be edited by:

**right-click the selection | Edit selection**

A dialog appears displaying the sequence. You can add, remove or change the text and click **OK**. The original selected part of the sequence is now replaced by the sequence entered in the dialog. This dialog also allows you to paste text into the sequence using Ctrl + V (⌘ + V on Mac). If you delete the text in the dialog and press **OK**, the selected text on the sequence will also be deleted. Another way to delete a part of the sequence is to:

**right-click the selection | Delete selection**

### 10.1.4 Removing annotations

Annotations can be hidden using the **Annotation Types** preferences in the **Side Panel** to the right of the view (see section 10.1.1). 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.

### 10.1.5 Sequence region types

The various annotations on sequences cover parts of the sequence. Some cover an interval, some cover intervals with unknown endpoints, some cover more than one interval etc. In the following, all of these will be referred to as *regions*. Regions are generally illustrated by markings (often arrows) on the sequences. An arrow pointing to the right indicates that the corresponding region is located on the positive strand of the sequence. Figure 10.2 is an example of three regions with separate colors.

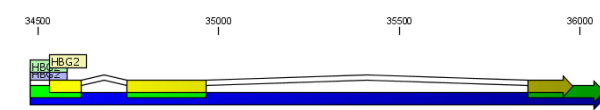


Figure 10.2: Three regions on a human beta globin DNA sequence (HUMHBB).

Figure 10.3 shows an artificial sequence with all the different kinds of regions.

## 10.2 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

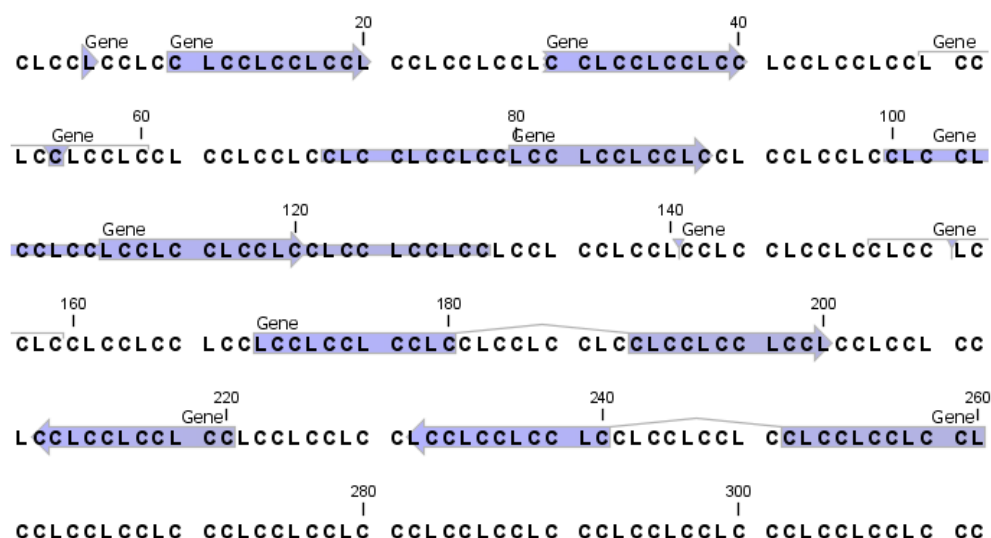


Figure 10.3: Region #1: A single residue, Region #2: A range of residues including both endpoints, Region #3: A range of residues starting somewhere before 30 and continuing up to and including 40, Region #4: A single residue somewhere between 50 and 60 inclusive, Region #5: A range of residues beginning somewhere between 70 and 80 inclusive and ending at 90 inclusive, Region #6: A range of residues beginning somewhere between 100 and 110 inclusive and ending somewhere between 120 and 130 inclusive, Region #7: A site between residues 140 and 141, Region #8: A site between two residues somewhere between 150 and 160 inclusive, Region #9: A region that covers ranges from 170 to 180 inclusive and 190 to 200 inclusive, Region #10: A region on negative strand that covers ranges from 210 to 220 inclusive, Region #11: A region on negative strand that covers ranges from 230 to 240 inclusive and 250 to 260 inclusive.

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 10.4.

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**.

### 10.2.1 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 10.5).

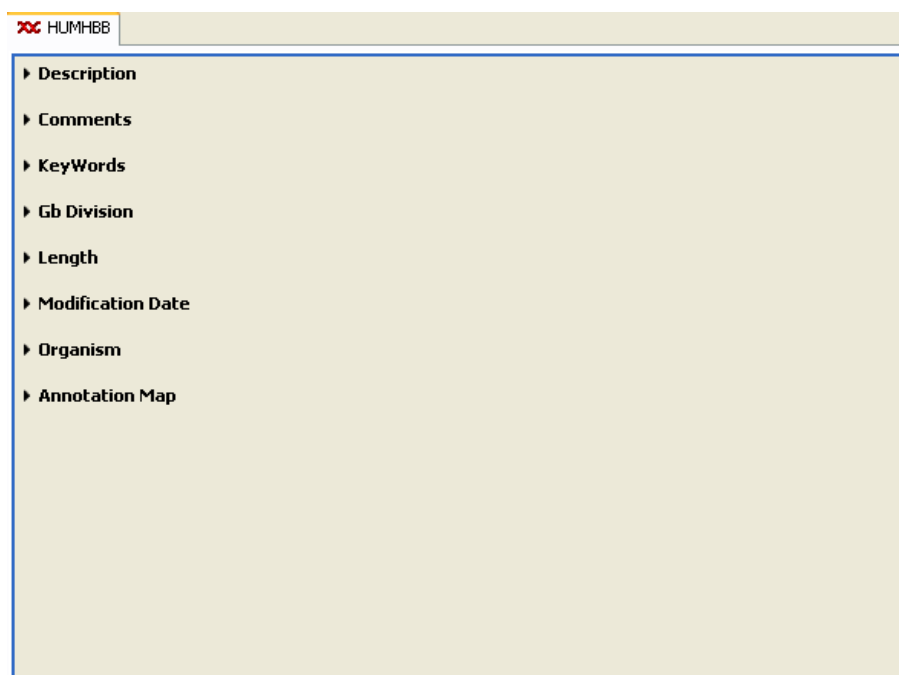


Figure 10.4: The initial display of sequence info for the HUMHBB DNA sequence from the Example data.

### 10.3 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 10.2.)

### 10.4 Creating a new sequence

A sequence can either be imported, downloaded from an online database or created in the *CLC Free Workbench 3.0*. This section explains how to create a new sequence:

**New(+) in the toolbar**

The **Create Sequence** dialog (figure 10.6) 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.



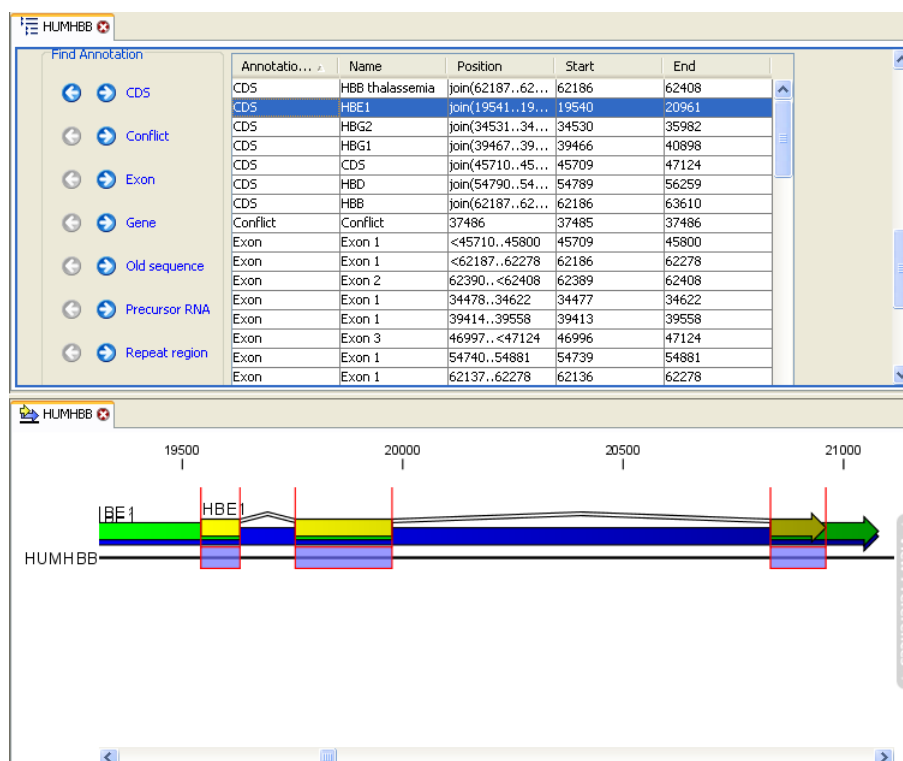


Figure 10.5: Clicking a sequence map annotation in the sequence information view, selects the annotation on the (normal) sequence view.

The 'Create Sequence' dialog box is shown. It has a title bar 'f Create Sequence' and a close button. The main area is divided into two panes: '1. Enter Sequence Data' and '2. Enter Sequence Data'. The '1. Enter Sequence Data' pane contains the following fields:

- Name: Globin
- Common name: Human
- Species: Homo sapiens
- Type: DNA (selected), RNA, Protein
- ☐ Circular
- Description: Globin sequence
- Keywords:
- Comments:

The '2. Enter Sequence Data' pane contains a text area with the sequence:

```
1 TCTAATCT
8 CCTCTCAACCTACAGTACCCATTGGTATATATAA
```

At the bottom, there are buttons for 'Previous', 'Next', 'Finish', and 'Cancel'.

Figure 10.6: Creating a sequence.

- **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**.

## 10.5 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:

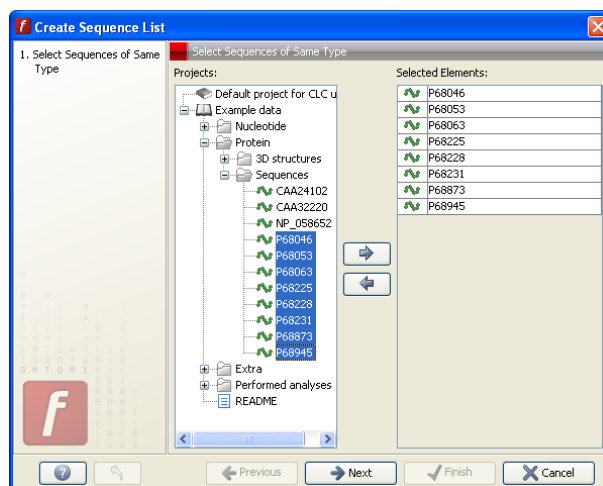


Figure 10.7: A Sequence List dialog.

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 10.8.

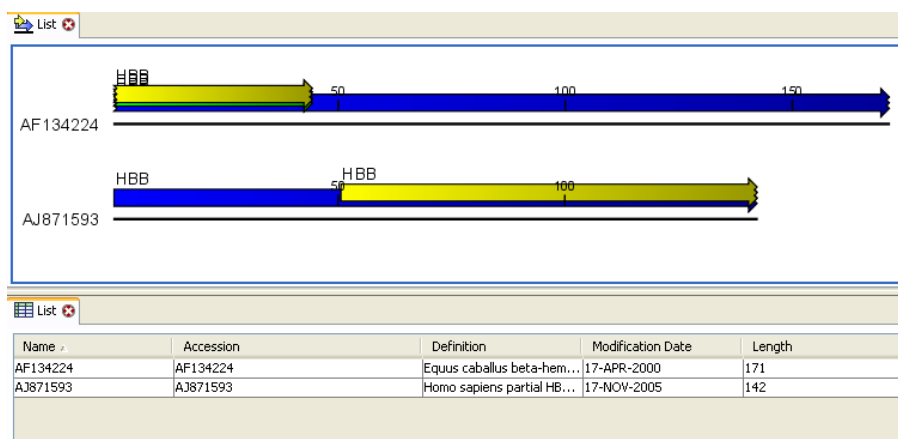


Figure 10.8: A sequence list of two sequences can be viewed in either a table or in a graphical sequence list.

### 10.5.1 Graphical view of sequence lists

The graphical view of sequence lists is almost identical to the view of single sequences (see section 10.1). The main difference is that you now can see more than one sequence in the same view.

However, you also have a few extra options for sorting, deleting and adding sequences:

- To add extra sequences to the list, right-click an empty (white) space in the view, and select **Add Sequences**.
- To delete a sequence from the list, right-click the sequence's 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**.

### 10.5.2 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.

### 10.5.3 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.

## 10.6 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 10.9.

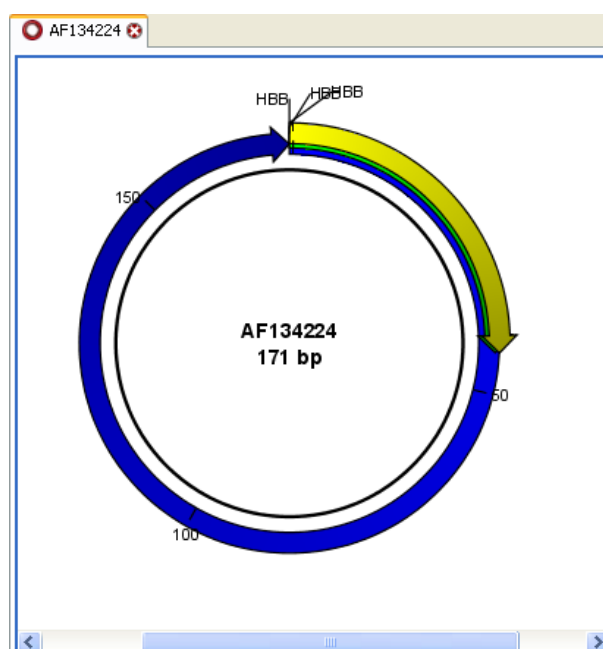


Figure 10.9: A molecule shown in a circular view.

This view of the sequence shares some of the properties of the linear view of sequences as described in section 10.1, but there are some differences. The similarities and differences are listed below:

- **Similarities:**

- **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).

### 10.6.1 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 10.10.

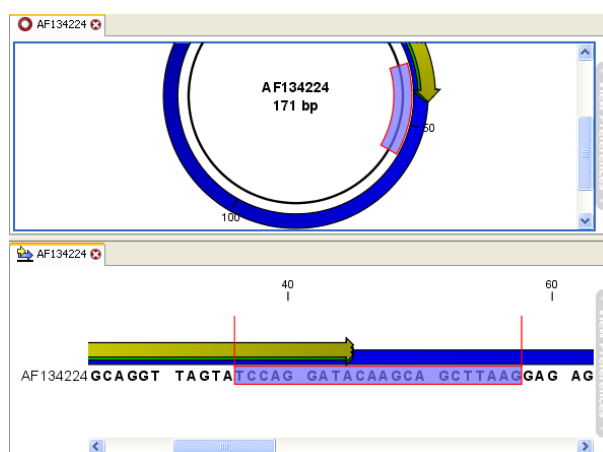


Figure 10.10: Two views showing the same sequence. The bottom view is zoomed in.

**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.

### 10.6.2 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 11

## General sequence analyses

### Contents

<b>11.1 Sequence statistics</b> . . . . .	<b>102</b>
11.1.1 Sequence statistics output . . . . .	105
<b>11.2 Shuffle sequence</b> . . . . .	<b>105</b>
<b>11.3 Join sequences</b> . . . . .	<b>105</b>

CLC Free Workbench 3.0 offers different kinds of sequence analyses, which apply to both protein and DNA.

### 11.1 Sequence statistics

CLC Free Workbench 3.0 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**.

This opens the dialog displayed in figure 11.1.

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.

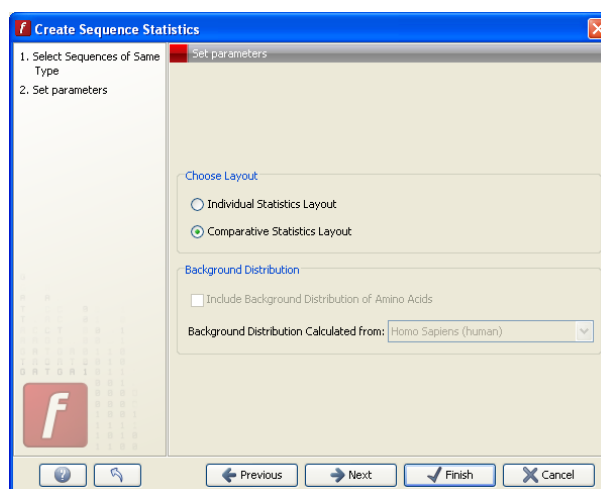


Figure 11.1: Setting parameters for the sequence statistics.

- **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](http://www.uniprot.org) version 6.0, dated September 13 2005.)

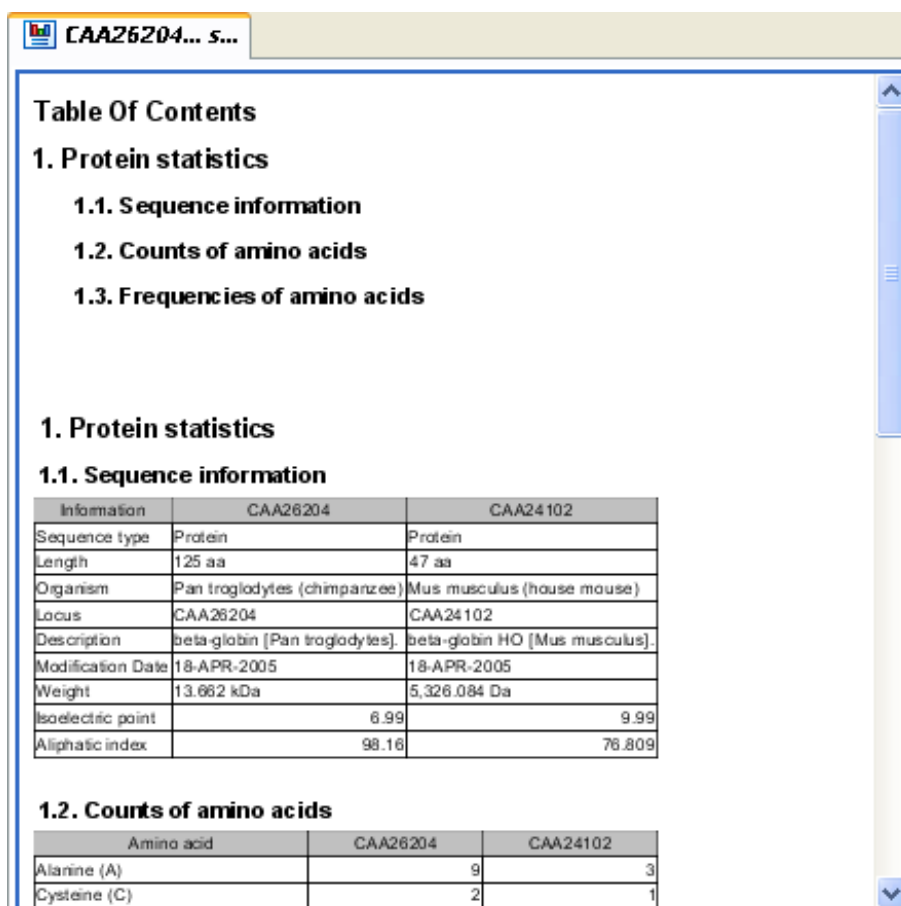
Click **Next** if you wish to adjust how to handle the results (see section 8.1). If not, click **Finish**. An example of protein sequence statistics is shown in figure 11.2.

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
  - Weight
  - Isoelectric point
  - Aliphatic index
- Amino acid distribution



**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 11.2: Comparative sequence statistics.

- Annotation table

The output of nucleotide sequence statistics include:

- General statistics:
  - Sequence type
  - Length
  - Organism
  - Locus
  - Description
  - Modification Date
  - Weight
- Nucleotide distribution table
- Annotation table

**Notice!** This section also describes statistics not available in *CLC Free Workbench*.



### 11.1.1 Sequence statistics output

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

## 11.2 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 11.3:

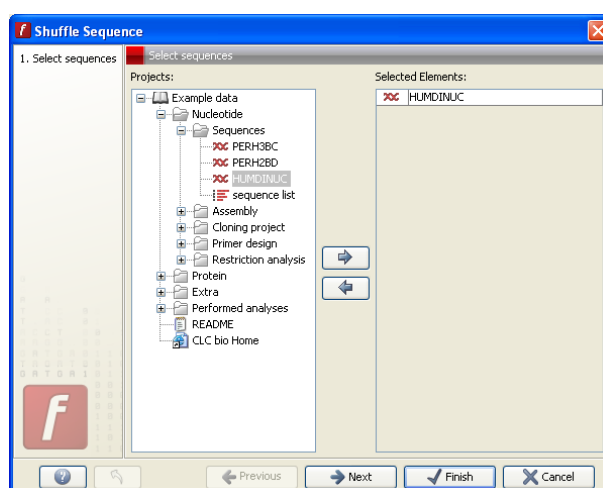


Figure 11.3: Choosing sequence for shuffling.

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.

## 11.3 Join sequences

CLC Free 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 11.4.

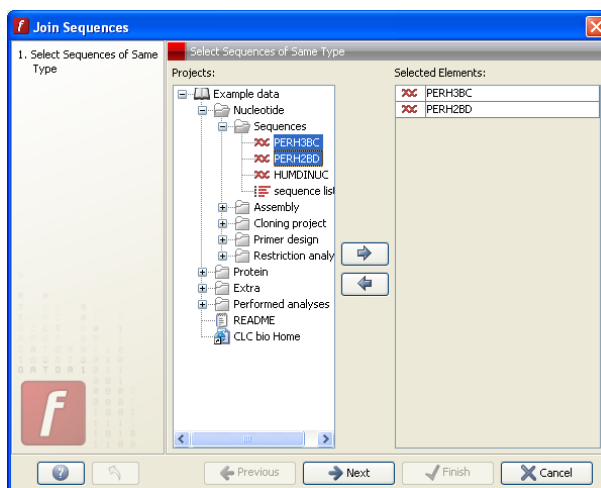


Figure 11.4: Selecting two alignments to be joined.

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 11.5.

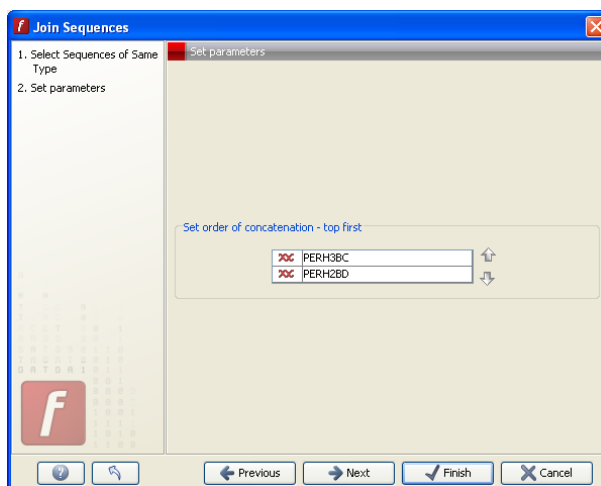


Figure 11.5: Setting the order in which sequences are joined.

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 11.6.

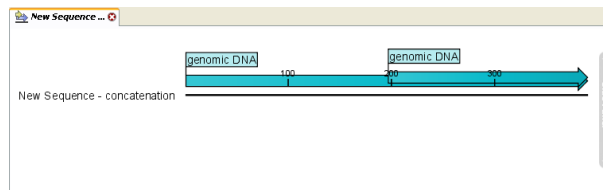


Figure 11.6: *The result of joining sequences is a new sequence containing all the annotations of the joined sequences.*

# Chapter 12

## Nucleotide analyses

### Contents

<b>12.1 Convert DNA to RNA</b>	<b>108</b>
<b>12.2 Convert RNA to DNA</b>	<b>109</b>
<b>12.3 Reverse complements of sequences</b>	<b>110</b>
<b>12.4 Translation of DNA or RNA to protein</b>	<b>111</b>
<b>12.5 Find open reading frames</b>	<b>111</b>
12.5.1 Open reading frame parameters	113

CLC Free Workbench 3.0 offers different kinds of sequence analyses, which only apply to DNA and RNA.

### 12.1 Convert DNA to RNA

CLC Free Workbench 3.0 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 12.1:

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.

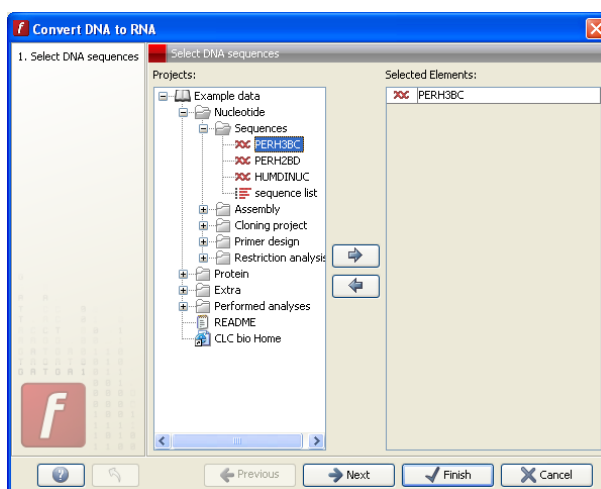


Figure 12.1: Translating DNA to RNA.

## 12.2 Convert RNA to DNA

CLC Free Workbench 3.0 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 12.2:

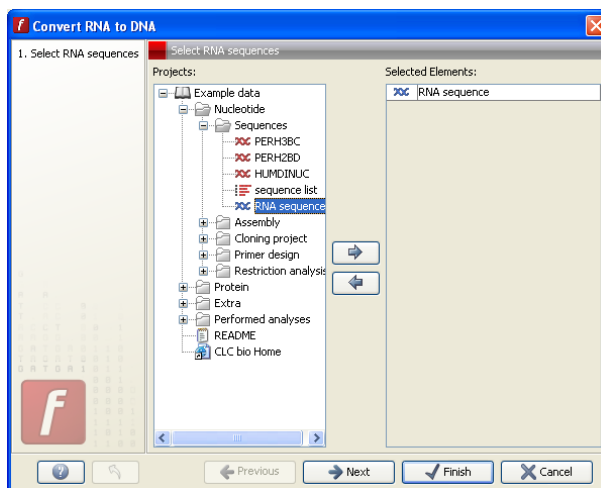


Figure 12.2: Translating RNA to DNA.

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.

## 12.3 Reverse complements of sequences

CLC Free Workbench 3.0 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 12.3:

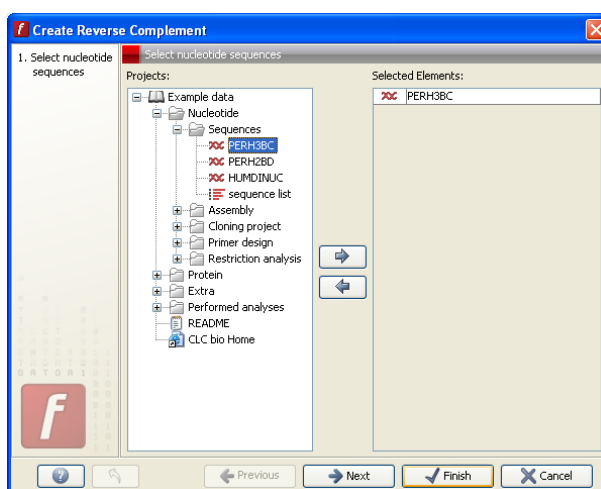


Figure 12.3: Creating a reverse complement sequence.

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.

## 12.4 Translation of DNA or RNA to protein

In *CLC Free Workbench 3.0* 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 12.4:

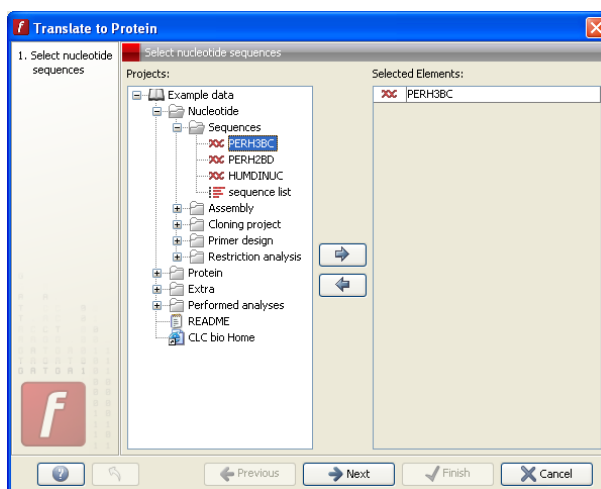


Figure 12.4: Choosing sequences for translation.

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 12.5:

The translation tables in *CLC Free 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.

To save a protein sequence, drag it into the **Navigation Area** or press Ctrl + S (⌘ + S on Mac) to activate a save dialog.

## 12.5 Find open reading frames

*CLC Free Workbench 3.0* 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

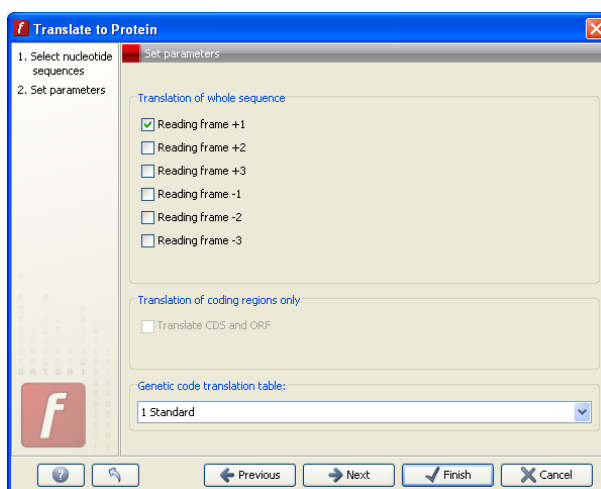


Figure 12.5: Choosing +1 and +3 reading frames, and the standard translation table.

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 12.6:

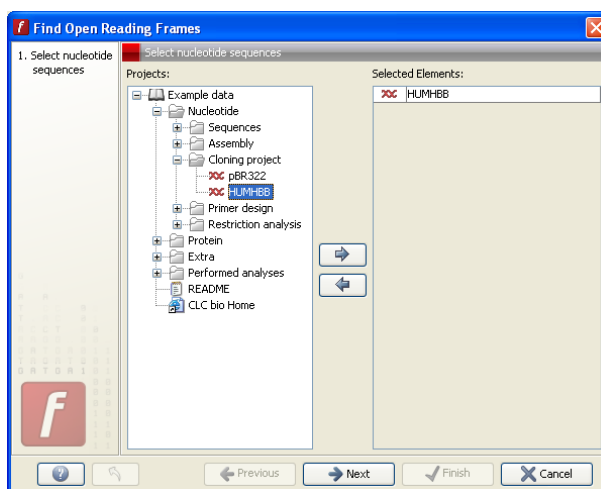


Figure 12.6: Create Reading Frame dialog.

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**.



### 12.5.1 Open reading frame parameters

This opens the dialog displayed in figure 12.7:

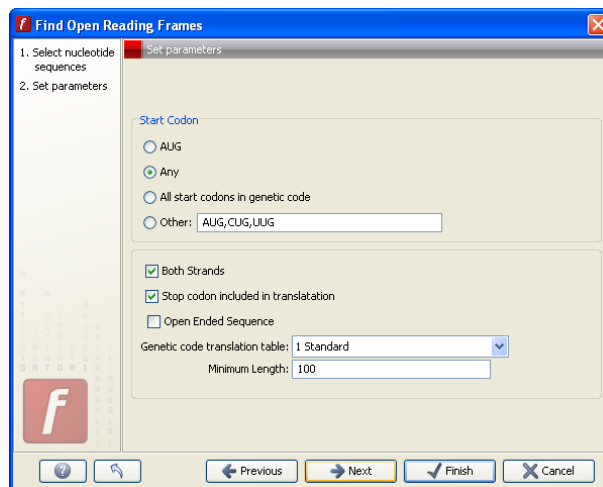


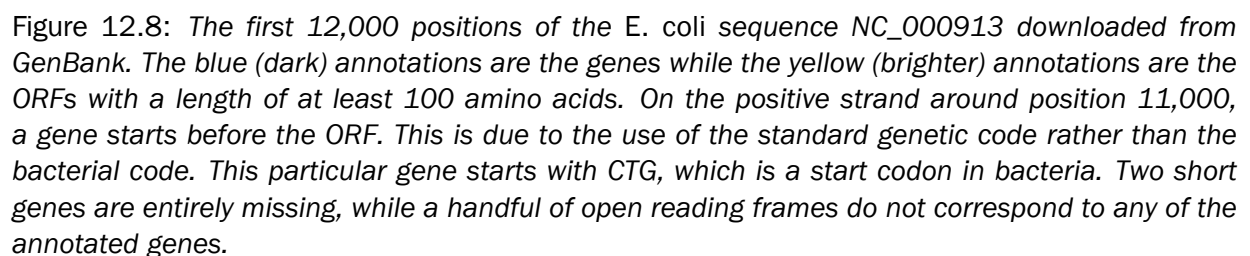
Figure 12.7: Create Reading Frame dialog.

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 12.8).

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 13

## Restriction site analyses

### Contents

<b>13.1 Restriction sites and enzyme lists</b>	<b>115</b>
<b>13.2 Restriction site analysis</b>	<b>115</b>
13.2.1 Restriction site parameters	115
<b>13.3 Restriction enzyme lists</b>	<b>118</b>
13.3.1 Create enzyme list	118
13.3.2 Modify enzyme list	119

### 13.1 Restriction sites and enzyme lists

*CLC Free Workbench 3.0* offers the opportunity to detect restriction sites. First the restriction site analysis is described and next, the functionalities regarding enzyme lists are explained.

### 13.2 Restriction site analysis

This section explains how to adjust the detection parameters and offers basic information with respect to restriction site algorithms.

#### 13.2.1 Restriction site parameters

Given a DNA sequence, *CLC Free Workbench 3.0* 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 (✂️)**

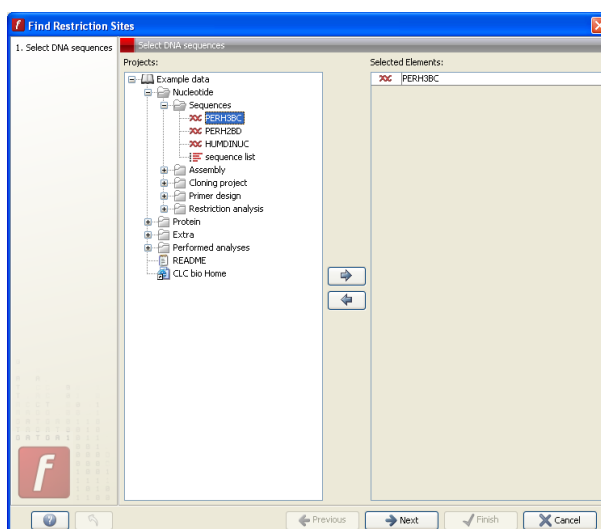


Figure 13.1: Choosing sequence PERH3BC.

The result of these steps can be seen in figure 13.1.

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**.

Clicking **Next** generates the dialog shown in figure 13.2.

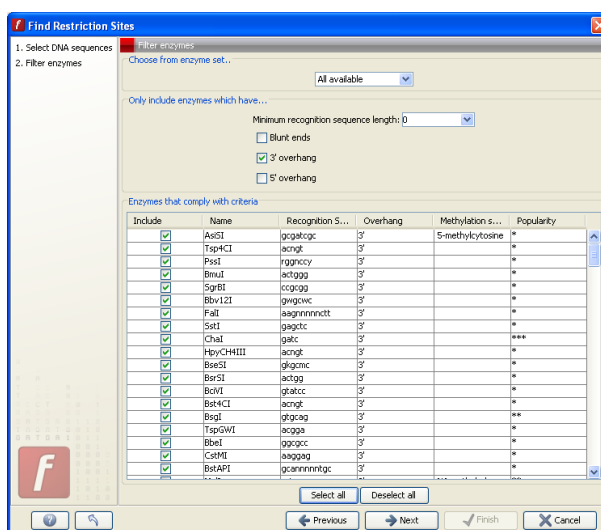


Figure 13.2: Selecting enzymes.

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 13.3 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 13.3.

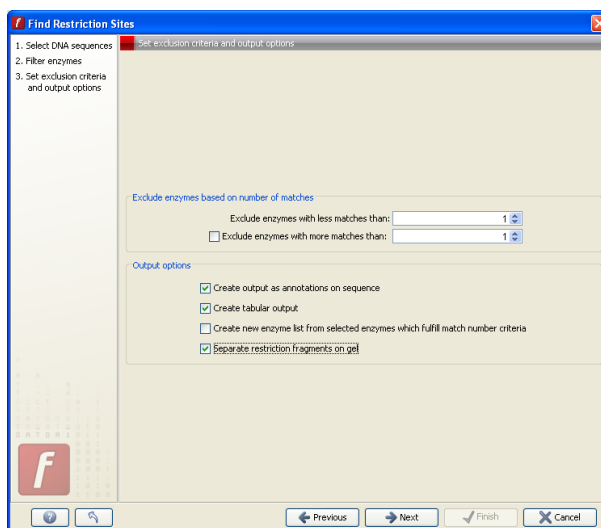


Figure 13.3: Exclusion criteria and output options.

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

In order to complete the analysis click **Finish**. The result is shown in figure 13.4.

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:

**Right-click the tab | File | Save()**

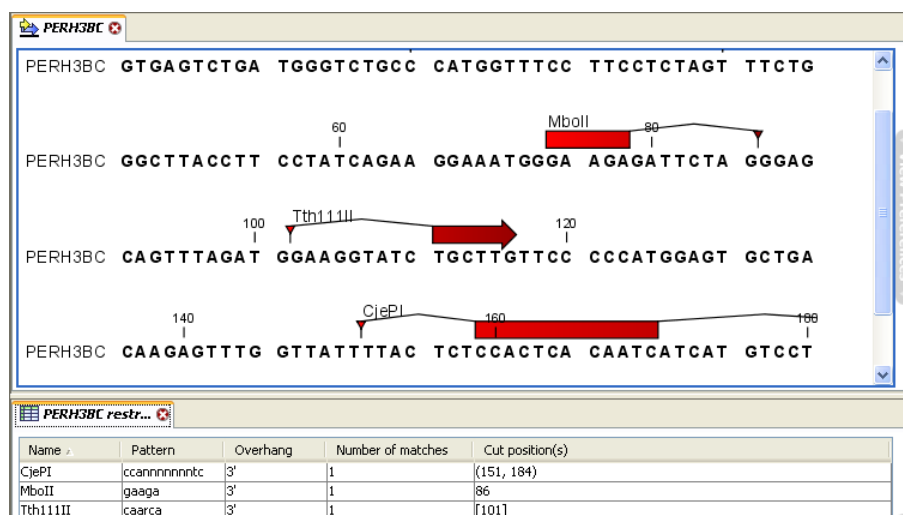


Figure 13.4: The result of the restriction site detection is displayed as text, and in this example the View Shares the View Area with a View of the PERH3BC sequence displaying the restriction sites (split-screen-view).

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 Free 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)**

## 13.3 Restriction enzyme lists

*CLC Free 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.

### 13.3.1 Create enzyme list

*CLC Free Workbench 3.0* 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 13.5

**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.

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**

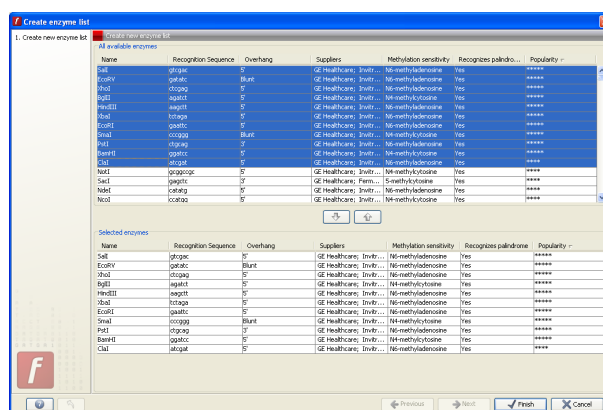


Figure 13.5: *Choosing enzymes for the new enzyme list.*

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.

### 13.3.2 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 13.6.

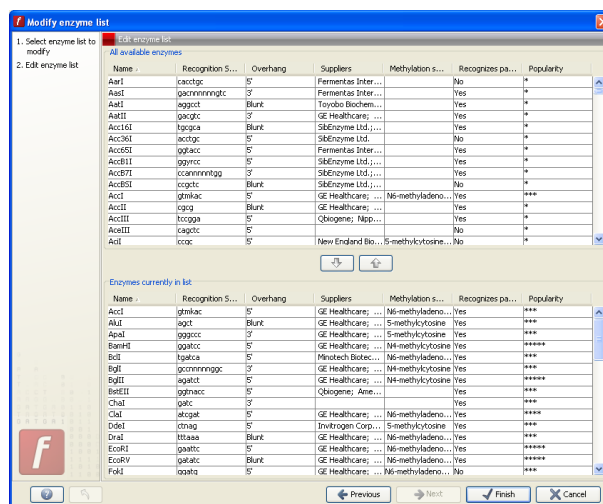


Figure 13.6: Adding and removing enzymes in the existing enzyme list.

Select sequences in either top or bottom table (see 13.3.1). Use the arrows to add and remove sequences. Click **Finish** to see the modified list.

# Chapter 14

## Sequence alignment

### Contents

<b>14.1 Create an alignment</b>	<b>120</b>
14.1.1 Gap costs	121
14.1.2 Fast or accurate alignment algorithm	122
<b>14.2 View alignments</b>	<b>123</b>
14.2.1 Conservation	124
<b>14.3 Edit alignments</b>	<b>124</b>
14.3.1 Move residues and gaps	124
14.3.2 Insert gap columns	125
14.3.3 Delete residues and gaps	125
14.3.4 Move sequences up and down	125
14.3.5 Delete sequences	126
<b>14.4 Bioinformatics explained: Multiple alignments</b>	<b>126</b>
14.4.1 Use of multiple alignments	126
14.4.2 Constructing multiple alignments	126

CLC Free Workbench 3.0 can align nucleotides and proteins using a *progressive alignment* algorithm (see section 14.4 or read the White paper on alignments in the **Science** section of <http://www.clcbio.com>).

This chapter describes how to use the program to align sequences. The chapter also describes alignment algorithms in more general terms.

### 14.1 Create an alignment

To create an alignment in CLC Free Workbench 3.0:

**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 14.1.



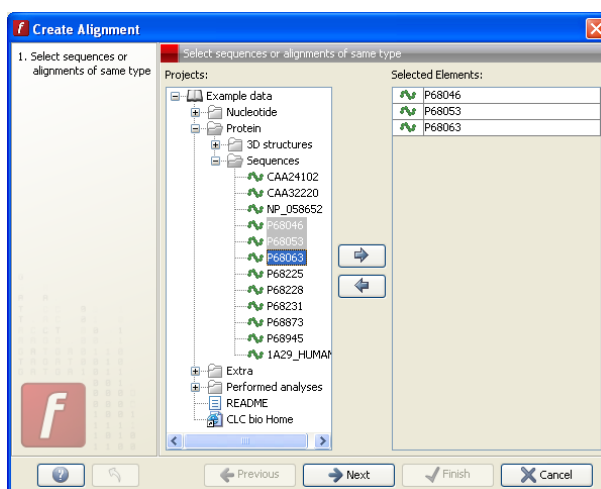


Figure 14.1: Creating an alignment.

If you have selected some elements before choosing the Toolbox action, they are now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences, sequence lists or alignments from the **Project Tree**. Click **Next** to adjust alignment algorithm parameters. Clicking **Next** opens the dialog shown in figure 14.2.

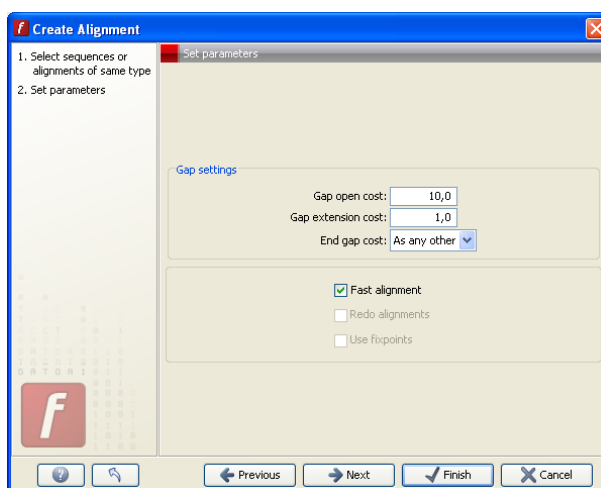


Figure 14.2: Adjusting alignment algorithm parameters.

### 14.1.1 Gap costs

The alignment algorithm has three parameters concerning gap costs: Gap open cost, Gap extension cost and End gap cost. The precision of these parameters is to one place of decimal.

- **Gap open cost.** The price for introducing gaps in an alignment.
- **Gap extension cost.** The price for every extension past the initial gap.

If you expect a lot of small gaps in your alignment, the Gap open cost should equal the Gap extension cost. On the other hand, if you expect few but large gaps, the Gap open cost should be set significantly higher than the Gap extension cost.

However, for most alignments it is a good idea to make the Gap open cost quite a bit higher than the Gap extension cost. The default values are 10.0 and 1.0 for the two parameters, respectively.

- **End gap cost.** The price of gaps at the beginning or the end of the alignment. One of the advantages of the *CLC Free Workbench 3.0* 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 14.3 and 14.4 illustrate the differences between the different gap scores at the sequence ends.

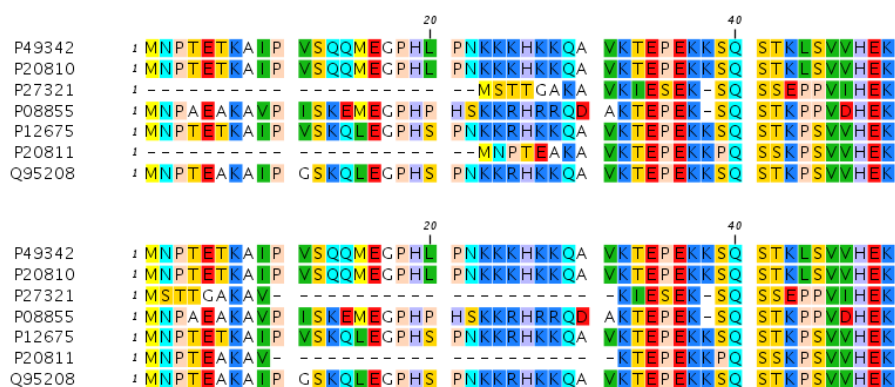


Figure 14.3: The first 50 positions of two different alignments of seven calpastatin sequences. The top alignment is made with cheap end gaps, while the bottom alignment is made with end gaps having the same price as any other gaps. In this case it seems that the latter scoring scheme gives the best result.

### 14.1.2 Fast or accurate alignment algorithm

*CLC Free Workbench* has two algorithms for calculating alignments:

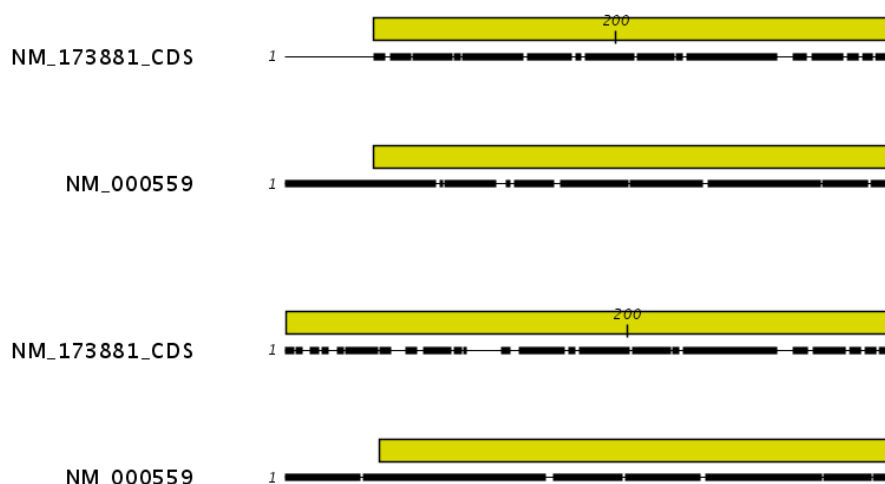


Figure 14.4: The alignment of the coding sequence of bovine myoglobin with the full mRNA of human gamma globin. The top alignment is made with free end gaps, while the bottom alignment is made with end gaps treated as any other. The yellow annotation is the coding sequence in both sequences. It is evident that free end gaps are ideal in this situation as the start codons are aligned correctly in the top alignment. Treating end gaps as any other gaps in the case of aligning distant homologs where one sequence is partial leads to a spreading out of the short sequence as in the bottom alignment.

- **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.

For a comprehensive explanation of the alignment algorithms see section [14.4](#).

## 14.2 View alignments

Since an alignment is a display of several sequences arranged in rows, the basic options for viewing alignments are the same as for viewing sequences. Therefore we refer to section [10.1](#) for an explanation of these basic options.

However, there are a number of alignment-specific view options in the **Alignment info** 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.
- **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.

### 14.2.1 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.

## 14.3 Edit alignments

### 14.3.1 Move residues and gaps

The placement of gaps in the alignment can be changed by modifying the parameters when creating the alignment (see section 14.1). However, gaps and residues can also be moved after the alignment is created:

**select one or more gaps or residues in the alignment | drag the selection to move**

This can be done both for single sequences, but also for multiple sequences by making a selection covering more than one sequence. When you have made the selection, the mouse pointer turns into a horizontal arrow indicating that the selection can be moved (see figure 14.5).

**Notice!** Residues can only be moved when they are next to a gap.

```

AGG GAGTCAT      AGG GAGTCAT
AGG GAGTCAT      AGG GAGTCAT
AGG GAGCAGT      AGG GAGCAGT
- - - - -        - - - - -
AGG GTACAGT      AGG GTACAGT
- - - GAGTAGC    - GA G - - TAGC
- - - GAGTAGC    - GA G - - TAGC
- - - GAGTAGG    - GA G - - TAGG
ATG GTGCACC      ATG GTGCACC
ATG GTGCATC      ATG GTGCATC

```

Figure 14.5: *Moving a part of an alignment. Notice the change of mouse pointer to a horizontal arrow.*

### 14.3.2 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.

### 14.3.3 Delete residues and gaps

Residues or gaps can be deleted for individual sequences or for the whole alignment. For individual sequences:

**select the part of the sequence you want to delete | right-click the selection | Edit selection | Delete the text in the dialog | Replace**

The selection shown in the dialog will be replaced by the text you enter. If you delete the text, the selection will be replaced by an empty text, i.e. deleted.

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.

### 14.3.4 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.

### 14.3.5 Delete 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.

## 14.4 Bioinformatics explained: Multiple alignments

Multiple alignments are at the core of bioinformatical analysis. Often the first step in a chain of bioinformatical analyses is to construct a multiple alignment of a number of homologs DNA or protein sequences. However, despite their frequent use, the development of multiple alignment algorithms remains one of the algorithmically most challenging areas in bioinformatical research.

Constructing a multiple alignment corresponds to developing a hypothesis of how a number of sequences have evolved through the processes of character substitution, insertion and deletion. The input to multiple alignment algorithms is a number of homologous sequences i.e. sequences that share a common ancestor and most often also share molecular function. The generated alignment is a table (see figure 14.6) where each row corresponds to an input sequence and each column corresponds to a position in the alignment. An individual column in this table represents residues that have all diverged from a common ancestral residue. Gaps in the table (commonly represented by a '-') represent positions where residues have been inserted or deleted and thus do not have ancestral counterparts in all sequences.

### 14.4.1 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.

### 14.4.2 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.

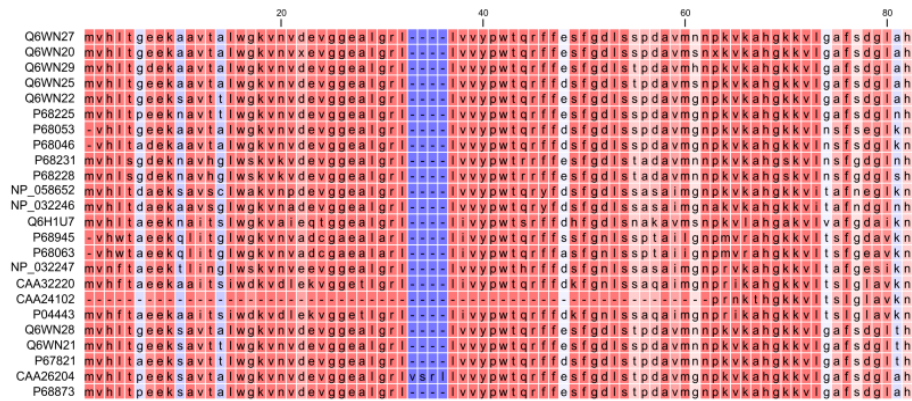


Figure 14.6: The tabular format of a multiple alignment of 24 Hemoglobin protein sequences. Sequence names appear at the beginning of each row and the residue position is indicated by the numbers at the top of the alignment columns. The level of sequence conservation is shown on a color scale with blue residues being the least conserved and red residues being the most conserved.

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 15

## Phylogenetic trees

### Contents

<b>15.1 Inferring phylogenetic trees</b>	<b>129</b>
15.1.1 Phylogenetic tree parameters	129
15.1.2 Tree View Preferences	131
<b>15.2 Bioinformatics explained: phylogenetics</b>	<b>132</b>
15.2.1 The phylogenetic tree	133
15.2.2 Modern usage of phylogenies	133
15.2.3 Reconstructing phylogenies from molecular data	134
15.2.4 Interpreting phylogenies	135

CLC Free Workbench 3.0 offers different ways of inferring phylogenetic trees. The first part of this chapter will briefly explain the different ways of inferring trees in CLC Free Workbench 3.0. The second part, "Bioinformatics explained", will give a more general introduction to the concept of phylogeny and the associated bioinformatics methods.

### 15.1 Inferring phylogenetic trees

For a given set of aligned sequences (see chapter 14) it is possible to infer their evolutionary relationships. In CLC Free Workbench 3.0 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 15.1:

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.

#### 15.1.1 Phylogenetic tree parameters

Figure 15.2 shows the parameters that can be set:

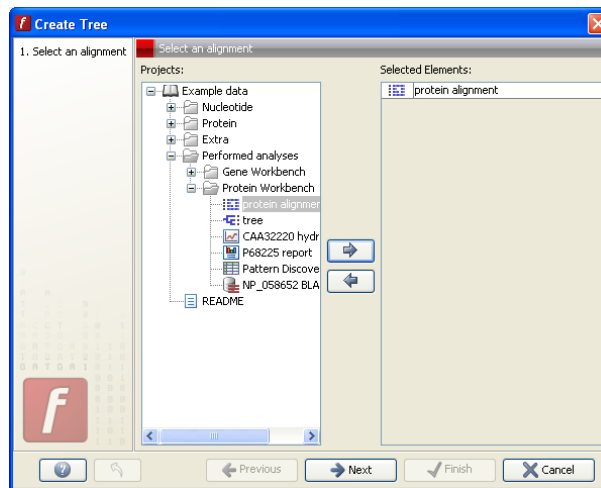


Figure 15.1: Creating a Tree.

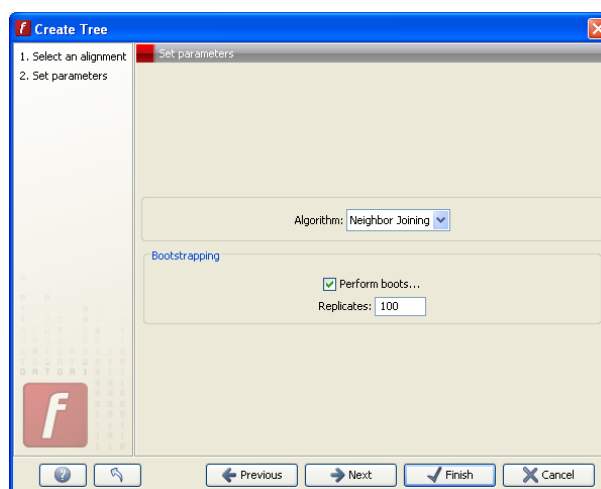


Figure 15.2: Adjusting parameters.

- 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 Free Workbench 3.0* 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 15.3 shows the difference between the two methods.
- To evaluate the reliability of the inferred trees, *CLC Free Workbench 3.0* 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 15.2.

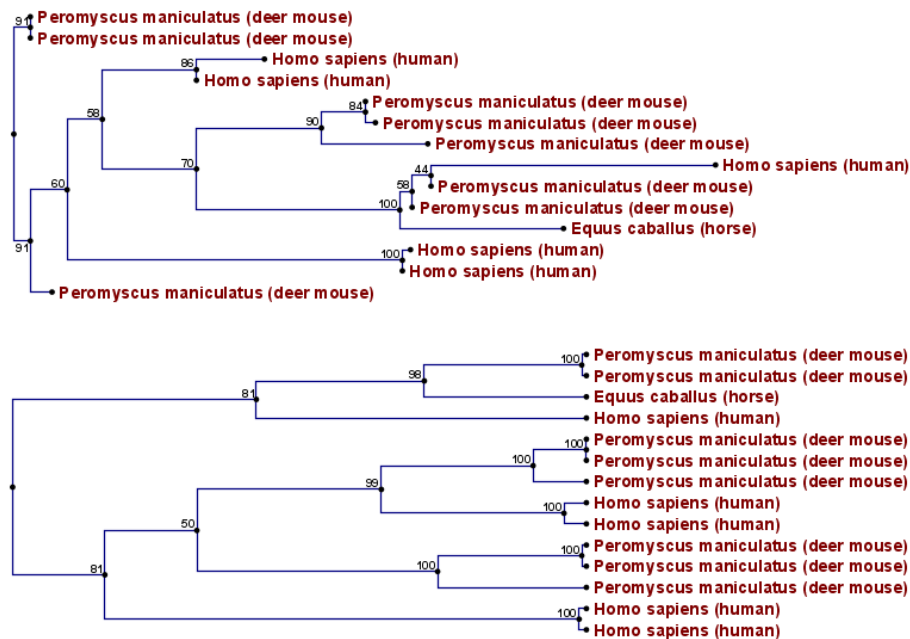


Figure 15.3: *Method choices for phylogenetic inference. The top shows a tree found by neighbor joining, while the bottom shows a tree found by UPGMA. The latter method assumes that the evolution occurs at a constant rate in different lineages.*

### 15.1.2 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.

## 15.2 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.

### 15.2.1 The phylogenetic tree

The evolutionary hypothesis of a phylogeny can be graphically represented by a phylogenetic tree.

Figure 15.4 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.

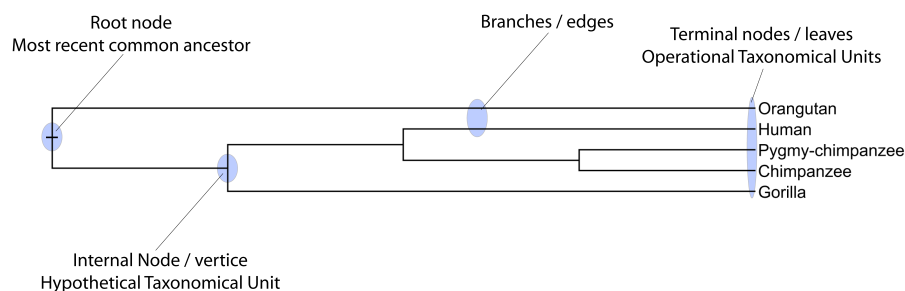


Figure 15.4: A proposed phylogeny of the great apes (Hominidae). Different components of the tree are marked, see text for description.

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 15.4 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.

### 15.2.2 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].

### 15.2.3 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,

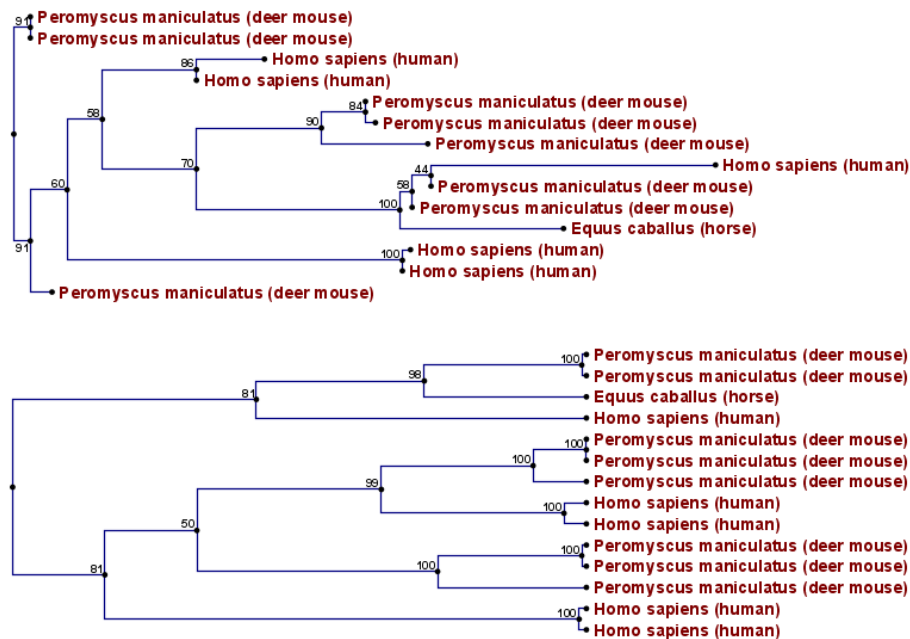


Figure 15.5: Algorithm choices for phylogenetic inference. The top shows a tree found by the neighbor joining algorithm, while the bottom shows a tree found by the UPGMA algorithm. The latter algorithm assumes that the evolution occurs at a constant rate in different lineages.

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].

## 15.2.4 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 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 (■)

<b>Batch processing</b>	Free	Protein	Gene	Combined
Processing of multiple analyses in one single work-step		■	■	■
<b>Database searches</b>	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		■	■	■
<b>General sequence analyses</b>	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>.

<b>Nucleotide analyses</b>	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			■	■
<b>Protein analyses</b>	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		■		■
<b>Sequence alignment</b>	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		■	■	■
<b>Dot plots</b>	Free	Protein	Gene	Combined
Dot plot based analyses		■	■	■
<b>Phylogenetic trees</b>	Free	Protein	Gene	Combined
Neighbor-joining and UPGMA phylogenies	■	■	■	■
<b>Pattern discovery</b>	Free	Protein	Gene	Combined
Search for sequence match		■	■	■
Motif search		■	■	■
Pattern discovery		■	■	■

<b>Primer design</b>	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			■	■

<b>Assembly of sequencing data</b>	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			■	■

<b>Molecular cloning</b>	Free	Protein	Gene	Combined
Advanced molecular cloning			■	■
Graphical display of in silico cloning			■	■
Advanced sequence manipulation			■	■

## Appendix B

# Formats for import and export

### B.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 Free Workbench* can import 'external' files, too. This means that *CLC Free 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 Free Workbench*.

## B.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

- [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.
- [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.
- [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.
- [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.
- [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.
- [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.
- [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.
- [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.
- [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

- AB1, file format, [23](#), [69](#), [141](#)
- ABI, file format, [23](#), [69](#), [141](#)
- About CLC Workbenches, [13](#)
- Accession number, display, [45](#)
- Add
  - annotations, [138](#)
- Advanced preferences, [62](#)
- Algorithm
  - alignment, [120](#)
  - neighbor joining, [134](#)
  - UPGMA, [134](#)
- Align
  - protein sequences, tutorial, [27](#)
  - sequences, [138](#)
- Alignments, [120](#), [138](#)
  - create, [120](#)
  - edit, [124](#)
  - fast algorithm, [122](#)
  - multiple, Bioinformatics explained, [126](#)
  - view, [123](#)
- .aln, file format, [69](#)
- Annotate with SNP's, [138](#)
- Annotation
  - layout, [90](#)
  - map, [95](#)
  - overview, [95](#)
  - types, [91](#)
- Antigenicity, [138](#)
- Append wildcard, search, [84](#)
- Arrange
  - layout of sequence, [24](#)
  - views in View Area, [51](#)
- Assembly, [138](#)
- Automatic parsing, [70](#)
- Back up, [73](#)
- Basic concepts of use, [14](#)
- Batch processing, [80](#), [138](#)
  - log of, [81](#)
- Bioinformatic data
  - export, [71](#)
  - formats, [68](#), [141](#)
- BLAST, [138](#)
- Bootstrap values, [135](#)
- Bug reporting, [13](#)
- CDS, translate to protein, [93](#)
- Cheap end gaps, [122](#)
- .cif, file format, [69](#)
- Circular view of sequence, [100](#), [138](#)
- .clc, file format, [69](#), [72](#)
- CLC Standard Settings, [62](#), [63](#)
- CLC Workbenches, [13](#)
- CLC, file format, [23](#), [69](#), [141](#)
- Cloning, [138](#)
- Close View, [49](#)
- Clustal, file format, [23](#), [69](#), [141](#)
- Coding sequence, translate to protein, [93](#)
- Compare workbenches, [138](#)
- Configure network, [16](#)
- Consensus sequence, [123](#), [138](#)
  - open, [123](#)
- Conservation, [124](#)
  - graphs, [138](#)
- Contact information, [9](#)
- Contig, [138](#)
- Convert old data, [70](#)
- Copy, [76](#)
  - elements in Navigation Area, [44](#)
  - into sequence, [94](#)
  - search results, GenBank, [87](#)
  - sequence, [96](#), [98](#)
  - sequence selection, [110](#)
  - text selection, [96](#)
- .cpf, file format, [62](#)
- Create
  - a project, tutorial, [21](#)
  - alignment, [120](#)
  - enzyme list, [118](#)
  - new folder, [43](#)
  - new project, [43](#)
  - workspace, [57](#)

- Data formats
  - bioinformatic, 141
  - graphics, 142
- Data structure, 42
- Database
  - GenBank, 84
  - local, 42
- Delete
  - element, 46
  - residues and gaps in alignment, 125
  - workspace, 57
- DNA translation, 111
- DNAstrider, file format, 23, 69, 141
- Dot plots, 138
- Double stranded DNA, 89
- Download and open
  - search results, GenBank, 87
- Download and save
  - search results, GenBank, 87
- Download of *CLC Free Workbench*, 9
- Drag and drop, 34
  - Navigation Area, 44
  - search results, GenBank, 86
- Edit
  - alignments, 124, 138
  - annotations, 138
  - enzymes, 91
  - sequence, 94
  - sequences, 138
- Element, 42
  - delete, 46
  - rename, 46
  - .embl, file format, 69
- Embl, file format, 23, 69, 141
- Encapsulated PostScript, export, 75
- End gap cost, 122
- End gap costs
  - cheap end caps, 122
  - free end gaps, 122
- Enzyme list
  - create, 118
  - modify, 119
  - .eps-format, export, 75
- Error reports, 13
- Evolutionary relationship, 129
- Example data, import, 16
- Export
  - bioinformatic data, 71
  - dependent objects, 72
  - folder, 71
  - graphics, 74
  - history, 72
  - list of formats, 141
  - multiple files, 71
  - preferences, 62
  - project, 71
- External files, import and export, 73
- Extract sequences, 100
- FASTA, file format, 23, 69, 141
- Feature request, 13
- Find open reading frames, 111
- Fit Width, 55
- Floating Side Panel, 63
- Format, of the manual, 19
- Free end gaps, 122
  - .fsa, file format, 69
- G/C content, 138
- Gap
  - delete, 125
  - extension cost, 121
  - fraction, 138
  - insert, 125
  - open cost, 121
  - .gbk, file format, 69
- GCG Alignment, file format, 23, 69, 141
- GCG Sequence, file format, 23, 69, 141
- GenBank
  - file format, 23, 69, 141
  - search, 84, 138
  - tutorial, 25
- Gene finding, 111
- General preferences, 61
- General Sequence Analyses, 102
- Getting started, 14
- Graphics
  - data formats, 142
  - export, 74
- Handling of results, 80
- Help, 14
- Hide/show Toolbox, 56
- History, 78
  - export, 72
  - preserve when exporting, 79
  - source elements, 79

- Hydrophobicity, 138
- Import
  - bioinformatic data, 69
  - data from older versions, 70
  - existing data, 22
  - external files, 73
  - FASTA-data, 22
  - list of formats, 141
  - preferences, 62
  - Vector NTI data, 70
- Infer Phylogenetic Tree, 129
- Insert
  - gaps, 125
- Installation, 9
- Join
  - sequences, 105
  - .jpg-format, export, 75
- Lasergene sequence
  - protein file format, 23, 69, 141
  - sequence file format, 23, 69, 141
- Linux
  - installation, 11
  - installation with RPM-package, 12
- List of sequences, 98
- Load enzymes, 91
- Local complexity plot, 138
- Locale setting, 61
- Location
  - of selection on sequence, 55
  - Side Panel, 61
- Log of batch processing, 81
- Logo, sequence, 138
- Mac OS X installation, 11
- Manipulate sequences, 138
- Manual format, 18
- Maximize size of view, 51
- Maximum memory, adjusting, 17
- Memory, adjust maximum amount, 17
- Menu Bar, illustration, 42
- mmCIF, file format, 23, 69, 141
- Mode toolbar, 53
- Modify enzyme list, 119
- Motif search, 138
- Mouse modes, 53
- Move
  - content of a view, 55
  - elements in Navigation Area, 44
  - sequences in alignment, 125
- .msf, file format, 69
- Multiple alignments, 126, 138
- Multiselecting, 44
- Navigation Area, 42
  - illustration, 42
- NCBI, 84
  - search, tutorial, 25
- Neighbor Joining algorithm, 134
- Neighbor-joining, 138
- Nested PCR primers, 138
- Network configuration, 16
- New
  - feature request, 13
  - folder, 22, 43
  - project, 22, 43
  - sequence, 96
- Newick, file format, 23, 69, 141
- .nexus, file format, 69
- Nexus, file format, 23, 69, 141
- Non-standard residues, 92
- Numbers on sequence, 89
- .nwk, file format, 69
- .nxs, file format, 69
- Old data, import, 70
- Open
  - consensus sequence, 123
  - files, 14
- Open reading frame determination, 111
- Open-ended sequence, 113
- Order primers, 138
- ORF, 111
- Origins from, 79
- Page setup, 66
- Parameters
  - search, 84
- Parsing, automatic, 70
- Paste/copy, 76
- Pattern discovery, 138
- PCR primers, 138
- .pdb, file format, 69
- .seq, file format, 69
- PDB, file format, 23, 69, 141
- .pdf-format, export, 75
- Personal information, 13

- Pfam domain search, 138
- Phred, file format, 23, 69, 141
  - .phy, file format, 69
- Phylip, file format, 23, 69, 141
- Phylogenetic tree, 129, 138
  - tutorial, 28
- Phylogenetics, Bioinformatics explained, 132
  - .pir, file format, 69
- PIR (NBRF), file format, 23, 69, 141
  - .png-format, export, 75
- Polarity colors, 92
- PostScript, export, 75
- Preferences, 60
  - advanced, 62
  - export, 62
  - General, 61
  - import, 62
  - style sheet, 62
  - toolbar, 61
  - View, 61
  - view, 52
- Primer
  - design, 138
- Print, 65
  - preview, 66
  - visible area, 65
  - whole view, 65
- .pro, file format, 69
- Problems when starting up, 14
- Processes, 56
- Project, create new, 22
- Protein
  - charge, 138
  - report, 138
- Proteolytic cleavage, 138
- Proxy server, 16
  - .ps-format, export, 75
- PubMed references, search, 138
- Quick start, 15
- Rasmol colors, 92
- Reading frame, 111
- Realign alignment, 138
- Rebase, restriction enzyme database, 118
- Recycle Bin, 46
- Redo/Undo, 50
- Reference sequence, 138
- Region
  - types, 94
- Remove
  - annotations, 94
  - terminated processes, 56
- Rename element, 46
- Replace file, 74
- Report program errors, 13
- Report, protein, 138
- Request new feature, 13
- Residue coloring, 91
- Restore
  - deleted elements, 46
  - size of view, 52
- Restriction enzymes, 115
- Restriction sites, 115, 138
  - enzyme database Rebase, 118
  - on sequence, 91
  - parameters, 115
  - tutorial, 30
- Results handling, 80
- Reverse complement, 110, 138
- Reverse translation, 138
- RNA translation, 111
- Safe mode, 14
- Save
  - changes in a view, 50
  - search, 26
  - sequence, 27
  - style sheet, 62
  - view preferences, 62
  - workspace, 57
- SCF2, file format, 23, 69, 141
- SCF3, file format, 23, 69, 141
- Search
  - GenBank, 84
  - handle results from GenBank, 86
  - hits, number of, 61
  - in a sequence, 92
  - in annotations, 92
  - options, GenBank, 84
  - parameters, 84
- Secondary structure prediction, 138
- Select
  - exact positions, 92
  - in sequence, 93
  - parts of a sequence, 93
  - workspace, 57
- Selection mode in the toolbar, 55

- Selection, location on sequence, 55
- Sequence
  - alignment, 120
  - analysis, 102
  - display different information, 45
  - extract from sequence list, 100
  - information, 94
  - information, tutorial, 31
  - join, 105
  - layout, 89
  - lists, 98
  - logo, 138
  - new, 96
  - region types, 94
  - search, 92
  - select, 93
  - shuffle, 105
  - statistics, 102
  - view, 88
  - view as text, 96
  - view circular, 100
  - view format, 45
- Sequencing data, 138
- Sequencing primers, 138
- Shortcuts, 58
- Show/hide Toolbox, 56
- Shuffle sequence, 105, 138
- Side Panel, location of, 61
- Signal peptide, 138
- SNP
  - annotation, 138
- Sort
  - sequences, 99
  - sequences alphabetically, 126
- Source element, 79
- Species, display sequence species, 45
- Staden, file format, 23, 69, 141
- Standard layout, trees, 132
- Standard Settings, CLC, 63
- Start Codon, 113
- Start-up problems, 14
- Statistics
  - about sequence, 138
  - sequence, 102
- Status Bar, 56, 57
  - illustration, 42
- .str, file format, 69
- Style sheet, preferences, 62
- Support mail, 9
  - .svg-format, export, 75
- Swiss-Prot, file format, 23, 69, 141
- Swiss-Prot/TrEMBL, 138
  - .swp, file format, 69
- System requirements, 12
- Tabs, use of, 48
- TaqMan primers, 138
- Terminated processes, 56
- Text format, 93
  - user manual, 19
  - view sequence, 96
- Text, file format, 23, 69, 141
  - .tif-format, export, 75
- Tips and tricks, tutorial, 33
- Toolbar
  - illustration, 42
  - preferences, 61
- Toolbox, 56
  - illustration, 42
  - show/hide, 56
- Topology layout, trees, 132
- Trace data, 138
- Translate
  - annotation to protein, 93
  - DNA to RNA, 108
  - nucleotide sequence, 111
  - RNA to DNA, 109
  - to DNA, 138
  - to protein, 111, 138
- Translation
  - tables, 111
- Transmembrane helix prediction, 138
- Trim, 138
  - .txt, file format, 69
- Undo limit, 61
- Undo/Redo, 50
- UniProt
  - search, 138
- UPGMA algorithm, 134, 138
- Urls, Navigation Area, 73
- User defined view settings, 62
- User interface, 42
- Vector graphics, export, 75
- VectorNTI
  - file format, 23, 69, 141

- import data from, 70
- View, 48
  - alignment, 123
  - preferences, 52
  - save changes, 50
  - sequence, 88
  - sequence as text, 96
- View Area, 48
  - illustration, 42
- View preferences, 61
  - show automatically, 61
  - style sheet, 62
- View settings
  - user defined, 62
- Wildcard, append to search, 84
- Windows installation, 10
- Workspace, 57
  - create, 57
  - delete, 57
  - save, 57
  - select, 57
- Wrap sequences, 89
- Zoom, 53
  - tutorial, 24
- Zoom In, 53
- Zoom Out, 55
- Zoom to 100% , 55