The easiest and most efficient way to conduct much bioinformatics is to use a suite of software dedicated to relevant applications. In this way, the input and output of various analyses can be saved, and sometimes logged.

This can then be used in conjunction with web servers, many of which offer applications complimentary to those offered within suite.
Contents

Data Analysis Packages............................................. 1

EMBOSS - A data analysis package........... 1

Jemboss - Graphical User Interface........... 1

Jemboss Menu .......................................................... 2

Program Form.......................................................... 3

Sequence Input .......................................................... 4

File Management.......................................................... 4

Running an Application................................................... 5

Batch Mode ............................................................... 7

Practical ........................................................................ 7

References ..................................................................... 9
Data Analysis Packages

EMBOSS - A data analysis package

Although many applications are on the World Wide Web, it is a lot easier if there is a package to hand that contains many of the staple applications used to verify information as a precursor to laboratory experimentation. The package used and developed at the HGMP-RC is one called EMBOSS\(^1\) - European Molecular Biology Open Software Suite. All programs are open source, meaning anyone can download the information and install it themselves. The application suite is offered through the HGMP-RC site and a support desk is available for any problems. This service is free to academics, but entails a fee for commercial users.

A list of current and proposed applications is maintained at http://www.uk.embnet.org/Software/EMBOSS/Apps/. We will be using the graphical interface to EMBOSS for the duration of this course, but should you wish to use it from the UNIX command line, you will find a copy of the EMBnet "Quick Guide" to EMBOSS in the Appendix.

The exercises in this chapter and many of those later in the course are based on a freely available sequence analysis package called EMBOSS. This is not the only sequence package available to you - you can also use GCG, Spin and various others. There are many commercial companies who also offer packages, specifically Informax (http://informaxinc.com), Paracel (http://www.pracel.com), Biotools (http://www.biotools.com/) and Molecular Mining (http://www.molecularmining.com/) to name but a few.

Jemboss - Graphical User Interface

The interface we will be using is called Jemboss\(^2\), and it has been written at the HGMP-RC by T. Carver in collaboration with the EMBoss team. It has been designed to run on both PC and UNIX systems\(^2\). Before you can use it, however, you must download a plug-in to allow this Java program to be downloaded and launched from your machine.

---

1 So called because it has been written in the programming language Java. This language allows programs to function on any java enabled computer platform.

2 Unfortunately, the old Mac operating system does not support Java and unless you have MacOSX (purchase since 2001), you must use Jemboss via VNC.

Human Genome Mapping Project Resource Centre http://www.hgmp.mrc.ac.uk
Go to the HGMP homepage at http://www.hgmp.mrc.ac.uk and follow the link to the WWW menu. Select the link to "EMBOSS" under the "Common Options" header and then the link to "Jemboss - Graphical Interface to EMBOSS". You now need to download the Java Web Start plug-in. Decide which type of computer software you are using and follow the link. We are using PCs here, so follow the "Windows" link to the Java Web Start page\(^3\). Go to Step 1 and click on the "Continue" button for the "Full package (US English only)". The file should be ~5.3 Mb in size. The link takes you to the Terms and Conditions of use. Read these and then "Accept" at the bottom of the page. Now select the "FTP Download" button from java.sun.com. You will be asked where you would like to save your file. This will depend on the room we are in - someone will tell you beforehand where to save it (usually the desktop). Remember where you saved it!

Now select the "exe" file you have just saved using a file manager and double click on the name of the program. This will proceed to install Java Web Start. Accept the license agreement and select the same place to save the data in\(^4\). When you have entered the correct path, click next and the program should install. You do not need to restart your computer, but you must close down your web browser and re-open it.

Go back to the original EMBOSS page in your browser, and select the link "Launch Jemboss". You may have to wait a few moments for the program to load when you first do this.

After a pause of a few seconds, during which time you will be asked to type in your username and password\(^5\), click on "OK" and the full Jemboss interface should appear on your screen.

**Jemboss Menu**

On the left hand side you will see analysis groups. Click on these to display other menus and programs belonging to that group. You will see the name of the program, along with its one line description. This should aid you in making a decision about the correct program to use. Alternatively, programs are listed in alphabetical order underneath these groups. If you know which program you wish

---

\(^3\) If you are using a UNIX system, you may need to ask your systems administrator where best to save the files you are going to download.

\(^4\) If you are doing this on your desktop computer in your laboratory, the best place to save all these files is in the folder C:\Program Files\.

\(^5\) For the purposes of this course, you will be assigned a username and password. Otherwise you should use the identification you received when you registered with the HGMP.
to use, you may scroll down the menu and select it by clicking once on the program name. Alternatively, you may start typing in the name of the program in the "GoTo" box, and when it appears in the menu, just click on it.

Program Form

The central panel is where the program form will appear. In red at the top of the form the name of the program is displayed, followed by its one-line description. Below this, there are the sequence entry options.\(^6\)

![Program Form Image](image)

Figure 1

Very often there is an "Advanced Options" button. This is not the case every time, and sometimes this button could obscure important parameters. It is important to check for each program.

---

\(^6\) Some programs will not require a sequence, and in these cases there is, obviously, no sequence entry field.

Human Genome Mapping Project Resource Centre  http://www.hgmp.mrc.ac.uk
**Sequence Input**

Sequences may be added to the program in a number of ways. They may be entered straight into the filename box using the conventional EMBOSS Universal Sequence Address (USA)\(^7\): files may be transferred from local or remote accounts by "drop & dragging" the relevant file into the filename box; sequences may be pasted into the filename box by selecting the "paste" option. For programs requiring multiple sequence options, files may be added separately. Sequence properties may also be entered using the "Input Sequence Options" menu. This latter option allows you to choose the database from which you would like to retrieve your sequence, but you must ensure that you TYPE the sequence identifier or accession number INTO THE SEQUENCE FILENAME BOX\(^8\).

Different applications will have various options available on the program form. Select, or fill in these options as required.

If an "Output Sequence Name" is entered, your output will automatically be saved onto the server, using that name. If you do not enter one, however, the file is automatically saved onto the HGMP server using a default output name, and you may retrieve it at a later date.

**File Management**

Both the remote and local file managers can be accessed using the "File" menu of the Jemboss interface and selecting "Show Local and Remote Files". An elongated window (figure 2) will pop up on your screen and you may split the screen horizontally or vertically. Files may be dragged between local and remote computers by picking up the file to move, and placing it over the name of a file in the relevant manager. You may also drag & drop files from this window to the sequence input field. The drop down menu at the top of this window is to allow you to toggle between files that are stored in your home directory, and your scratch directory.

Jemboss has an inbuilt facility to allow you to transfer files between your local disk and your HGMP account. Just drag and drop the required files from one list to another to move them between accounts. Each file manager also has an options menu associated with it. By clicking the right hand mouse button once a

---

\(^7\) The sequence address communicates to the program which sequence you would like to retrieve from which database. It is written in the form **database:sequence** where sequence represents the sequence identifier or accession number.

\(^8\) If you do not type in a sequence identifier, the program will try to download the whole of your specified database, when using programs like "seqret", for example.
menu will pop up, offering the option to refresh the list of files. This is particularly useful if you have just created a new file or folder. Options to delete a file or folder are also on the menu, together with renaming a file, or creating a new folder.

![Local and Remote file manager](image)

Figure 2

Running an Application

Hit the “Go” button to run an application. If you do not understand the significance of a parameter option, use the “mouse over” facility. If the parameter has a specific help text attached to it, it will appear on the screen.

Alternatively, click on the button, which will take you to the program documentation.

Your results will be presented to you in a new tabbed window on the screen (Figure 3). The first tab should contain the results of the program that you ran. The name on the tab is the default output name, or a filename specified by you. You may save your results to your local computer. A second tab (cmd) will give you the command line. If you are saving graphical output, you may save it as a “.png” file (the default). This format may be inserted as a picture into Microsoft Word documents. Other tabs contain any sequences you may have pasted into the “input sequence” window.

---

9 As EMBOSS runs on a UNIX operating system, the optimal way of executing a program is to write out all the commands you wish the program to do on one line – the command line. The interface will actually do this for you, but if you work with the UNIX system, or would like to start, it may be useful for you to know the exact command line used.
You can then view your file using Jemboss, by double-clicking on the filename\textsuperscript{10}. Alternatively, you can save the file to your local machine and then view the output using a WWW browser. The file manager on Jemboss should automatically update to show your saved files. When you have finished with the window, it can be closed.

If you have entered your input sequence and filled in any extra sequence parameters you wish to use in the "Input Sequence Options" box, you may then click the "LOAD SEQUENCE ATTRIBUTES" button\textsuperscript{11}. This will then display all remaining sequence parameters for the program that are pertinent to your sequence only. Parameters that are not appropriate for your sequence will be greyed out. If you would prefer redundant (for a particular sequence) parameters to be absent on the screen, de-select "Shade unused parameters" from the "Preferences" menu of the Jemboss interface.

\textsuperscript{10} Files may also be edited in this way.
\textsuperscript{11} It is important to input all sequence options first, otherwise the sequence attributes will be overwritten.
Saved results can be accessed by using the "Saved Results" option on the pull-down "File" menu in the Jemboss window. This will display all previous applications you have run. Details are displayed in the right hand pane in command line format to offer a reminder of exact parameters used for a particular analysis. Results can be displayed by selecting the appropriate entry and clicking on the "Display" button. Information and results files can be discarded by clicking first on the program and then on the "Delete" button.

**Batch Mode**

Programs may be run interactively, or in batch mode. This allows the analysis to be carried out in the background, whilst you get on with doing something else. The job manager is located at the bottom of the Jemboss window, and will keep a tally of all batch programs running, and all those completed. For each session, the job manager will also record all analyses run, and you may recall and display them as you would using the program manager. The job manager will automatically refresh every 15 seconds, to let you know whether your analysis is still running, or whether it has finished. This time frame may be changed, by choosing the "Advanced Options" of the Jemboss "File" menu and altering the options in the pull down job manager menu.

**Practical**

Here are a couple of simple programs to familiarise you with Jemboss.

---

Type in the "GoTo" field on the left hand side of the interface as many letters as you need to, to make "wosname" visible on the scroll menu. Click on the program "wosname" and wait for the program form to load. Type database into "keyword" field, and ensure "Search EMBoss programs" and "Search EMBASSY programs" are selected (they should be ticked). Leave the output as the default Launch the program by clicking on the "Go" button.

A tabbed window should appear giving details of all the programs in EMBoss which have the word "database" in their one line description. Although Jemboss has grouped programs according to functionality, if you are new to the package, it isn’t always obvious which program to use. Wosname acts like a dictionary and can be used to look up relevant programs. Towards the bottom of the list is a program called "showdb". According to its one line description, this program displays information on the currently available databases. We will use this program now.
Close the "wossname" results window and select "Showdb" either from the grouped menus (in "Display") or from the scroll menu. Do not fill anything in the database field, and leave the "protein databases" and "nucleic acid databases" options ticked. Select the "Display Column Headings" and click on "Go".

The left hand column of this list details all the databases that are available for use with EMBOSS at the HGMP\textsuperscript{12}. The names contained in this list represent their identifier to EMBOSS applications.

\begin{itemize}
  \item EMBL identifiers
  \item Swissprot identifiers
\end{itemize}

Letters in the "Type" column indicate whether the database contains Protein or Nucleotide sequences and an "OK" in the subsequent three columns "ID", "Qry" and "All\textsuperscript{13}" represent whether you may query the database for a single sequence, a set of sequences, or all sequences in that database respectively. The "Comment" column provides a description of the database. If you do not need all the information provided in the output, you may select the "Display specified columns" option and choose the information you require.

You do not need this window anymore, so close it down.

\textsuperscript{12} Like SRS, this may change depending on where you are using the service. If you university is providing EMBOSS as a service, the database selection may be more limited.

\textsuperscript{13} For more than one sequence, you can use a wildcard. This is represented by an asterisk inserted after any identifying characters. So searching for all pax sequences in Swissprot, you could type \texttt{sw:pax*} into the sequence field of an application. If you wanted to search all the files in this database, you would enter \texttt{swissprot:*} into the field.
References

EMBOSS: The European Molecular Biology Open Software Suite

2 Carver T.J., Mullan L.J., (2002) Comparative and Functional Genomics 3 (1) 75-78,
A new graphical user interface to EMBOSS