A Quick Guide To EMBOSS

http://www.emboss.org

This is a Quick reference Guide for EMBOSS version 2.8.0

Rice, P. Longden, I. and Bleasby, A. (2000) "EMBOSS: The European Molecular Biology Open Software Suite" *Trends in Genetics* **16**(6):276-277.

Introduction

EMBOSS (European Molecular Biology Open Software Suite) is a freely available suite of programs and libraries for sequence analysis. It incorporates many tools originating from the EGCG package created in 1988. All EMBOSS programs are designed to run on a UNIX command-line or behind graphical interfaces (*e.g.,* Jemboss, wEMBOSS).

Obtaining EMBOSS

To install EMBOSS: download the current version from `ftp://ftp.uk.embnet.org/pub/EMBOSS/EMBOSS-2.8.0.tar.gz`, then follow the instructions at:
`http://www.rfcgr.mrc.ac.uk/Software/EMBOSS/download.html`

Graphical User Interfaces

There are a number of graphical interfaces to EMBOSS:
`http://www.rfcgr.mrc.ac.uk/Software/EMBOSS/interfaces.html`

Jemboss is a java interface and is distributed with EMBOSS. If you are installing with the Jemboss interface you should use the installation script in the `EMBOSS-x.x.x/jemboss/utils` directory. Instructions for Jemboss installation are given at:
`http://www.rfcgr.mrc.ac.uk/Software/EMBOSS/Jemboss`

Support and Mailing lists

The mailing list `emboss@embnet.org` is used for discussions of user problems. To subscribe to this list, send a mail to `majordomo@embnet.org` with the message text: *subscribe emboss*. The mailing list archive is:
`http://www.rfcgr.mrc.ac.uk/Emboss/HYPERMAIL/emboss`

Please send bug reports to `emboss-bug@embnet.org`

Help on a program

A program can be found using a keyword search of the description of all the programs by running the EMBOSS application `wossname`.

`wossname` *keyword*	displays list of all programs with keyword in description
`wossname` *-alphabetic -auto*	displays a list of all programs
programname -help	gives the available parameters for the *programname*
`tfm` *programname*	displays the documentation of *programname*

Documentation is also given online at:
`http://www.rfcgr.mrc.ac.uk/Software/EMBOSS/Apps`

Sequence formats

Sequences are stored in databases or in files as simple text. EMBOSS does not support sequences in word-processor files! The default sequence file format is **fasta**. This format has an initial title line consisting of a ">" followed by the sequence description on the first line. The second and subsequent lines contain the sequence, *e.g.:*

```
>fau Human FAU gene fragment
GACCGGCCAGGAAACGGCATGTAGCCTCACTGGAGGGCATTGCCCCGGA
AGATCAAGT
```

EMBOSS currently supports 42 formats, including: **Clustal, EMBL, GCG, Genbank, PIR, MSF, Phylip, SwissProt, Text (raw).**
The default output can be altered for all programs by an environment setting:
`setenv EMBOSS_OUTFORMAT` *format*

Alignment Formats

Several formats have been written or adopted for EMBOSS output.

Multiple Alignment

`simple`	Displays names, positions and sequences, markup line underneath [default]
`fasta`	Standard fasta display. Gaps displayed as "-"
`msf`	Standard MSF format. Gaps displayed as "." for intrinsic and for terminal ones
`srs`	Similar to simple. No markup line
`trace`	Verbose form for de-bugging

Pairwise Alignment

`pair`	Simple format for pairwise output [default]
`markx`	Standard output from FASTA program suite
`srspair`	Similar to pair format
`score`	Score output. No sequence display

Any program derived from Bill Pearson FASTA suite of programs has a markx default format.

`-aformat`	Alters output format
`-awidth`	Displays alignment width
`-ausashow`	Displays the full USA (see below) in the alignment

Feature Formats

`gff`	General Feature format defined by the Sanger Institute [default]
`embl`	Feature table used by EMBL database (em)
`swissprot`	Feature table used by SwissProt database (swiss) (sw)
`-ufo`	UFO (uniform features object) features
`-fformat`	Opens features format

These flags can be applied to the output by using "o" as a prefix, e.g. -oufo

`-fbegin`	Specifies first position
`-fend`	Specifies final position
`-freverse`	Reverses features (DNA only)

Graphic Formats

`-graph`	Static graphics using PLP plot. Output as X11 [default], PNG, ps, tektronics amongst others

Sequence Databases

Your local EMBOSS installation may have many sequence databases set up. The program `showdb` will indicate the available databases.

Uniform Sequence Address (USA)

A USA is an unambiguous means of specifying sequences in EMBOSS. It has the following syntax:

format::database:entry

Only raw (text) or IntelliGenetics format need to be specified. EMBOSS identifies the rest automatically.

You may also use:

`filename`	all sequences in a file
`filename:entry`	an entry in a file
`@listfilename`	a list file (see below)
`asis::ACGACTGACGG`	a specific short sequence

The entry can include '*' characters for wildcard matches of several entries and sequence may be specified by adding [`start:end:rev`] positions to the USA. The `rev` keyword will reverse complement a DNA sequence. Command lines using these characters must be encased in double quotes :

e.g. `seqret "embl:hs*"`

A part of the sequence can be specified by adding the range:

e.g. `seqret "embl:hsfau[1:57]"`

The last 100 bases of a sequence can be specified by a negative start:

e.g. `seqret "embl:hsfau[-100:]"`

List Files

A list file contains a list of USAs (one per line). The list file input is `@listfile`. A list file may be read in wherever a program can read multiple sequences. Blank lines and USAs starting with a '#' character are ignored. There is no limit on different sequence formats within one list file.

Format Conversion

The format of an output sequence file can be specified. `seqret` can read in sequences in one format and write them in the other format, for example to convert a sequence to GCG format:

```
seqret in.seq  gcg::out.seq
```

The command line and parameters

EMBOSS programs are designed to be run from the command-line, as well as within scripts. To customise their behaviour, each has a distinct set of parameters, also known as options or flags.

There are 3 classes of parameters: *standard, additional, advanced.* Information on allowable flags for each program is given in the help files.

If values for *standard* (mandatory) parameters are not specified, the programs will prompt for them.

If *additional* (optional) parameters are missed out, default values will be used unless you put `options` (or `opt`) on the command line.

EMBOSS programs never prompt for *advanced* parameters; these must be explicitly specified. They are defined in the program documentation.

General qualifiers

These can be used with any program:

`-auto`	Turns off prompts and descriptions. Used when in running programs scripts
`-stdout`	Writes to standard output (screen) by default
`-filter`	Reads from standard input (keyboard), writes to standard output (screen) by default
`-options`	Prompts for all required and additional values
`-debug`	Writes debug output to the file *programname.dbg*
`-help`	Reports command line options. Or help verbose for more information on associated and general qualifiers
`-warning`	Reports warnings
`-error`	Reports errors
`-fatal`	Reports fatal errors
`-die`	Reports deaths

Each of these can be prefixed with "no" to negate the action.

e.g. `-nowarning`

`-sbegin`	States the first position of the sequence
`-send`	States the final position of the sequence

Some major programs

EMBOSS currently offers approximately 200 applications Use `wossname` to see them all together with below a selection of interesting tools:

TOOLS (examples)

`seqret`	Reads and writes (returns) sequences
`est2genome`	Aligns EST and genomic DNA sequences
`needle`	Needleman-Wunsch global alignment
`water`	Smith-Waterman local alignment
`dotmatcher`	Displays a thresholded dotplot of two sequences
`remap`	Displays a sequence with restriction cut sites, translation etc
`prettyplot`	Displays aligned sequences, with colouring and boxing
`extractseq`	Extracts regions from a sequence
`revseq`	Reverses and complements a sequence
`plotorf`	Plots potential open reading frames
and many other	

UTILS MISC

`embossdata`	Finds or fetches the data files read in by the EMBOSS programs
`embossversion`	Writes the current EMBOSS version number

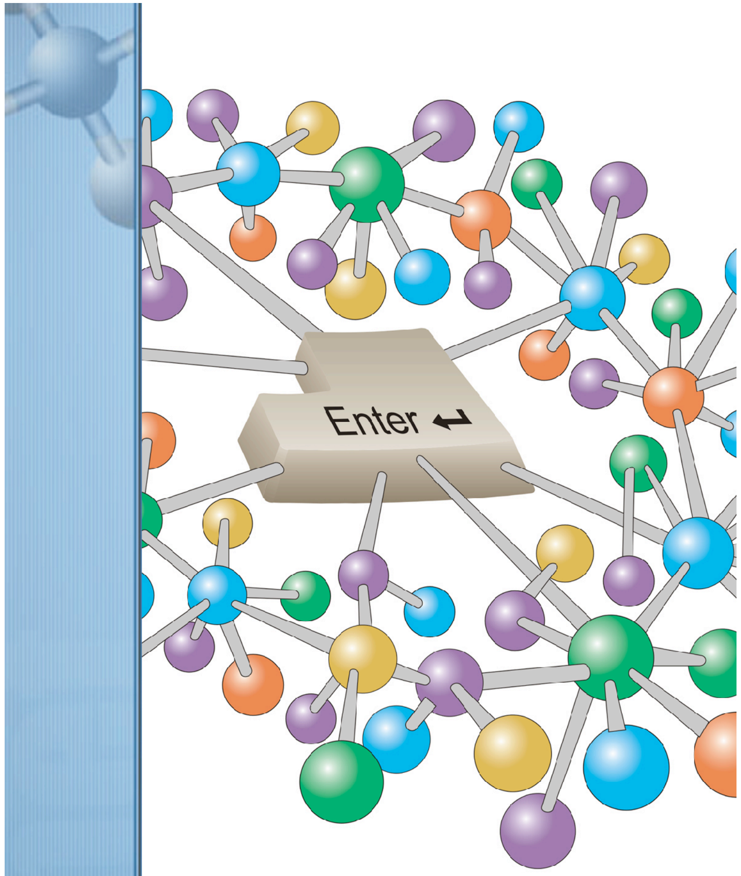
This document was written and designed by Lisa Mullan from the UK EMBnet node and being distributed by P&PR Publications Committee of EMBnet.

EMBnet - European Molecular Biology Network - is a bioinformatics support network of bioinformatics support centers situated primarily in Europe. Most countries have a national node which can provide training courses and other forms of help for users of bioinformatics software.

You can find information about your national node from the EMBnet site:

http://www.embnet.org/

A Quick Guide To EMBOSS
First edition © 2004



EMBnet

A Quick Guide
EMBOSS