EMBOSS

A Quick Guide

European Molecular Biology Open Software Suite

History

Since 1988, the sequence analysis package EGCG has provided extensions to the market leading commercial sequence analysis package GCG. EGCG development was a collaboration of groups within EMBnet and elsewhere.
That project has reached the limits of what we can achieve using the GCG package. Specifically, it is no longer possible to distribute academic software source code which uses the GCG libraries and has become difficult even to distribute binaries.
As a result, the former EGCG developers have been designing a totally new generation of academic sequence analysis software. This has resulted in the present EMBOSS project.
EMBOSS is a new suite of freely available programs and libraries for sequence analysis. It incorporates and integrates a range of currently available public packages and tools into a general, publicly available, suite specially developed for the needs of the Sanger Centre and the EMBnet user community.

Licensing

The EMBOSS core application suite is licensed under the General Public License (GPL) allowing free copying, modification and distribution of the package.
The EMBOSS Libraries are licensed under the the Library General Public License.
Associated packages may be licensed under different terms, all of which permit free redistribution of the software.

Obtaining EMBOSS

EMBOSS and the associated packages can be obtained via FTP from the Sanger Centre, UK at ftp.sanger.ac.uk/pub/EMBOSS

EMBOSS home page

http://www.sanger.ac.uk/Software/EMBOSS

Running EMBOSS

All EMBOSS programs are designed to be run from the comamnd line. Each program has a specific description file (ACD file) that describes the input and output parameters. All the parameters can be specified on the command line, allowing modular integration into graphical interfaces.
To run an EMBOSS program, just type its name. Your system administrator should ensure that the programs are available in your `$PATH`.

The Uniform Sequence Address (USA)

The USA is a method of specifying the location of a sequence and its format. The general form is:
Format:::database:sequencename
eg. **`embl:em:scact`**
EMBOSS is normally very good at identifying sequence *formats* automatically but occasionally needs a hint. *Database* will be one of the databases already set up at your site. The command
% **`showdb`**
lists the databases available on your system.
The *sequencename* can be either its name, accession number, the filename in which the sequence is found, or the sequence itself if **`asis::`** format is specified. If you are taking one sequence from a multiple sequence file, put the sequence number in braces after the filename, eg:
`allmyseqs.fasta{32}`

EMBOSS programs

You can obtain a list of EMBOSS programs with the command **`wossname`**. Useful qualifiers for **`wossname`** are :

`-alphabet`	List all programs in alphabetical order
`-auto`	List all programs without asking for a keyword.

% **`wossname –alphabet –auto`**
will list all the available emboss programs with a short description of the function of each program

EMBOSS will by default only prompt you for the minimal input it needs to run the program. The default behaviour can be changed using command line qualifiers.

Important qualifiers

The behaviour of EMBOSS programs can be modified by using a large number of qualifiers. This is a list of the more useful ones.

`-help`	Prints a summary of the options the program can take. With **`-verbose`** it gives a more detailed list.
`-options`	Prompt the user for the optional parameters
`-auto`	Accept all the default settings and run without prompting the user.
`-sask`	Ask for the start, end and reverse of the sequence input
`-stdout`	Print output to `stdout` (the screen) instead of to a file.
`-filter`	Take input from `stdin` (keyboard) and output to `stdout`

What `-help` tells you

The **`–help`** option lists the inputs to the program along with the input type (sequence, integer etc). There are additional qualifiers associated with many types. **`–verbose`** will list all the additional qualifiers related to the input types for the program.

The qualifiers are listed in three sections:

Mandatory Qualifiers	These are the minimum inputs the program needs to run. Some of these have default values which can be selected using **`-auto`**
Optional Qualifiers	These are qualifiers for which you will be prompted if you use the **`–option`** qualifier. All these qualifiers have default values.
Advanced Qualifiers	You will never be prompted for these. If you wish to use them you must specify them on the command line.

EMBOSS parameter types

Type	Allowed values
`bool`	yes: **`-param`** no: **`-noparam`**
`integer`	Whole numbers **`–param=5`**
`float`	decimal numbers **`–param=23.9`**
`range`	sequence ranges. eg. **`-param=1-12,35-99`**
`regexp`	a regular expression pattern
`string`	ordinary text. **`-param='text with *'`**
`infile`	path of a file
`matrix`	integer scoring matrix for alignments
`matrixf`	floating point scoring matrix
`codon`	codon usage table
`sequence` **`seqset`** **`seqall`**	Uniform sequence address (USA) for the sequence or set of sequences.
`features`	Feature table
`list`	list of options
`selection`	selection list of options
`outfile`	path to a file for nonsequence output
`seqout`	output sequence USA
`seqoutset`	multiple sequence file for output
`seqoutall`	multiple or single sequence output files
`featout`	output feature table
`graph`	output device for graphics images
`xygraph`	output device for XY graphs

See the descriptions below for many of these.

Associated qualifiers: sequence, seqset, seqall

-sbegin	integer	first base used [start]
-send	integer	last base used [end]
-sreverse	bool	reverse sequence [N]
-sask	bool	prompt for begin/end/reverse [N]
-snucleotide	bool	Sequence is nucleotide [N]
-sprotein	bool	Sequence is protein [N]
-slower	bool	Make sequence lowercase[N]
-supper	bool	Make sequence uppercase[N]
-sformat	string	input sequence format
-sopenfile	string	input filename
-sdbname	string	database name
-sentry	string	entry name/accession number
-ufo	string	Feature table (UFO)
-fformat	string	features format

Associated qualifiers: seqout, seqoutset, seqoutall

-osformat	string	output sequence format
-osextension	string	filename extension
-osname	string	base filename
-osdbname	string	database name to add
-ossingle	bool	seperate file for each entry[N]
-oufo	string	features UFO
-offormat	string	features format
-ofname	string	features filename

Associated qualifiers: features

-fformat	string	features format
-fopenfile	string	features filename
-fask	bool	prompt for **fbegin, fend,** and **freverse**
-fbegin	integer	features starting position
-fend	integer	features end position
-freverse	boon	features on the reverse strand [N]

Associated qualiifers: featout

-offormat	string	feature format
-ofopenfile	string	output filename
-ofextension	string	filename extension
-ofname	string	filename
-ofsingle	bool	write one feature per file

Associated qualifiers: graph, xygraph

-gprompt	bool	graph prompting
-gtitle	string	graph title
-gsubtitle	string	graph subtitle
-gxtitle	string	x axis title
-gytitle	string	y axis title
-grtitle	string	right axis (y2) title
-gpages	integer	number of pages
-goutfile	string	output filename

EMBOSS and Graphics

EMBOSS can support a number of different graphics output types depending on the features available on your system. It will prompt for a graphics device:

Graphics device [x11]:

Typing rubbish here then pressing return will give a lengthy list of devices, many of which are equivalent.

The main graphics options are:

[X]	x11	Output to an X-window
	postscript	Output to a postscript file (good for printing on a laser printer)
	cps	Output to a colour postscript file
	text	Output to a text file
	data	Output XY data points to a file. (good for importing into a graphing package)
[P]	png	Output to a PNG image file (good for web pages)
[X]	Tek	Output to tektronics terminal
[X]	xterm	Output to an Xterm window

[X]- requires X-windows [P] – requires PNG support

The default filename is *prog.format* eg. **octanol.ps**

Some useful programs

General

wossname	lists all EMBOSS programs
showdb	Shows the available databases

Sequence retreival

seqret	retreives and/or changes format of a sequence
seqretset	retrieve and or change formats of a number
seqretall	of sequences at once
transeq	translate a DNA sequence to protein
backtranseq	translate a protein sequence to DNA
extractseq	extract regions from a sequence
cutseq	remove a region from a sequence
pasteseq	inserts a sequence into another sequence
infoseq	display information about a sequence
splitter	split a sequence into smaller sequences

Sequence comparison

needle	Needleman-Wusch sequence alignment
water	Smith-Waterman sequence alignment
stretcher	Myers and Miller global alignment
matcher	Huang and Miller local alignment
dottup	dotplot comparisons of two sequences.
dotmatcher	
prettyplot	plots multiple sequence alignments
polydot	dotplot comparisons of multiple sequences.
supermatcher	

Sequence parameters

cusp	generates a codon usage table
syco	synonymous codon usage plot
dan	calculates DNA/RNA melting temperature
compseq	sequence composition tables

DNA Sequence features

remap	restriction map of the sequence
cpgplot	CpG island detection
cpgreport	
etandem	finds tandem and inverted repeats
einverted	
plotorf	plots potential ORFs
showorf	pretty display of potential ORFs
fuzznuc	DNA pattern search
tfscan	scans sequence for TF binding sites

Protein Sequence features

ief	Isoelectric point calculation
antigenic	Finds potential antigenic sites
digest	protein digestion map
findkm	Vmax and Km calculations
fuzzpro	protein pattern search
garnier	protein 2D structure prediction
helixturnhelix	finds nucleic acid binding motifs
octanol	displays protein hydropathy
pepwindow	
patmatdb	searching with motifs vs protein sequences
patmatmotifs	
pepcoil	predicts coiled coil regions
pepinfo	Protein information
pepstats	
pepwheel	shows protein sequences as a helix.

File formats supported by EMBOSS

IntelliGenetics, Genbank, NBRF, EMBL, GCG, DNAStrider, Fitch, FASTA, Phylip, PIR, MSF, ASN.1, PAUP, ClustalW

This Quick Guide was written by and is copytight Dr David Martin at the Norwegian EMBnet node.
Comments and suggestions for improving this guide should be addressed to him at david.martin@biotek.uio.no

EMBnet is a network of academic and commercial bioinformatics institutes, supporting bioinformatics research and collaboration in more than countries worldwide.
More information about EMBnet and details of your local node can be found at http://www.embnet.org

An unlimited noncommercial right to redistribute the unamended document in printed or electronic form is granted without restriction.