

EMBOSS:

Alternative software to GCG

April 17, 2002

Sequence Analysis & Consulting Service

Seminar Series

Susan Johns

(johns@cgl.ucsf.edu)

To work along during this seminar session you need to be logged into your socrates account.

Log into your socrates account by clicking on the socrates icon in the small window on the left hand side of the screen and responding with your account name and password.

```
emboss_setup<rtm>  
cd embossy<rtm>
```

As a result of these commands, you are now in a new sub-directory called **embossy** in your account where the necessary files are located.

EMBOSS

EMBOSS is a new, free Open Source software analysis package specially developed for the needs of the molecular biology (e.g. EMBnet) user community. The software automatically copes with data in a variety of formats and even allows transparent retrieval of sequence data from the web. Also, as extensive libraries are provided with the package, it is a platform to allow other scientists to develop and release software in true open source spirit. EMBOSS also integrates a range of currently available packages and tools for sequence analysis into a seamless whole. EMBOSS breaks the historical trend towards commercial software packages.

EMBOSS

The name EMBOSS stands for European Molecular Biology Open Software Suite. As indicated in the history link, its creation is the direct result of a change in software distribution policy at GCG.

The suite had been in a state of constant revision until this past July (2001) when a stable

release became available. Its evolution took some time, since it needed to insure that all its code was indeed "open source" and freely available to all.

The programs developed concentrated first on those which produced text output, since graphical output required more intergradation of various software libraries with decisions as to which code should be supported. Graphical programs are still the weakest part of the suite.

EMBOSS programs fall into the following general categories.

ALIGNMENT

- CONSENSUS
- DIFFERENCES
- DOT PLOTS
- GLOBAL
- LOCAL
- MULTIPLE

DISPLAY

- EDIT
- ENZYME KINETICS
- FEATURE TABLES
- INFORMATION
- NUCLEIC

- 2D STRUCTURE
- CODON USAGE
- COMPOSITION
- CPG ISLANDS
- GENE FINDING
- MOTIFS
- MUTATION
- PRIMERS
- PROFILES
- REPEATS
- RESTRICTION
- TRANSCRIPTION
- TRANSLATION

PHYLOGENY

- PROTEIN
- 2D STRUCTURE
- 3D STRUCTURE
- COMPOSITION
- MOTIFS
- MUTATION
- PROFILES

UTILS

- DATABASE CREATION
- DATABASE INDEXING
- MISC

EMBOSS

EMBOSS uses a different approach to writing software than is usually used. The package

employs an application command definition file for input testing of any application. A new program is written starting with ACD file and then added to the resource prior to any code actually being written. Once this has been done, the source code for the new program is added and debugged.

EMBOSS programs are written in C. The graphics library is an LGPL library that was developed outside of the EMBOSS project. Information on developing code is available as well as programming standards to be used.

Other programs than have been converted into EMBOSS format are also distributed with EMBOSS and are known as the EMBASSY suite. Currently these programs or software packages include, HMMER-2.1.1, MEME-2.3.1, MSE-0.0.4, PHYLIP-3.573c and TOPO-0.1.

EMBOSS

EMBOSS was really created because a number of folks had spent a lot of time developing code to extent the functionality of GCG and were burned when GCG changed its software distribution policy. But, it does have some advantages as well.

Now that data for whole chromosomes and genomes exists, all the sequence analysis packages are struggling to deal with these large data files. GCG has some real problems using these data files. GCG requires that they be broken up into smaller blocks in order to be used. This process is not only time consuming, but, could also leads to problems when a sought for feature extents over created data file junctions.

Because EMBOSS is newer than the other packages, its handling of large files is more up-to-date. It will handle these large files without any additional data file processing.

EMBOSS also has a more relaxed approach to dealing with the various sequence analysis formats. It has the software determine the format of the data to be used instead of requiring the user to be format conversion literate.

EMBOSS has an online tutorial in its general documentation page. This tutorial can be downloaded as a postscript file which is 11 MBytes in size. There is also a discussion of themes in Bioinformatics from EMBOSS's point of view.

EMBOSS programs are run by typing them at the UNIX prompt. Any required information that you have not already given on the command-line will be prompted for.

If in doubt, type:

programname -help to get some help on the options

or

programname -opt to make the program prompt you for common options

or

tfm programname to get the full help on a program

Using EMBOSS means becoming familiar with a whole new set of program names. A real effort has been made to make the programs names in EMBOSS as distinctive as possible. However, this also increases the learning curve in getting to be familiar with the software suite. To get around this problem, use the page to find the general function topic you are interested in.

The EMBOSS folks realize that GCG will still be used by the sequence analysis community. Many people depend on it and its formatted databases to do their research. Therefore, EMBOSS can be configured to read GCG formatted databases.

SACS has a number of databases. Some are updated every night from NCBI and some are only updated when a new release comes along. EMBOSS requires that its indexing be run whenever the data is updated. This process hasn't been put into the automatic scripts yet. However, the necessary control file has been created and initial indexes have been run on the following databases.

database	name	contents
genbank	genbank	regular genbank divisions
tags	tags	genbank est, htgs, sts divisions
genpept	genpept	all genpept divisions
pir	pir	pir, nrl_3d, patchx data
swissprotein	swiss	swissprot and update data

Making the indexes automatically should be in place by the beginning of the summer.

Getting graphical applications to work has been a problem for EMBOSS. The selection of a "open source" graphics library was difficult and still seems to be in a state of flux. The package concentrated on getting text applications to work first and then turned its attention to graphical ones.

EMBOSS has a list of known devices, including postscript, ps, hpgl, hp7470, hp7580, meta, colourps, cps, xwindows, x11, tektronics, tekt, tek4107t, tek, none, null, text, data, xterm, png. The default output device is X11.

Running some of the graphical programs currently available shows them to be rather basic. Experimentation has shown that X11 graphics (at least with MacX) is not very good. Postscript works. But, you really need to explore the documentation for a graphical program to decide if it will be helpful to you or not. Getting the output to be just what you want can also be a trick.

```
program_name -graph=device_name
```

```
pepnet -graph=ps
```

```
pepnet -graph=png
```

Let's start exploring EMBOSS by doing some simple alignment tasks. Use the group page to find those general topics that deal with alignments. Then go the desired topic to find out the name of a program to use for the desired task.

Do a global and then a local alignment with the ccho and cca tfa files in your current directory location. Look at your results. How does this compare to GCG output for the same task?

Then do a multiple alignment with the [veggie.tfa](#) file.

While the clustalw job runs fine, there isn't a real way of getting a good look at your output. The default output format for EMBOSS is a fasta collection file of the alignments.

To look at the text version of the alignment you are supposed to use showalign. This program takes the output file and makes it more like a msf file. Care needs to be taken with the command switches to get what you are looking for.

```
showalign -nobottom -show=all
```

The prettyplot program is what should work to display the data. During testing with MacX and the default X11 output option there was only the first page of the output. Changing to ps for postscript results created output that is difficult to colorize.

Let's continue exploring EMBOSS by doing

some simple protein motif tasks. Use the group page to find this general topic. Then go the topic to find out the names of programs to use.

There are a number of programs here that look for patterns or motifs in a sequence. Check for PROSITE motifs and PRINTS fingerprints in the [pir:ccca](#). Also check for possible antigenic sites in this protein.

On to protein structure information. From the realm of 2D analysis, do a garnier secondary structure prediction on [swiss:ag22_mouse](#).

Continue exploring EMBOSS by doing some simple nucleotide analysis tasks. Use the group page to find this general topic. Then go the topic to find out the names of programs to use.

First, look into nucleic restriction programs. Use [restrict](#) and then [remap](#) on the [genbank:dmsnail](#) sequence. How do the results compare with one another? What features are the same as GCG programs?

Then, check out nucleic gene finding programs. Use [getorf](#) to find the open reading frames in the lamcg sequence from genbank. Next, use [wobble](#) to find the location of the possible open reading frames in the [genbank:dmsnail](#) sequence. Have wobble create ps output and print the results on the local printer (class).

```
wobble -graph=ps
```

```
lpr -Pclass wobble.ps
```

As you have seen, the EMBOSS suite of programs is a very diverse set of tools. Some depend on the user going through a number of steps to get desired result. Others are very straight forward and do very simple tasks.

There is a learning curve with the package. The documentation is not as complete as one would like. The software has kept its original command line switch emphasis which works well with advanced users, but is a problem for beginners. Not all the switches appear to be given which confuses things.

As time permits the auxiliary parts of the suite will be brought online. The W2H interface to the package is now available. The EMBOSS package creates a more complete working environment for doing sequence analysis tasks. EMBOSS is even branching out into true 3D structure and mass spec programs.

URLs used in this seminar:

EMBOSS home page:

<http://www.emboss.org/>

EMBOSS overview :

<http://www.uk.embnet.org/Software/EMBOSS/overview.html>

EMBOSS history page:

<http://www.uk.embnet.org/Software/EMBOSS/history.html>

EMBOSS development page:

<http://www.uk.embnet.org/Software/EMBOSS/Doc/index.html>

EMBOSS themes in Bioinformatics page:

<http://www.uk.embnet.org/Software/EMBOSS/Themes/>

EMBOSS approach to sequence formats page:

<http://www.uk.embnet.org/Software/EMBOSS/Themes/UniformSequenceAddress.html>

EMBOSS sequence formats discussion page:

<http://www.uk.embnet.org/Software/EMBOSS/Themes/SequenceFormats.html>

EMBOSS tutorial page:

<http://www.uk.embnet.org/Software/EMBOSS/Doc/Tutorial/>

local EMBOSS documentation:

<http://www.sacs.ucsf.edu/Documentation/emboss/>

local EMBOSS group page:

<http://www.sacs.ucsf.edu/Documentation/emboss/groups.html>

local EMBOSS tutorial postscript file:

http://wwwtest.sacs.ucsf.edu/Training/emboss/emboss_tutorial.ps

local EMBOSS W2H interface page:

<https://wwwtest.sacs.ucsf.edu/secure/cgi-bin/w2h-emboss/w2h.start>

EMBnet home page:

<http://www.embnet.org/>