Software review

Emboss opens up sequence analysis

Free software for analysing DNA and protein sequences on the UNIX platform tends to be difficult to use and does not always interact seamlessly with other programs. Comprehensive packages of integrated applications from commercial vendors are easier to use, but may be too expensive for many laboratories and cannot be customised for local use because the source code is not available. The creators of the European Molecular Biology Open Source Software Suite (EMBOSS) aim to bridge this gap with a comprehensive package of integrated sequence analysis applications released under the open source model on UNIX platforms. This review examines EMBOSS as a sequence analysis package and also as a platform for developing new analysis programs.

EMBOSS (European Molecular Biology Open Software Suite[1]) is an open source software project with two goals: to provide the molecular biology community with a freely available software package for analysing DNA and protein sequences and to provide a set of software libraries for scientists to use to develop their own applications. It is actively developed by researchers associated with the European Molecular Biology network (EMBnet), predominately from the Sanger Centre and the UK Human Genome Mapping Project (HGMP) Resource Centre. EMBOSS has its origins in the EGCG package of extensions to the commercial Genetics Computer Group (GCG) Wisconsin Package.[2] GCG originally released its source code to the public, allowing outside programmers to develop applications that used the GCG libraries and worked within the GCG data environment. When GCG closed access to its source code, the EGCG developers set out to design and implement an open source replacement.

Version 2.01 of EMBOSS provides over 100 analysis programs and a set of core libraries. In addition, several publicly available software packages have been integrated into the EMBOSS environment; these are collectively referred to as EMBASSY. The package runs on most major UNIX platforms (IRIX, Solaris, Tru64 UNIX, Linux, FreeBSD and Macintosh OS X). The software may be obtained by anonymous FTP from the UK EMBnet node.[3] Documentation and e-mail support are available through the EMBOSS web page.[1] For this review, the software was installed on a Sun Ultra-Sparc under Solaris 5.8.

INSTALLATION AND CONFIGURATION

The software is installed using the GNU[4] configure and build system. Installation and configuration is straightforward for anyone familiar with UNIX system administration, but if you have little experience with UNIX, it would be best to seek the assistance of your local expert and to read the Tutorial section in this issue which is devoted to EMBOSS. An *Administrator's Guide* provided with the package explains the process in detail, including platform-specific modifications. (A more recent version of this guide is available on the EMBOSS web page – make sure to get this newer document if you will be using EMBOSS on Mac OS X.) Graphics output uses the

PLplot library,[5] which is provided with the package. You may need to obtain and install additional libraries (X11, z, gd, png) for certain types of graphics.

Databases are not distributed with EMBOSS, but the software can index sequence databases obtained from the database centres or their mirror sites. It can also index GCG-format databases, so you do not need to keep multiple copies of a database if you also use the Wisconsin Package. Specialised databases, such as REBASE,[6] PROSITE[7] and PRINTS,[8] can also be used with EMBOSS.

APPLICATION OVERVIEW

EMBOSS contains the expected standard set of applications for analysing DNA and protein sequences: restriction/proteolytic enzyme mapping, translation and reverse translation, pair-wise global and local alignment, open reading frame analysis, and secondary structure prediction, among others. An interface (emma) to CLUSTAL W[9] allows multiple sequence alignment within the EMBOSS environment. There are also a number of applications not commonly found in sequence analysis packages, such as plotting isochores and CpG-rich regions, finding MAR/SAR sites and calculating the twisting of a B-DNA sequence.

No sequence editors are included in EMBOSS proper, although the MSE multiple sequence editor is available as an EMBASSY program (not tested). In practice, this is not a drawback. New short sequences may be entered from the keyboard and EMBOSS's rich array of sequence manipulation tools can be used to perform many editing functions such as removing gap characters, trimming ambiguous characters from sequence ends, deleting or extracting specified regions of a sequence and inserting one sequence into another.

Programs are also provided to extract sequence entries from a database, to extract sequence fragments from databases based on an entry's features table information, and to convert sequence files from one format to another. Format conversion is usually only necessary to prepare sequences for working with other programs, because EMBOSS programs automatically recognise and read many sequence formats. There are no programs for sequence assembly, and with the exception of a distance matrix program, EMBOSS contains no phylogenetic analysis programs. Many of Joseph Felsenstein's PHYLIP[10] phylogenetics programs are included in EMBASSY however, and these work smoothly within the EMBOSS environment.

The most noticeable gap in analysis coverage is database-searching programs. There are few programs that can use a database as a search set, and they are not as useful as they might be. For example, the output files for the programs stssearch (STS primers *v.* a DNA database) and profit (a profile *v.* a database) list every sequence in the order it was encountered in the database. Poor matches are not dropped, so the output files can be very large. Since the profit output is not sorted in the order of best matches, you must either write a script to sort the output, or scan the file by eye to find the highest-scoring sequences. We could not get the program est2genome (aligns a genomic sequence to ESTs) to work with more than a few EST sequences – attempts to compare a genomic sequence to the EST division of GenBank resulted in out-of-memory messages and segmentation faults.

WORKING WITH EMBOSS

As mentioned above, EMBOSS programs automatically recognise many single and multiple sequence formats, so you can use sequence files obtained from other sources without having to convert them first. (This is also true for EMBASSY programs.) Thus EMBOSS works well in an environment where many different programs are in use. There are no sequence length limits imposed by the programs. All of the programs can be run entirely from the command line, and thus are scriptable. (To see a list of all of a program's qualifiers, type -help after the

program name.) EMBOSS programs warn you if they do not recognise a program qualifier, so you cannot run a program with unintended options because of typographical errors. A detailed explanation of the command-line syntax is available on the EMBOSS web page.

EMBOSS programs can also be run interactively, with two levels of program interaction available. By default, a program will prompt you only for required information. If you add -options to the command line, the program will also prompt you for optional qualifiers that can affect the way the analysis works. For example, sigcleave's results depend on whether or not the sequence is from a eukaryotic or prokaryotic organism. The program assumes eukaryotic, and will not prompt you to change this unless -options is on the command line. Advanced qualifiers are never prompted for. For example, programs that use codon usage tables read a human table by default. You cannot specify another table interactively – this information can only be entered on the command line.

While EMBOSS programs can use list files (files of filenames) as input, none of the programs produce list files as program outputs, so you cannot link programs together with list files. However, if you put the -filter qualifier on the command line, an EMBOSS program will get its first-named input from standard input and send its first-named output to standard output, allowing programs to be linked via UNIX pipes. For example, to extract the translations of all *Mycoplasma genitalium* open reading frames at least 500 nucleotides long, search each resulting protein against the PRINTS database of protein signatures, and save the output in a file named hits, type the command:

```
getorf L43967.gbk -minsize
                500 -filter | pscan -filter > hits
```

Program outputs are not copiously annotated. In particular, we missed having details about the program parameters recorded in the output. In some cases, the minimal annotation causes problems. For example, the output from pscan (searches protein sequence(s) against the PRINTS database) does not contain the name of the sequence used as the query. If many query sequences are used in a single program run, it can be difficult to relate each signature in the output to the corresponding query sequence.

The EMBOSS distribution comes with a tutorial that guides you through some common analyses. This tutorial (*Introduction to Sequence Analysis using EMBOSS*) is also available in HTML form on the EMBOSS web page. There are also several types of on-line help. As mentioned above, typing -help after the name of a program displays a list of its command-line options. Adding -verbose as well causes global command-line options to appear in the list. (Global options are not program-specific and may be used with any program.) Typing wossname displays a list of all EMBOSS and EMBASSY programs along with short descriptions of their functions, and typing showdb displays information about all databases that are accessible to EMBOSS. To view the documentation for an individual program, type tfm followed by the program name, for example, tfm sigcleave. Each program document gives an overview of the program and a brief description of each of its command-line options, but does not provide much detailed information about the analysis or how changes in the values of the options can affect the analysis. Literature references are cited so that users can look up the details themselves.

A number of user interfaces have been developed to replace the EMBOSS command line. A text-based menu interface (emnu) is provided as an EMBASSY program. (It is similar to the DOS menu interfaces that were common in pre-Windows days.) We also installed two web interfaces, Catherine Letondal's Pise[11] and the W2H interface of Senger *et al.*,[12] both of which work well. Several other web interfaces, X11 interfaces and a Java interface are under development (not

tested). See the EMBOSS web page for links to these.

PROGRAMMING ENVIRONMENT

One of the EMBOSS goals is to provide software libraries that researchers can use to build their own applications. We programmed a simple EMBOSS application in order to investigate this aspect of the package. EMBOSS is built on two libraries. AJAX contains low-level functions for string handling, reading and writing sequence files, memory management, mathematical operations, list management and so forth. NUCLEUS contains high-level functions that are mostly molecular biology algorithms for alignments, pattern matching, etc. The code is written in ANSI C for portability reasons, but it borrows concepts from C++, and encourages manipulating objects using object methods rather than accessing the underlying data directly. The library code is well documented, and functions have a defined header format so that the function documentation can be extracted easily by means of a script. In order to integrate your application into EMBOSS, you must create an AJAX command definition (ACD) file that defines the user interface for the application. This is the system by which EMBOSS can detect mistyped command-line qualifiers. (The ACD files also make it easier to create alternative user interfaces for the EMBOSS package.)

A *Programmer's Guide* is distributed with EMBOSS as a PostScript file, or it may be read in HTML form on the EMBOSS web page. With this document, and the examples afforded by the existing code, an experienced programmer can create a simple EMBOSS application or modify an existing EMBOSS application fairly rapidly. For more complex programs, consult the EMBOSS web page for detailed documentation on creating ACD files and using the AJAX and NUCLEUS library functions. The makefiles supplied with EMBOSS assume that all code and all ACD files are in the EMBOSS distribution directories. To minimise the chance of losing local code when a new release is installed, it would be useful to modify the makefiles so that local applications and modifications to existing EMBOSS applications can be built from a source code directory separate from the distribution directories.

CONCLUSIONS

The EMBOSS package contains the most commonly used sequence analysis applications, can handle sequences of unlimited length, and can seamlessly use sequences and databases from many sources. Installing and administering the package is not difficult for someone familiar with UNIX. It is ideal for computer-savvy individuals wanting a free sequence analysis package for their own desktop UNIX computers. It is also suitable for a departmental server, especially when coupled with one of the web interfaces geared toward less experienced users. Access to the source code is a plus for those who want to modify the existing programs (if only to alter an output format) or who want to develop their own applications without having to spend a lot of time on 'housekeeping' aspects, such as reading different file and database formats. As with other open source software projects, the EMBOSS originators hope that it will become a community project, and they encourage any programmers interested in sequence analysis applications to become involved in the development of EMBOSS.

If you are looking for a sequence analysis package that contains both analysis programs and database searching programs in a single integrated environment, the current version of EMBOSS will not quite fill the bill. And microcomputer users may prefer software designed specifically for their machines over EMBOSS, even with one of its alternative interfaces. But if EMBOSS's collection of applications is sufficient, if you do not want the expense of a commercial package, or if you want a

supplement to a package you already have, EMBOSS is worth considering – especially if you want the flexibility provided by access to source code, want to write our own analysis programs or want to contribute to a community project.

Sue A. Olson
Center for Genomics and Bioinformatics,
Indiana University,
Bloomington,
Indiana 47405, USA

References

1. EMBOSS home page: URL: http://www.uk.embnet.org/Software/EMBOSS/
2. Devereux, J., Haeberli, P. and Smithies, O. (1984), 'A comprehensive set of sequence analysis programs for the VAX', *Nucleic Acids Res.*, Vol. 12, pp. 387–395.
3. The UK EMBnet node: URL: ftp://ftp.uk.embnet.org/pub/EMBOSS/
4 The GNU Project home page: URL: http://www.gnu.org/
5. The Plplot home page: URL: http://plplot.sourceforge.net/
6. Roberts, R. and Macelis, D. (2001), 'REBASE – restriction enzymes and methylases', *Nucleic Acids Res.*, Vol. 29, pp. 268–269. URL: http://rebase.neb.com
7. Hofmann, K., Bucher, P., Falquet, L. and Bairoch, A. (1999), 'The PROSITE database, its status in 1999', *Nucleic Acids Res.*, Vol. 27, pp. 25–219. URL: http://www.expasy.ch/prosite/
8. Attwood, T. K., Flower, D. R., Lewis, A. P. *et al.* (1999), 'PRINTS prepares for the new millennium', *Nucleic Acids Res.*, Vol. 27, pp. 220–225. URL: http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
9. Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994), 'CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice', *Nucleic Acids Res.*, Vol. 22, pp. 4673–4680.
10. Felsenstein, J. (1993), 'PHYLIP (Phylogeny Inference Package) version 3.5c', Distributed by the author. Department of Genetics, University of Washington, Seattle WA, USA. URL: http://evolution.genetics.washington.edu/phylip.html
11. Letondal, C. (2001), 'A Web interface generator for molecular biology programs in Unix', *Bioinformatics*, Vol. 17, pp. 73–82. URL: http://www-alt.pasteur.fr/~letondal/Pise/
12. Senger, M., Flores, T., Glatting, K.-H. *et al.* (1998), 'W2H: WWW interface to the GCG sequence analysis package', *Bioinformatics*, Vol. 14, pp. 452–457. URL: http://industry.ebi.ac.uk/w2h/