EMBOSS: The European Molecular Biology Open Software Suite

Peter Rice
pmr@sanger.ac.uk

Ian Longden
il@sanger.ac.uk

***Alan Bleasby**
ableasby@hgmp.mrc.ac.uk

The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK CB10 1SA.
*HGMP-RC, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK CB10 1SD.

In the area of sequence analysis, biologists find that existing software developed in the early days of sequencing often falls short of their needs. The rapid increase in large-scale sequencing and the challenges of the post-genomic era lead to a need for the rapid development of new applications, or the enhancement of existing software. This has been limited by the need for a general purpose-framework for the academic development of sequence analysis software. A survey by Pitt in 1998 (http://www.hgmp.mrc.ac.uk/CCP11/newsletter/vol2_1/questionnaire/) found that most bioinformatics developers were using C and that, although only 18% were using a free software library, a further 54% were likely to do so in future.

There is a further need to integrate existing packages and databases more effectively than present methods are able to. These problems have been debated at length by the members of EMBnet[1], with the conclusion at the 1996 meeting in Helsinki that there should be a concerted effort to provide an integrated sequence analysis software suite.

The EMBOSS suite (http://www.sanger.ac.uk/Software/EMBOSS/) is the result of this effort. With over 100 applications (Table 1) and the capability to be run from the traditional command line, through a web browser, or under more advanced graphical user interfaces, EMBOSS is suited to the needs of both expert users and practising scientists. The package runs on all Unix platforms, and it has also been built for Microsoft Windows NT (R. Bruskiewich, pers. commun.).

TABLE 1. Some of the most popular EMBOSS applications[a]

Program group	Selected applications
Alignment	Local (matcher, water) and global (stretcher, needle) alignment
Coding regions	Synonymous codon use (syco), codon statistics (chips)
Comparison	Large sequence word comparisons (dottup, polydot, wordmatch) and alignment (supermatcher)
Composition	General (compseq), frequent words (wordcount), graphical representation (chaos)
CpG islands	Report (cpgreport) and plot (cpgplot)
DNA features	Repeats (einverted, etandem), DNA melting (dan)
Editing	General editing utilities (cutseq, splitter), features (maskfeat), etc.
Indexing	Database indexing (dbiflat, dbigcg, dbiblast)
Motifs	Searching prosite (patmatdb, motifsearch), prints (pscan), transfac (tfscan), general patterns (fuzznuc, fuzzpro)
Multiple alignment	Interface to clustalw (emma), display (prettyplot), editing (mse)
Protein features	Functional motifs (antigenic, sigcleave), structural motifs (pepcoil, helixturnhelix), amphipathic regions (pepnet, pepwheel), transmembrane prediction (tmap) and display (topo)
Protein properties	Hydropathy (pepwindow, pepwindowall, octanol), protease sites (digest), general (pepstats)
Sequence formats	Sequence reading/writing/format conversion (seqret, seqretall) and feature format conversion (seqretfeat)
Translation	Codon usage (cusp), reading frames (getorf, showorf, backtranseq)
Utilities	Motif database indexing (rebaseextract, tfextract, prosextract, printsextract), listing databases (showdb), searching for applications (wossname)

[a]For a full listing, see http://www.sanger.ac.uk/Software/EMBOSS/Apps/

TABLE 2. EMBOSS interfaces under development

Interface	Developer(s)	URL
W2H	EMBnet Germany	http://www.uk.embnet.org/embnet.news/vol4_1/w2gcg.html
www2gcg	EMBnet Belgium	http://www.uk.embnet.org/embnet.news/vol4_2/www2gcg.html
DisGUIse	EMBnet UK	http://www.hgmp.mrc.ac.uk/
PISE	Institute Pasteur	http://www.pasteur.fr/~letondal/Pise/ http://bioweb.pasteur.fr/seqanal/EMBOSS/
SRSWWW	EBI/Lion	http://srs.ebi.ac.uk/
ACEDB	Sanger Centre, UK and CNRS, Montpellier, France	http://www.acedb.org/
CINEMA	EMBnet Manchester, UK	http://www.umber.embnet.org/dbbrowser/
SeqPup	Indiana, USA	http://iubio.bio.indiana.edu/soft/molbio/seqpup/java
Staden	LMB, Cambridge, UK	http://www.mrc-lmb.cam.ac.uk/pubseq/
BioNavigator	EMBnet Australia and Encompass Pty	http://www.bionavigator.com/

Data sources

Sequence data can be used in any of the common sequence formats, with new formats easily added. For institutes where the public sequence databases are installed, EMBOSS can use these with a choice of formats, including GCG (Wisconsin Package Version 10, Genetics Computer Group, Madison, WI, USA; http://www.gcg.com/), SRS (Ref. 2), STADEN (http://www.mrc-lmb.cam.ac.uk/pubseq/), ACEDB (Ref. 3, see http://www.acedb.org/) and BLAST databases[4]. Failing this, individual sequence entries can be retrieved automatically over the web. Private local databases can also be used, for example by industrial users. Work is in progress on converting annotation (feature) formats.

The user interface

The basic interface is the command line. This is needed for many reasons, including the ability to run applications as part of larger automated analyses. The command line is defined in 'ACD' definitions (http://www.sanger.ac.uk/Software/EMBOSS/Acd/).

Most of the more friendly user interfaces in bioinformatics are simple forms which build the command line for the user, run the program, and present the results. EMBOSS is easy to convert to

 PII: S0168-9525(00)02024-2

FIGURE 1. EMBOSS

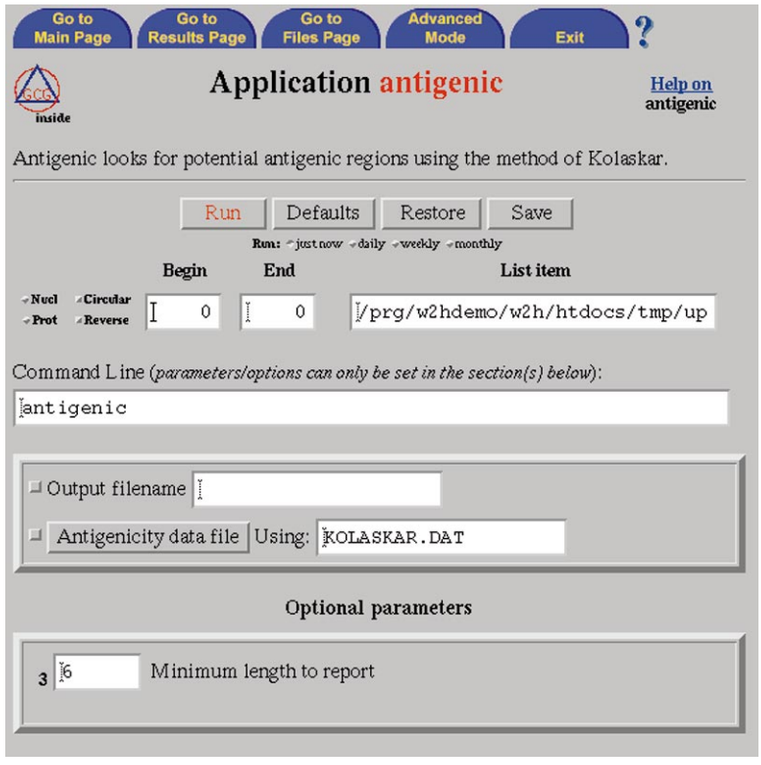
(a)

```
% antigenic
Finds antigenic sites in proteins
Input sequence(s): swissprot:act1_fugru
Minimum length [6]:
Output file: [act1_fugru.antigenic]: actin.anti
%
```

(b)

```
% antigenic swissprot:act1_fugru -out actin.anti -auto
```

(c)



trends in Genetics

(a) The command line interface, with prompts to the user. (b) Automated, with everything on the command line. (c) The W2H Web Interface, generated automatically.

these interfaces by using the ACD definitions to generate the forms for each application automatically. Figure 1a, b shows an example application running from the command line, and Fig. 1c shows the web form for running the same program under the W2H web interface. Similar forms are available for other interfaces, including PISE (http://biowebpasteur.fr/seqanal/EMBOSS/). Table 2 lists the web and GUI interfaces we are currently collaborating with.

EMBOSS for developers

Fashions change in bioinformatics, and a number of programming languages have been favoured. In EMBOSS we have chosen to use standard C but we also support development in other languages. The first of these are from the Birkbeck Template Library (http://www.cryst.bbk.ac.uk/~classlib bioinf/BTL99.html). The licensing is open source (Gnu General Public Licence; http://www.gnu.org) so that EMBOSS is freely available. The software libraries are under the library GPL to allow other packages to link to EMBOSS. Examples so far include Phylip (http://evolution.genetics.washington.edu/phylip.html) and the MSE sequence editor.

Applications

The original goal of EMBOSS was to re-implement applications in the 'EGCG package' (see http://www.uk.embnet.org/embnet.news/vol3_1/software.html) of extensions to GCG. This has now been achieved, with all the popular EGCG applications either replaced by EMBOSS or by other free software.

Additional applications have been contributed by several EMBnet members with their own development groups. The major contributors to date have been the UK node at Human Genome Mapping Project in Hinxton, Cambridge UK, with further programs from Norway, Germany, Italy, The Netherlands and the European Bioinformatics Institute. For most of the past year new applications have been added at the rate of ten per month. Most other EMBnet nodes are actively contributing to the project with interface developments, testing and documentation. The additional resources provided in this way are a typical benefit of open source software. Full documentation on all programs is available at http://www.sanger.ac.uk/Software/EMBOSS/Apps/.

Releases

The first developers' release was in the summer of 1998, together with an EMBnet workshop to review developments. The full beta-test release was made available in summer 1999, with a second EMBnet workshop where user feedback was gathered. We have implemented most of the recommendations from that workshop, and the first official release of EMBOSS followed in Spring 2000. The latest details are available from the EMBOSS Web pages (http://www.sanger.ac.uk/Software/EMBOSS/).

References

1 Harper, R. (1997) EMBnet, an institute without walls. *Trends Biochem. Sci.* 21, 150–152
2 Etzold, T. and Argos, P. (1993) SRS, an indexing and retrieval tool for flat file data libraries. *Comp. Appl. Biosci.* 9, 49–57
3 Stein, D.L. (1999) Internet access to the *C. elegans* genome. *Trends Genet.* 15, 425–427
4 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
5 Rice, P. *et al.* (1996) EGCG 8.1 Release Notes. *Embnet.news* 3, 2–4