

Bench philosophy (13): EMBOSS Suite

Open Source for Open Minds

Commercial bioinformatic programmes like the genome database and sequence analysis programme GCG, sold by the software company Accelrys, are both expensive and restricted in use or programme modification. There is no chance to independently change the source code of such proprietary programmes or even to fix the toughest bugs. In recent years, however, highly efficient open source programmes, such as the EMBOSS suite, have emerged that challenge commercial bioinformatic programmes like GCG.

EMBOSS, which is short for European Molecular Biology Open Software Suite, is a completely free, open source bioinformatic software package covering more than 160 programmes for genetic analysis, sequence alignments and other bioinformatic tasks. The EMBOSS project was launched in 1996 by Peter Rice, then at the Sanger Centre, and Alan Bleasby at the Rosalind Franklin Centre for Genomics Research in Hinxton near Cambridge, UK. In 2005, EMBOSS moved to the European Bioinformatics Institute (EBI) located on the Wellcome Trust Genome Campus in Hinxton. EMBOSS is more of a framework than a programme, being pushed forward by about 20 groups coordinated by Rice and Bleasby, working on software solutions for molecular biology. The suite covers all the needs of genomic analysis from DNA sequence alignments and data mining to in depth analysis of enzyme kinetics of

the proteins under study. All EMBOSS programmes can communicate with each other and have access to the same libraries. Additionally, they are able to understand many different sequence codes, including those used by GCG.

Rising volume

Members of the EMBOSS staff are scattered around the globe, meeting each other only occasionally. To coordinate such a huge project, the EMBOSS people apply a Concurrent Versions System (CVS) approach, which means in simple words, that the source code is organised in a tree-like structure. Anytime a new code is integrated into an EMBOSS programme, its algorithm is carefully counter-checked by the EMBOSS community. That's crucial to the whole programme, since the risk that a new feature could destabilise the software increases when the volume of the source code rises. Any new programme lines and features are, therefore, first checked by other members of the working staff and subsequently by the EMBOSS users. In addition, newsgroups can discuss problems of a new release in a usenet forum and try to fix them. In contrast to commercial software companies, the working staff responds very quickly if problems occur with distinct programmes of the package. To avoid major problems, EMBOSS versions are divided into stable editions and developer releases. Usually, the developer releases cover new features, however, they may also contain some hidden errors.

Back-end and front-end

The structure of the EMBOSS Suite programmes can be divided into a back-end and a front-end part. The back-end contains the algorithms necessary for, e.g. plotting a sequence, finding a consensus sequence and establishing connections to database interfaces. It can be run on almost any UNIX equipment. The installation of the back-end



Jemboss, the graphical user interface.

programme is done by compilation. The code is delivered in the computer language C and should run on every target machine, whether it's a mainframe or a small laptop. However, if you try to code an EMBOSS programme on a Windows operating system, you will need an additional C++ compiler.

You may choose between several EMBOSS front-ends called graphical user interfaces (GUI). The Java front-end Jemboss is designed for operating on a local computer whereas the Web front-end, called EMBOSS Explorer, is restricted in speed and usage. Nevertheless, it will run on any thin computing front-end, such as an embedded Linux machine like the Asus Eee PC. Web interfaces support the installation of EMBOSS core programmes on a central computer, localised, for example, at a research institute. If genome databases are also implemented on this computer, the search utilities of EMBOSS have direct access to the information stored in the databases.

The EMBOSS code and the routines for installation are distributed from many European EMBOSS centres. Some of them not only offer free download of the installation package, but they also provide additional information. The MPI for Molecular Genetics in Berlin, for example, allows access to both code and programme (www.molgen.de).



From the Equator...

mpg.de/~beck/embossfaq.shtml). EMBOSS portals such as the Swiss EMBNet node, the Computational Biology Research Group in Oxford (UK) or the Finnish IT Centre for Sciences, maintain fully installed versions, provide access to EMBOSS applications and you may also obtain advice from well-trained EMBOSS specialists.

Finding the optimal alignment

The EMBOSS package focuses on sequence comparison. Alignments of sequences, usually written as DNA codes, are based on the Needleman-Wunsch algorithm. Basically, two sequences are typed into a matrix and the algorithm returns a score close to one when the two sequences are highly similar. Let's take, for example, the two sequences A G T C and A C G T C. The algorithm recursively aligns the sequences and searches for the best alignment giving the highest value of similarity. In the example above, the Needleman-Wunsch algorithm will predict that ACGTC matches A-GTC. The alignment programme of EMBOSS is based on an optimised version of the genuine Needleman-Wunsch algorithm leading to better results, especially when comparing longer sequences. The aim is to find the optimal alignment, including gaps, of two sequences.

EMBOSS is on its way to becoming the leading software for genome analysis. More and more institutes and bioinformatic projects are migrating from commercial



... to the Arctic. EMBOSS is spreading all over the world.



Meeting of EMBOSS developers at the European Bioinformatics Institute.

sequence analysis programmes, like GCG, to EMBOSS. One major reason for that is given in the Questions and Answers section at the EMBOSS portal of the Max Planck Institute for Molecular Genetics (MPIMG) in Berlin: *"It is no longer possible to distribute an academic software source code, which uses the GCG libraries and [it] even has become difficult to distribute binaries"*. The Max Planck Institute for Plant Breeding Research (MPIZ) in Cologne, where I finished my own studies on a zinc finger protein a decade ago using GCG libraries, has also switched to EMBOSS. Martin Weber from Bernd Weishaar's group at the MPIZ (prior to 2003 at the MPIZ, now Professor for Genome Research at the Bielefeld University) has established a so called TF workbench based on EMBOSS, serving both as a repository and a tool to analyse data from MYB-type transcription factors of *Arabidopsis thaliana*. Since there are more than 130 members in this gene family, a lot of data has to be collected. Weber and Weishaar integrated the EMBOSS sequence analysis tools PHYLIP and MEME/MAST into the TF workbench. They complemented these tools with batch scripts in order to ease the work on great data volumes.

Cosmetic highlights

According to Kurt Stüber, member of the Bioinformatic Services at the MPIZ, the contract to the company Accelrys, which commercialises the GCG sequence analysis package, has been cancelled. Though the new release of GCG brought some "cosmetic" highlights like a modern GUI, there were

no real improvements. In his opinion, the complete work at the MPIZ could be performed using open source software. That's probably the better solution, since the total dependence on commercial bioinformatic software may bear a risk. Accelrys, for example, has recently announced on the company's website that: *"GCG is no longer being developed and it will no longer be offered after June 2008."*

While working with EMBOSS, one might wish to get involved in the developmental process. Before starting to programme, however, take a brief look at the style guide and documentation for developers published on the EMBOSS Web page. Working on the same project with people residing on different continents is a challenge. It's a good idea to initially attend an EMBOSS meeting after contacting the development crew by email. An introduction to the code is given at EMBOSS coordination meetings, which take place every fortnight at Hinxton Hall together with hands-on courses. If the application fee of around 160 euros (125 pounds sterling, excluding accommodation) is too expensive, one may download course tutorials and documentation instead. However, the direct contact to developers may be important.

MATTHIAS FAIX

Fancy composing an installment of "Bench Philosophy"?

Contact Lab Times
E-mail: editors@lab-times.org