embnet

embnet.news Volume 8 Nr1 (ISSN1023-4144) September 1, 2002

Editorial

It has been some time since the last edition of EMBnet news, and at the last EMBnet AGM in Vienna it was decided that the production of the HTML format would be abandoned and that the newsletter would now only be produced in PDF format. In the past it was always a time consuming process to collect the contributions in plain ASCII, and convert it to HTML and from there generate postscript files and finally produce PDF files. Many people only wanted to have a paper copy so the editorial board decided that the PDF format was deemed the best and only option.To get back into the swing of things it was decided to do an edition that concentrated on Education and Training. So there is a report on the EMBER project and also the work of the EMBnet node in China who have recently been teaching courses on EMBOSS, Jemboss, Staden, Stack, and the philosophy behind open source software.

China joined EMBnet in 1996 under the direction of Prof. Jingchu Luo and Centre of Bioinformatics at Peking University has made remarkable progress under his direction.

The centre has an SRS server running with over 100 indexed databases, which compares very favourably with the number of databases that are installed at the average European EMBnet node. In addition the Chinese EMBnet node mirrors many of the major databases held at the EBI, on their FTP server and thus serves as a primary point of distribution for databases and software in the pacific region.

The Peking University campus is undergoing a large reconstruction programme and as well was many new buildings springing up the campus has preserved and refurbished many of the older building of architectural interest. A large area of the campus is given over to parkland, and students take full advantage of this by spending their free time by the lakes under the trees. It was equally interesting to note that the campus now attracts many western students. So the recent EMBnet course in China should be viewed as a pioneering effort and should open the doors to many more cooperative efforts between China and EMBnet.

Contents

EMBnet Course
18 to 27 May 2002

Prof. Jingchu Luo
Embnet China Node
Centre of Bioinformatics
Peking University
Beijing 100871
China
Jingchu Luo <luojc@plum.lsc.pku.edu.cn>

The EMBnet course was held at the Peking University campus and was attended by 120 students. The course was divided into lectures, demos, and hands on practical sessions. The students came from many different Universities, and even a few of them came from Taiwan. The students were split into three groups of 40 and were taught in well-equipped computer rooms, which had the latest PC models running Windows

For Demos the teacher could broadcast their own session from their computer to all of the computers in the computer room, so the students could easily follow the steps required to run analysis with the Jemboss, Staden or Stack packages. Each computer also had a set of headphones so even those students that sat at the back of the PC rooms could hear the lectures without any difficulty. The maintenance of the computer rooms is under the control of the computer science department at Peking University.

Jemboss was taught by Lisa Mullan from the HGMP-RC at the Wellcome Trust Genome Campus in Hinxton. The Staden package was taught by James Bonfield from MRC-LMB in Cambridge, and the Stack package was taught by Janet Kelso, from the South African EMBnet node at SANBI. Johann Visagie, also from South Africa and who works at Electric Genetics gave a talk on the history of open source code, and many of the graduate students were eager to attend special lectures on this subject.

It then became obvious that the students after hearing the lecture on open sources were interested to learn more about software development, so separate sessions were organised for PhD students to learn more about the various bioperl, biojava, and biopython initiatives.

Students were highly motivated and many overcame the language barrier to learn about the various packages used on this course. Many of the students worked with their own data, to obtain results that were meaningful in their research. Individual help given to one or two students was soon turned into a demonstration for 6 or more, as everyone was eager to learn what their neighbour was being taught. It is refreshing to teach a class where every student is keen to learn as much as they can.

Jemboss

Lisa Mullan
Tim Carver
HGMP-RC
Hinxton, Cambridge
CB10 1SB
UK
lmullan@hgmp.mrc.ac.uk
tcarver@hgmp.mrc.ac.uk

Jemboss,[1,2] http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Jemboss/, is a new graphical interface for the EMBOSS suite of open source bioinformatics programs. It is written in Java to be platform independent. It uses the EMBOSS command line syntax information from the applications ACD (Ajax Command Definition) files to automatically generate a program form. In this way any new analysis programs incorporated into EMBOSS will appear in Jemboss. The forms are used by the users to enter sequences and parameters required for their analysis, see figure 2.

As part of the standard EMBOSS distribution, Jemboss can be downloaded from http://www.uk.embnet.org/Software/EMBOSS/ and installed locally. Installation notes can be found at the Jemboss home page. This allows the server to be setup by system administrators for their users. A script is provided to assist the installation of a server at other sites. The Jemboss server is being set up by a number of EMBnet nodes and has been installed at the Centre of BioInformatics in Beijing. The client may then be launched from the local Jemboss web page and EMBOSS applications run on the server local to that site. To launch the client from a web page the user needs to download Java Web Start for their operating system.

Jemboss has mainly been developed as a client-server application and is run as such on all the HGMP-RC training courses. When the network connections are too slow to allow course participants to use the HGMP-RC service, Jemboss is installed on a Sony VAIO laptop (256 M RAM, 20 GB hard drive, 800 MHz), which is then used as a Jemboss server for up to 20 users.

The interface itself is divided into three panes. The left pane consists of a list of EMBOSS and EMBASSY programs arranged into menus and sub menus representing different application categories, see figure 1. There is also an alphabetical list of programs which can be scrolled. A 'Go To' field allows the user to jump to a specific program by entering in the first few unambiguous letters of the application name. The right hand pane contains the local file manage, see figure 2. It is here that data stored on the local disk may be accessed, and the local home directory determines where any program output will automatically be saved. A 'remote file manager' is also accessible, from the "File" menu on the tool bar. This displays the user's directory on the

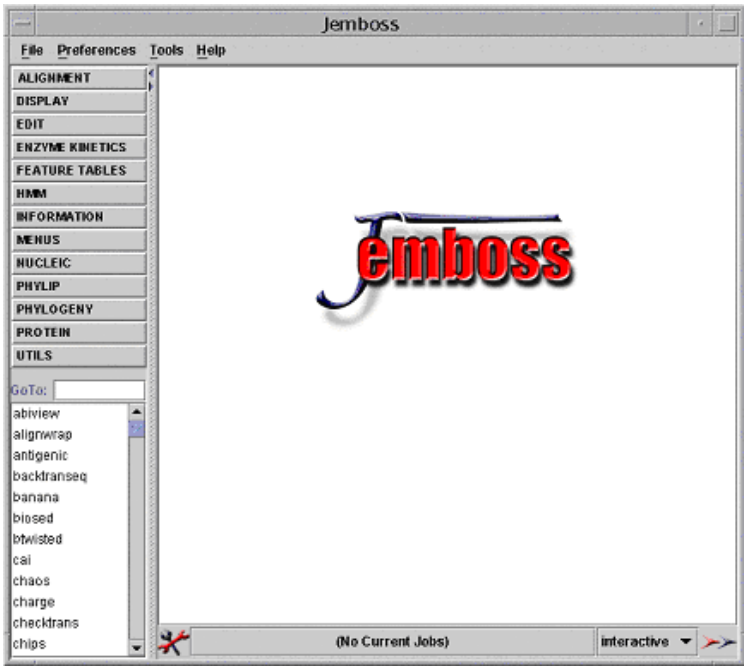


Figure 1 ***The Jemboss interface with program and group listings on the left. The Job manager is the tool bar at the bottom which indicates the number of batch/background jobs.***

server and files may be dragged between this file system and the local file system.

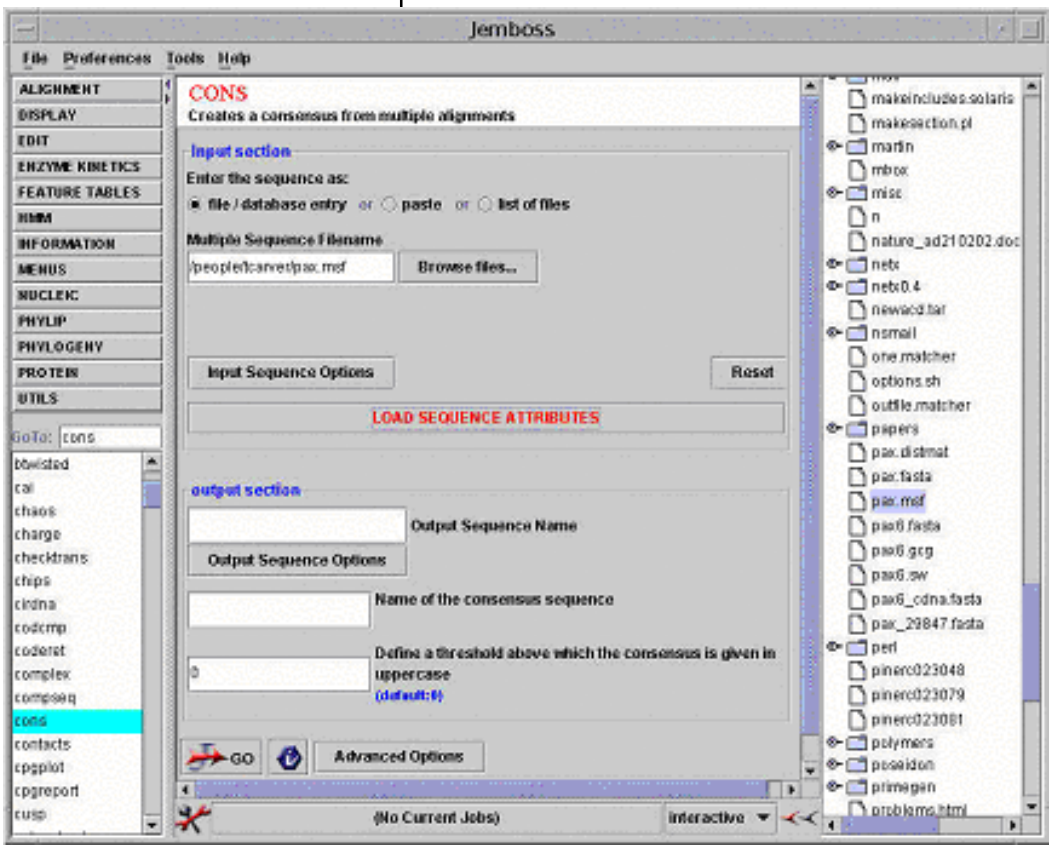


Figure 2
The EMBOSS application, 'cons', has been selected. The program form is shown in the central form and the local file manager is displayed on the right hand side of the window,

The central pane is the area in which the program form for each application is presented. This offers an area to input a sequence or filename, and also allows the input data to be pasted into the application. Once the input data has been entered, sequence attributes must be loaded in order for the program to recognize the type of data is expected to process, and, where applicable, calculate default parameters. Optional parameters may be entered, or, in the case of personal matrices or codon usage tables, dragged from their position in the local home or remote directory.

The "LOAD SEQUENCE ATTRIBUTES" button is used in some of the applications in Jemboss. This is used to mimic the command line EMBOSS implementation. It is used to contact the server with the input sequence and the type and length of the sequence is determined. The client then uses this information to update the client parameters. It is therefore important to click on this before running the application.

Once analysis parameters have been completed, the job is started using the "Go" button and the input data is converted to EMBOSS command line syntax to be sent to the server where the analysis is carried out. The program can be run in both an interactive and a batch mode. The former is default for many programs not requiring large computing resources and disables the Jemboss interface until the results are displayed on the user's screen. Alternatively, the job can be carried out in the background (batch mode). In this case, the process is handed to a job manager, which keeps the user informed of the status of the program ("running" or "completed"). Results are then be accessed through the job manager, which refreshes by default every 15 seconds. This can be altered to as little as 5 seconds, or as much as 60 seconds. Alternatively, the job manager list may also be refreshed manually.

Results appear in a tabbed pane. Together with the results pane, the parameters of the analysis may also be displayed in the form of EMBOSS command line syntax. Any input files are also displayed in individual tab windows. Any of the data represented in the results window may be saved,

although the actual analysis results are possibly the most useful data to save. To ensure that the correct data is saved, the pane containing that data must be selected in the results window. The "Save" option is in the "File" menu of the results window toolbar and will automatically save the data to the local home directory. Other directories may be specified using the "Browse" facility, although it is only possible to save to the local computer, and not the remote files. If it becomes necessary to save the data onto the server, the file can be dragged from the local file manger to the remote one.

For every analysis run through Jemboss, all the results are saved on the server. So the results of previous analyses can be accessed by selecting "Saved Results" from the "File" menu. The results list contains the name of the program, and the date and time at which it was run, together with the parameters of the analysis. Results may be displayed, and saved locally. Files may only be deleted manually by the user, thus the server results provide a back up system of data for the user.

1. Tim J Carver, Lisa J Mullan, Comparative and Functional Genomics, "A new graphical user interface to EMBOSS" vol. 3, 1, pp. 75-78, 2002.

2. Lisa Mullan and Tim Carver, Biochemist, "Point, click and analyse!", vol. 24, No. 2, p. 39, 2002.

EMBnet Node News

BEN - The Belgian EMBnet Node

Personnel shifts

During last year, the system management at BEN was within the hands of computer engineer Vincent Ndi Owona. His position was taken over in November 2001 by David Coornaert, Ph.D. in biology and self trained bioinformatician. Today, in addition of David, BEN's personnel consists of Guy Bottu, Ph. D. in chemistry, in charge of general software management and user support and Valérie Ledent, Ph. D. in biology, in charge of user support and BEN's courses and training sessions.

New mission

The OSTC, our Belgian Federal Office for Scientific, Technical and Cultural Affairs, the financial support of the Belgian EMBNet Node, asked us to run over a one year period a prototype action in the field of Biodiversity. This is part of the efforts of the Belgian authorities to meet the goals of GBIF, the Global Biodiversity Information Facility, a worldwide initiative to make biodiversity information easily available and interoperable. To reach this goal, BEN's team was complemented by the so-called "BeBIF team", presently consisting of Patricia Mergen Ph. D. in Biology, and the informaticians Marc Dubrowski and Johan Duflost. See the web site bebif.ulb.ac.be for further descriptions and the current status of the project. BEN's intention in this field is to implement the SRS technology (Lion Bioscience) to try and integrate available biodiversity data, together with Zope technology for general website management. A first prototype about the fish populations in artificial water reservoirs was built at BeBIF, and the next efforts will concentrate on various national data sources (wildlife in Flanders and the Walloon region, the North sea), as well as data from our museum of Central Africa and data brought back from the recent European missions to the Antarctic regions.

Interface developments

BEN efforts in the field of software interfaces date from many years in the past. Back in the times of command line interface, Marc Colet built a ***menu*** system on top of GCG. Later, he was among the first to provide a web interface to GCG ***WWW2GCG***. In the present days where EMBOSS is taking over most sequence analysis tasks, BEN provides the Jemboss interface to EMBOSS, but Marc is working in cooperation with Luke Mc Carthy from the Canadian EMBnet node, author of ***EMBOSS GUI,*** to provide a Web browser-oriented interface to EMBOSS, devoid of the burdens associated with the Java implementation of Jemboss. This project is reaching maturity and allows registered BEN users to access all classical EMBOSS applications and store/manage their biocomputing projects, databanks and sequences directly on the UNIX server. Ease of use and absence of any specific installation are the keywords of this user interface. It will soon be available form BEN's ftp server (ftp.be.embnet.org)

Staden Package

James Bonfield
Medical Research Council
Laboratory of Molecular Biology
Hills Road, Cambridge, CB2 2QH, UK.
jkb@mrc-lmb.cam.ac.uk

The Staden Package covers two main areas. The most widely known is the DNA sequence assembly tools ("Pregap4" and "Gap4"). However the origins of the package are with nucleotide and amino-acid sequence analysis. These functions are now covered by the (much updated since then) "Spin" program.

Availability

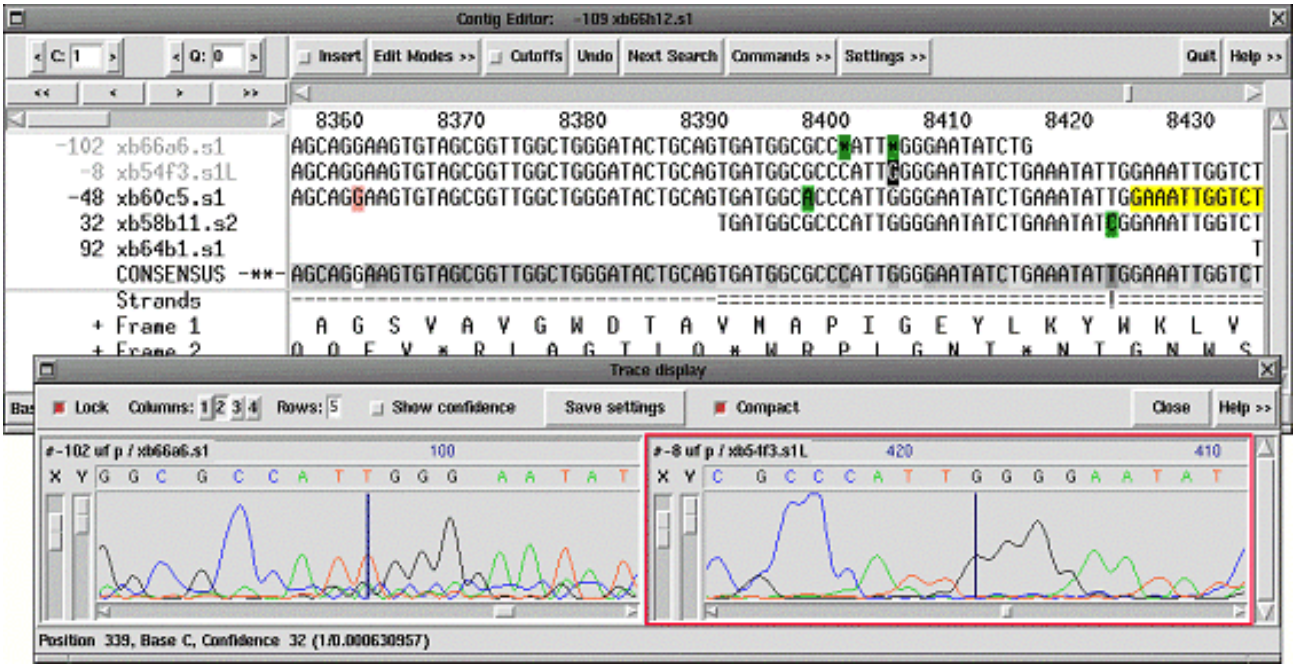
The complete package is available for Unix (Sun Solaris, Linux (on intel systems), Compaq Tru64 Unix (previously known as Digital Unix), SGI Irix), Microsoft Windows (all win32 systems, but best run on NT, 2000 or XP), and is in testing for Apple MacOS X.

To obtain the package follow the links from the package web site at http://www.mrc-lmb.cam.ac.uk/pubseq/downloads.html. Once installed (which is simply a matter of unpacking the archive or running the self-extracting installer) you will need to obtain a licence key to unlock the software. This is freely available for academic or not-for-profit users. Requesting a licence is easy - simply run one of the programs (eg Gap4) and cut and paste the host information into the web form at http://www.mrc-lmb.cam.ac.uk/pubseq/licence.html. The Staden Package team will then email you a licence file with instructions on where to save it.

Full course data and course notes are included with the package. This is the same two day Gap4 course run by David Judge and James Bonfield.

Sequence Assembly

Handling sequence assembly projects is split into two programs. The first of these, Pregap4, deals with the automatic processing that needs to be applied to DNA "trace" files. This includes quality assessment, sequencing vector identification, screening for contamination, marking known repeat sequences and the sequence assembly itself. Pregap4's philosophy is to allow integration with third party external tools (such as the Phred



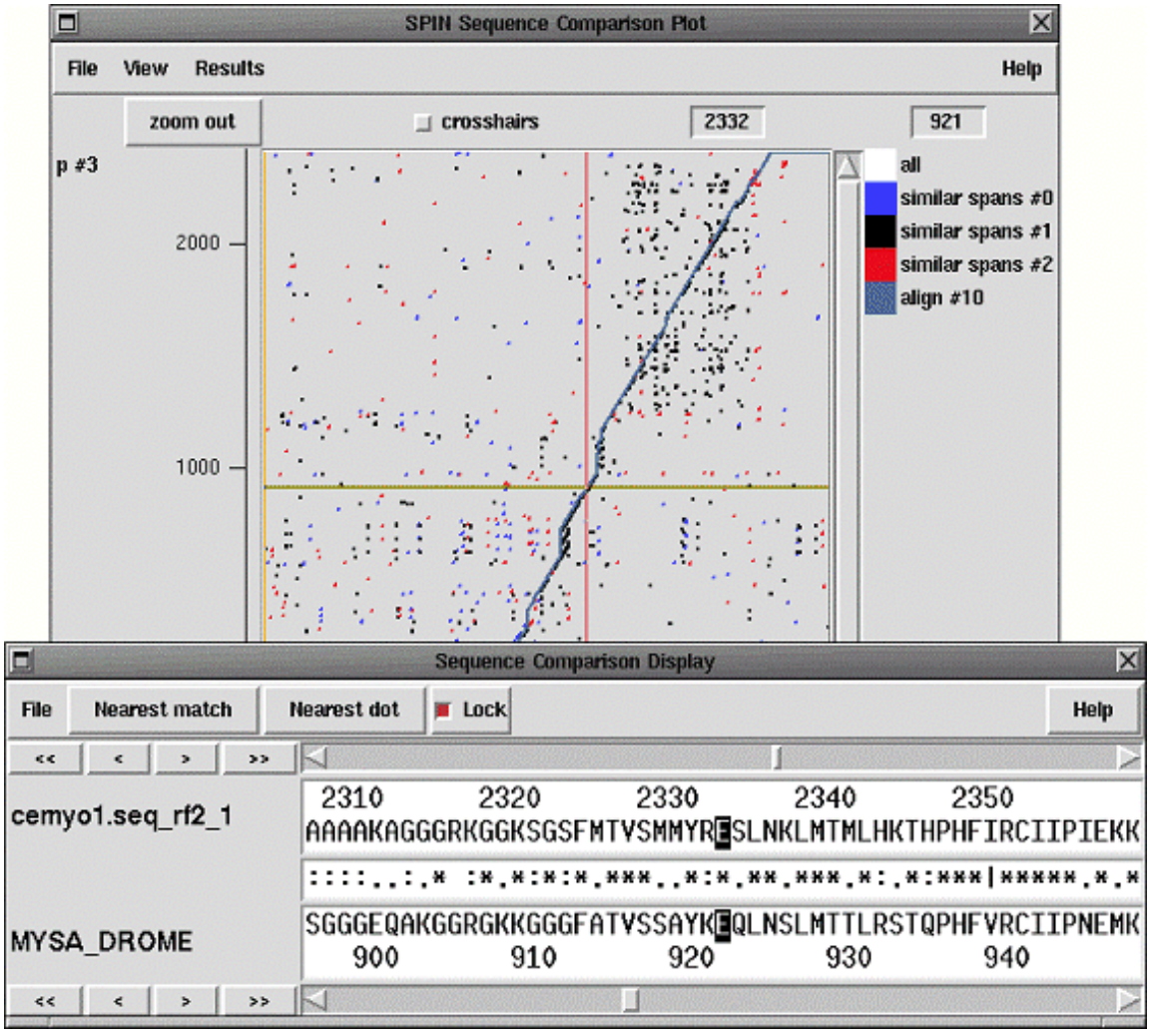
The Contig Editor window showing the aligned sequences, consensus bases with confidence, and two trace files.

base caller and the Phrap, Cap or Fak assembly algorithms) so that the user may use the best parts of each software suite.
The primary output of Pregap4 is an assembled sequencing project, but rarely does this give the final answer without lots of additional experiments and editing. This is where the Gap4 program is used. Gap4 is a large program containing many graphical overviews of the assembly project (reading positions, template coverage, read-pair coverage, consensus quality, strand conflicts) along with the textual contig editor in which most of the user's time is likely to be spent.

Sequence Analysis

Analysis is covered by the "Spin" program. Spin's main strength is in its graphical displays, with may be zoomed, scrolled, overlayed with other plots and by use of common x-y cursors may be used to control other plots or sequence displays.

Spin also includes a variety of DNA analysis tools including restriction enzymes (presented in both textual and graphical displays) and several types of codon usage plots. Spin's greatest drawback is the lack of analysis algorithms, in particular protein sequence analysis. However Spin also includes an interface to the EMBOSS suite of tools which for Spin users provides many more algorithms and for EMBOSS users provides an improved graphical interface.



A Spin dot-plot showing matches between a protein and a three-phase translation of a nucleotide sequence. Superimposed on these dot-plots is a graphical view of a sequence alignment. The top-most window is the pair-wise sequence display showing a detailed sequence comparison.

Transcript reconstruction and analysis using STACKdb and stackPACK

Janet Kelso
South African National Bioinformatics Institute
University of the Western Cape
Modderdam Road
Bellville 7535
janet@sanbi.ac.za

Summary

StackPACK is an integrated suite of transcript reconstruction and analysis tools for the production of gene indices from transcript data. The toolset is generic in that it can be used to cluster transcript data from any organism

STACKdb is a human gene index built from more than 3 million publically available human ESTs and mRNAs using the stackPACK toolset.

Introduction

STACKdb and stackPACK were developed at the South African National Bioinformatics Institute as part of an academic research project into using and managing EST and mRNA data to understand gene expression. In 1999 Electric Genetics, a spin-off company, undertook long-term commercial support and development of the STACK project and subsequently re-engineered the academic software to produce a stable and integrated software suite which uses a CORBA layer and a database back-end.

STACKdb is a human gene index reconstructed from EST and mRNA data. stackPACK is the software toolset used to process the transcript data to produce the index. Both the clustering method [1] and the database [2] have been published in peer-reviewed journals.

The aims of the project at its outset were:

- To use the large quantities of publically available EST data to produce a high quality gene index. A gene index is a set of all gene products. Expressed sequence tags can be grouped together based on sequence similarity and annotation information to produce a single cluster per expressed gene.
- To do this in such a way as to maximise the production of long consensus sequences which cover as much of the gene as possible.
- To capture as many transcript isoforms of each gene as possible
- To produce tools for clustering, viewing and extraction of data
- To understand gene expression profiles (ie: in which tissues, developmental stages and diseases specific genes and gene isoforms are expressed)

stackPACK

stackPACK is a generic set of tools for clustering and analyzing transcript data. Data is processed through a pipeline which is made up of different proprietary and external software designed to cluster and align transcripts belonging to a common gene, to produce a consensus sequence for each cluster and to detect instances of potential alternative splicing. A relational database is used to store the input and output of each stage of processing.

The stackPACK system can be accessed using a single-step web-based interface, or more complex analyses can be performed via the command-line interface. The command-line is more modular, providing users with the ability to modify parameters, use custom-made masking files and to undo steps in the pipeline.

Key features which make this system unique are its ability to:

- Maximise consensus length
- Identify potential transcript variation

The next release of stackPACK (v2.2) will include:

- the ability to add sequences to existing projects
- the use of phred quality scores
- additional data extraction reports
- further speed enhancements

STACKdb

STACKdb is a gene index built from more than 3 million publically available human ESTs and mRNAs using the stackPACK toolset.

STACKdb differs from other gene indices in that it organises ESTs by their tissue of origin, and making an index for each tissue, thus providing an index (a list of the genes expressed) for each tissue. A total, or whole body index, is created by merging the individual tissue indices to make a index of all the expressed genes.

STACKdb is available in two formats – as mySQL database dumps which can be loaded into local mySQL databases, or as non-redundant flat files for rapid searching. All necessary database and visualisation tools are provided and users therefore have access to all stackPACK functions, except the actual processing pipeline.

Users can query the latest release of STACKdb via the web-based BLAST search engine at: `http://juju.egenetics.com/stackpack/webblast.html`

Key features which make STACKdb unique are:

- ESTs and mRNAs are organized by tissue & disease state of origin
- Inclusive consensus sequences which are linked based on clone ID information
- Has built in visualization tools
- Is stored in a relational database
- Provides list of genes which may demonstrate alternative splicing

Future STACKdb releases will incorporate all mRNAs and will use the fully sequenced mRNA's as a basis upon which to build the fully 'whole gene index'. In addition, current work at SANBI using controlled gene expression vocabularies will allow for STACK full body index queries that can extract transcript-sets based upon any combination of expression terms. Mapping of STACK indices to the genome is currently underway.

Userbase

stackPACK and STACKdb are used by more than 150 academic and commercial groups worldwide including the Wellcome Trust, Max Planck Institute, Sanger Centre, NIH and Affymetrix. For a full listing of stackPACK and STACKdb users see: `http://www.egenetics.com/?Section=users&Parent=products`

Availability

The software is available free of charge to researchers at non-profit institutions using it for non-profit purposes. Download stackPACK and / or STACKdb from `http://www.sanbi.ac.za/CODES/`

A full user manual and installation instructions are included in the download. If you need any further help with installing or using the software you can contact support@egenetics.com

STACKdb is available free to academics via download from the kind co-operation of the Swiss Institute of Bioinformatics (Europe). Download requests made to SANBI are automatically redirected to SIB's ftp site in order to improve download speeds. STACKdb and STACKpack can be served from EMBnet nodes upon request.

Reference List

1. A. Christoffels et al., *Nucleic Acids Res.* 29, 234-238 (2001).

2. R. T. Miller et al., *Genome Res.* 9, 1143-1155 (1999).

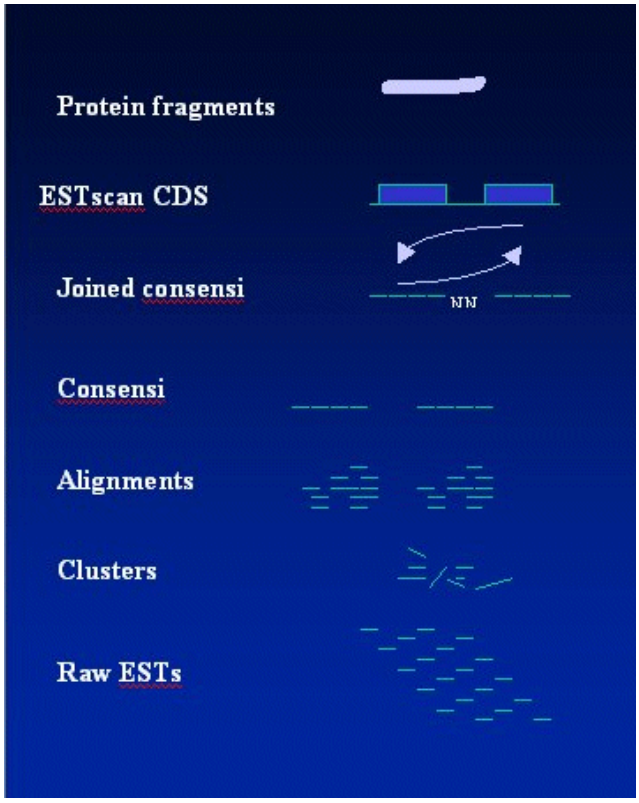


Figure 1. ***Transcript reconstruction and analysis: Raw ESTs aregrouped into clusters based on sequence similarity, clustered transcripts are then aligned and a consensus sequence is built from the alignment. If clone ID information is available this can be used to link consensi which share ESTs from opposite ends of the same clone. Further processing can then be performed – detection of the CDS and prediction of virtual proteins from consensus sequences can be useful in determining the identity and function of previously uncharacterised genes.***

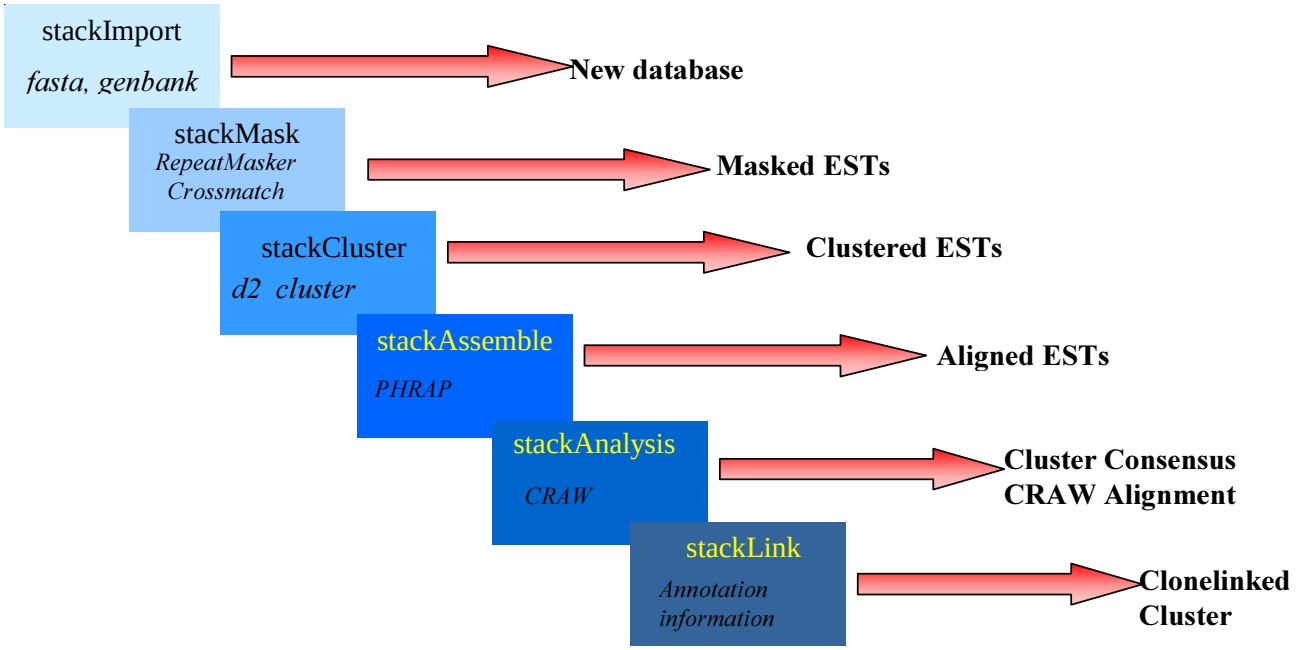
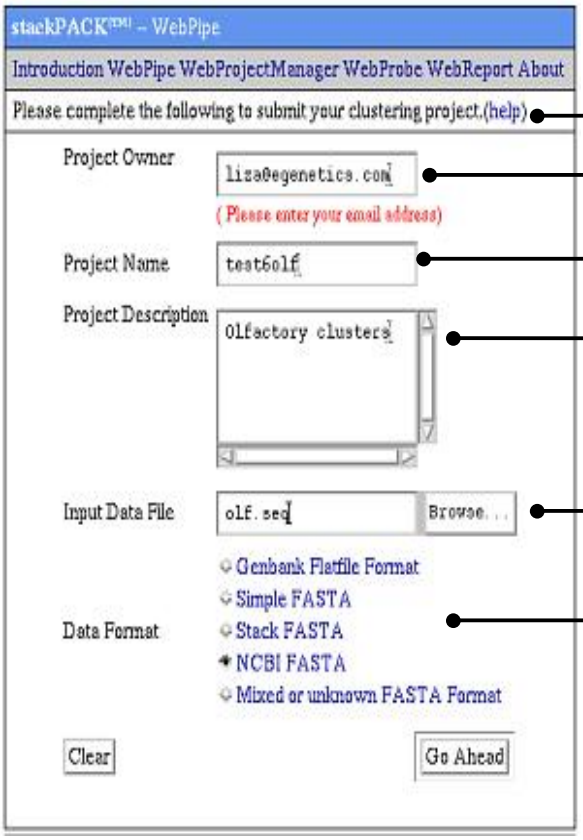


Figure 2. ***The stackPACK data processing pipeline. Raw transcripts are subjected to consecutive processing by different algorithms in order to maximize cluster size and consensus length while maintaining quality. Using the command-line each step can be run separately. Using the web-interface the user can submit data for processing through all six stages in a single click.***



Context sensitive help.

Project owner is emailed when clustering completes.

Project name

Detailed description of project for your reference.

Browse for input data file on local or networked computers.

Supports commonly used sequence file formats, attempting to parse as much information as possible.

Figure 3. *The stackPACK web interface. Users provide project information, sequence data and specify the sequence format. Email is sent to the user upon completion of Context sensitive help.*

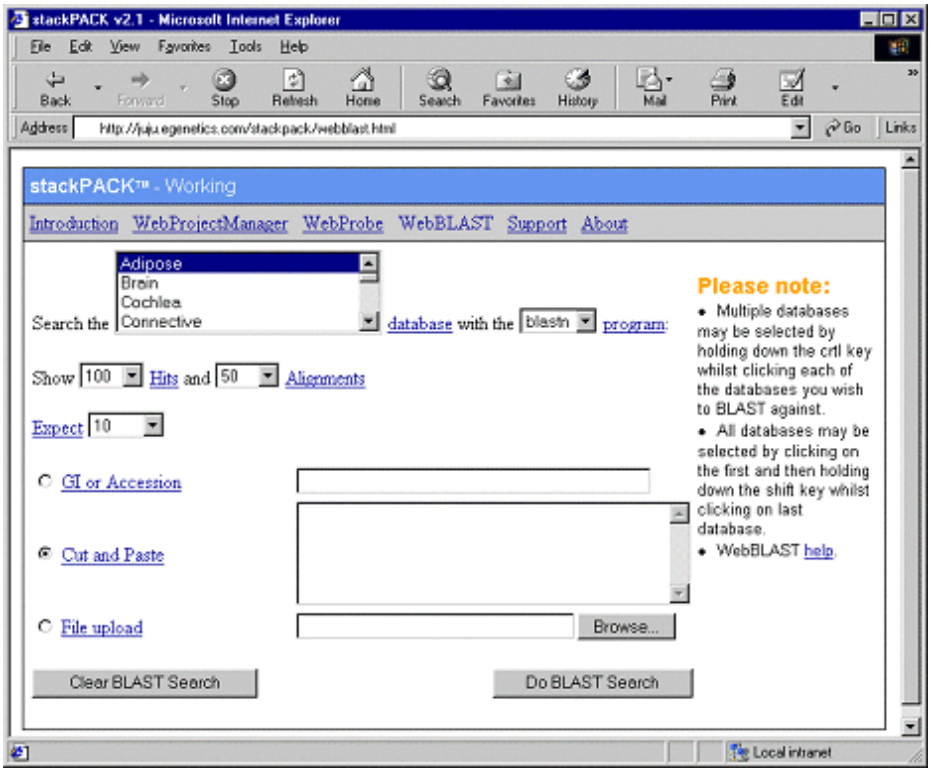


Figure 4. ***STACK Web-BLAST***

The latest release of STACKdb is available for online homology searching at http://juju.egenetics.com/stackpack/webblast.html. Searching STACKdb will provide researchers with a list of the matching transcripts, cluster information and the tissue distribution of the matching sequences.

Open Source code

Johann Visagie
Electric Genetics
PO Box 3912
Cape Town 8000
South Africa
johann@egenetics.com

Introductory remarks

Basic tenets of open source

The basic precept of open source software – namely that it is software that is distributed *with* its source code – is widely understood. That alone only tells a part of the story, though. This article will look at some of the issues surrounding open source, and specifically how open source relates to Bioinformatics.

"Open Source" and "Free Software"

It should be noted that there has long been a difference of opinion as to whether "open source" or "free software" is the more correct term to describe what they layman would regard as essentially the same concept. This article will not stray into this argument, and will admittedly use the term "open source" very loosely.

History of open source

Early Internet

To some extent, the history of open source is intertwined with the early history of the Internet. The technical standards that form the basis of this network were developed in a collaborative fashion and soon spawned a community to which the ideals of openness and cooperation were of paramount importance. Years before explicit definitions were attached to "open source" or "free software", this community was in the habit of sharing large swathes of source code – often reference implementations for some or other new protocols that was under discussion. "Intellectual property" was not an issue to these researchers; the widest possible dissemination of the code was. The concepts of *community* and *collaboration* remain central to open source today.

Richard Stallman

It would be difficult to overstate the role of Richard M Stallman played in the history of free software. A denizen of MIT's famed AI Lab during the 1970's, Stallman became disheartened by the demise of his beloved hacker culture when commercial software companies started poaching talent from the AI Lab in the early 80's.

Stallman's solution was to start writing his own free UNIX-like operating system, GNU (an acronym for "GNU's Not UNIX"). To counter the restrictive licences imposed on users by commercial vendors, Stallman authored the GNU General Public Licence (GPL) – a licence which uses the concept of copyright to ensure the freedom of a software product in perpetuity.

Stallman's work led to the formation of the Free Software Foundation (FSF), a jealous guardian of the concept of software freedom which operates from premises on the MIT campus to this day.
http://www.fsf.org/

Berkeley's CSRG

Strangely, the role of the Computer Science Research Group (CSRG) at the University of California's Berkeley campus in the development of open source software is often overlooked. To appreciate this role, one has to look back in history to several chains of events during the early 70's:

Firstly, there was the release by AT&T Bell Labs of their in-house research operating system, UNIX. Since AT&T was not at the time in the software business, UNIX was made available to universities at nominal cost and soon formed the basis of numerous research-oriented operating systems in academia.

Secondly, the US Advanced Research Projects Agency (ARPA) – and later its successor, DARPA – was making swift progress in laying the foundation of what would come to be known as the Internet.

In the late 70's DARPA felt the need to standardise the implementations of their various network protocols on a single operating system. Eschewing commercial alternatives, they elected

to provide funding UC Berkeley to integrate these protocols into their increasingly popular UNIX implementation, BSD. And so, for the first time, the Internet's TCP/IP protocol suite was married to UNIX. Moreover, since BSD was distributed under a licence with virtually no strings attached, both UNIX and the Internet protocols gained an ever-stronger foothold in the academic world. For the first time, open source hackers had their own platform to work on, and their own network through which to coordinate their efforts. http://www.freebsd.org/

Linus Torvalds and the rise of Linux

The story of a young Finnish student who decided to write his own operating system in order to learn more about his new Intel 386-based PC system has become well known. When Linus Torvalds released the first version of the fledgling operating system he called Linux in 1991, nobody – least of all he – could foresee the sort of underground revolution it would incite in the computer industry.

But the timing was right. With AT&T (which finally realised the value of UNIX) and the University of California locked in a legal wrangle, Linux had the freedom to build up enough of a following to become a *de facto* standard – a free operating system for the masses and a widely available platform on which to develop open source software.
More than any other single product, Linux has been responsible for the mainstream nature of the open source movement today. http://www.linux.org/

The Open Source Initiative

During the latter part of the 1990's, the Open Source Initiative (OSI) was founded to promote a more "business-friendly" image of open source – something its advocates found lacking in the FSF. The founding of the OSI is based at least partially on the writings of Eric Raymond, whose paper *The Cathedral and the Bazaar* provided some form of philosophical framework wherein discussion of the principles of open source can take place.
http://www.opensource.org/
http://tuxedo.org/~esr/writings/cathedral-bazaar/

What is open source?

Open source is a development methodology

Over the years, mainstream open source projects have started to exhibit a characteristic development methodology.

- Open source projects tend to rely on a common, free development toolset. They are most often developed on a free operating system like a Linux distribution or one of the BSDs. They tend to use free language implementations like the GNU C++ compiler, gcc, the Perl language or the Python language. They tend to use free development and build tools. In short, they rely on a toolset that ensures that there is no barrier to participation in the project.
- Open source projects tend to be licensed under one of a number of existing, recognised open source licenses.
- Open source projects tend to be entirely transparent even down to the way in which they share source code. Source code is often committed to a version controlled repository where it is widely available for public scrutiny… and participation.
- Open source projects focus strongly on community building via web sites, mailing lists and the numerous other channels provided by the Internet today.
- Open source projects tend to make releases, to paraphrase Eric Raymond, "early and often".
- Open source projects tend not to reinvent the wheel, but rather to cooperate with or incorporate other open source projects that provide needed functionality.

Open source is a licence to be free

Licensing issues are often a point of conflict amongst open source devotees, and their importance are almost certainly overstated by some sectors of the community. Still, in an increasingly litiginous society it would be equally unwise to *under*estimate their importance. Specifically, open source developers should be aware of the implications of various open source licences to the point where he can make an informed decision when it comes to licensing his

own software. Briefly, two of the most prominent groups of competing licences are:

- The Free Software Foundation's General Public Licence and Lesser General Public Licence (LGPL). These licences are long and rather complex, and this very complexity has in the past led to various "compatibility" issues. These licences enshrine the freedom of the software above all else by allowing redistribution only if the source code is also redistributed.
- The Berkeley (BSD), MIT (X), or Apache-style licences tend towards extreme simplicity by contrast. They *do* allow redistribution without source code, and therefore can accommodate commercial exploitation of open source software.

Open source is a community

Perhaps the single most important characteristic of open source development remains the fact that it is done in a communal fashion, often by close-knit communities, for the public benefit.
Open source in Bioinformatics

Why Bioinformatics is different

Most existing open source projects are aimed at providing the very tools required by the technical people who tend to participate in them – operating systems, network services, and the like. However, Bioinformatics (and the many other scientific disciplines that are waking up to the advantages of open source) is different in the intended end users often from a slightly different demographic from the developers.

Why open source is a good fit

- Open source software tends to have a much faster response or turnaround time than commercial software, since it is written by the very people who need the software "in the field". By contrast, commercial software typically has an 18 month product cycle.
- Open source development mirrors the ideal of collaboration in scientific research: Knowledge is shared freely, and everyone is free to build on the foundations laid by others.
- Due to the vast quantities of data under consideration, Bioinformatics research takes place almost exclusively on UNIX platforms. Open source too is still essentially a UNIX-bound endeavour.

The world of open source Bioinformatics

Free tools from academic groups

Over the years, many academic groups have been providing software under a variety of licenses, sometimes partially free (in one sense or another). Unfortunately, it appears that many of these academic groups were unaware of (or unconcerned with) developments in the broader open source movement, resulting in a large body widely used of software with often very murky licensing conditions, giving rise to the possibility of trapping a user in a veritable legal quagmire.

The Open Bioinformatics Foundation (O|B|F)

The Open Bioinformatics Foundation is a not-for-profit organisation that was formed as an umbrella group for the various open source Bioinformatics projects that were developed by the extended community that was originally responsible for Bioperl.
The O|B|F is mainly performs a support role. As such, it provides computing infrastructure to these open source projects, and organises events such as the annual Bioinformatics Open Source Conference (BOSC).
For anyone not familiar with current developments in open source Bioinformatics, the O|B|F website provides the perfect springboard.
http://www.open-bio.org/

Open source Bioinformatics projects

The Bioperl project was one of the earliest coordinated open source initiatives in Bioinformatics. It arose out of the perceived need for a basic Bioinformatics toolset for Perl programmers. (The Perl language has long been popular in the field due to its flexible and powerful text processing features.) Today, Bioperl is regarded as stable enough to merit a "1.0" version number.

It's also noteworthy that the well-known Ensembl automatic annotation system (itself an open source project) is largely built on Bioperl. The Bioperl

project inspired users of other popular programming languages to develop analogous toolsets and today we also have – in addition to Bioperl – projects such as BioJava, Biopython and BioRuby.

Fortunately, the groups responsible for these projects have consistently "stuck together" and today we find that these projects largely overlap in terms of design philosophy. Moreover, new open source initiatives (such as BioSQL and the DAS protocol) which provide shared infrastructure for all these toolsets have been launched and supported communally by member projects of the O|B|F. The success (and wide acceptance) of these projects has resulted in other projects joining this broad community, notably the Omnigene suite.

The projects directly associated with the O|B|F form a core community of developers, but this community is in no way exclusive. There are a number of other noteworthy open source initiatives in Bioinformatics – some large; some small. Many of the smaller ones have found the free project hosting services offered by Bioinformatics.Org to be of great use.
http://www.bioperl.org/
http://www.biojava.org/
http://www.bioruby.org/
http://www.ensembl.org/
http://omnigene.sourceforge.net/
http://bioinformaics.org/

EMBER - A European Multimedia Bioinformatics Educational Resource

T.K.Attwood
(On behalf of the EMBER consortium)
School of Biological Sciences
and Department of Computer Science
University of Manchester
Oxford Road
Manchester M13 9PT
attwood@bioinf.man.ac.uk

Introduction

Bioinformatics has taken centre stage in the post-genome era. The data over-load arising from the many now-fruitful genome projects has created an insatiable demand for suitably qualified people to build and maintain databases, to design more incisive analysis software, to use disparate databases and software tools, and to understand both the statistical and biological significance of results generated *in silico*. It is rare to find individuals with such a range of skills, yet such scientists are now needed urgently in sequencing centres, research/academic institutes, pharmaceutical/agrochemical companies, software houses and start-up companies. But the rate of growth of the field, and its cross-disciplinary nature, has created a problem: while there are many trained biologists and computer scientists, there are few computer-literate biologists or biology-literate computer scientists. Consequently, there is a dearth of skilled staff in bioinformatics. This is especially problematic for universities, which are less able than large multinational companies to compete for the small numbers of trained individuals emerging, for example, from MSc or MRes courses.

As a step towards addressing some of these issues, the European Commission has recently funded an innovative new educational project (EMBER) that strives both to bridge the current skills gap and to foster academic and private-sector collaboration in the development of new standardised, multimedia courses in bioinformatics.

The Concept and the Consortium

EMBER will comprise a suite of multimedia bioinformatics educational tools, including a self-contained, interactive Web-tutorial in bioinformatics, the equivalent stand-alone course on CD-ROM, and an accompanying introductory text-book. The use of conventional text, coupled with Web- and CD-based media, will ensure that students in European domains for whom Internet access is not optimal also have access to the same fundamental level of bioinformatics education. The project, which is being coordinated at the University of Manchester, involves a world-

Table 1. The EMBER Consortium

Developers	
UMan	The University of Manchester, UK
SIB	The Swiss Institute of Bioinformatics, Switzerland
SANBI	University of the Western Cape, South Africa
IMB	National Research Council of Canada Institute for Marine Biosciences
ETI	Expert Centre for Taxonomic Identification, The Netherlands*
Evaluators	
SIB	The Swiss Institute of Bioinformatics, Switzerland
KUN	The University of Nijmegen, The Netherlands
EBI	The European Bioinformatics Institute, UK
IGC	Institute Gulbenkian de Ciencia, Portugal
BEN	The University of Bruxelles, Belgium
RIGEB	The Research Institute for Genetic Engineering and Biotechnology, Turkey
IMB	National Research Council of Canada Institute for Marine Biosciences

* ETI are technical developers rather than educational content developers

wide consortium of bioinformatics organisations, as shown in Table 1.

Some of these participants will help to develop the new teaching materials (the developers in Table 1), while others will test them through user trials conducted across Europe (the evaluators). Collaboration with a professional multimedia publisher in the Netherlands will ensure production of the Web- and CD-ROM-based materials to the highest standards.

The Project

Where possible, EMBER will build on teaching materials currently used by the consortium partners. At Manchester, part of our MSc teaching hinges on the use of an interactive bioinformatics practical, termed BioActivity (see Figure 1) [1-3], which is supported by an introductory text-book [4].

BioActivity is now used world-wide and must be revised and updated constantly, as embedded Web links change and the field moves on; hence its maintenance is both cumbersome and time-consuming. Moreover, the book was written 3 years ago, and not only are its contents in urgent need of update, but they must also be extended in scope. Therefore, EMBER will both revise and add new dimensions to these materials by drawing on the complementary research expertise and

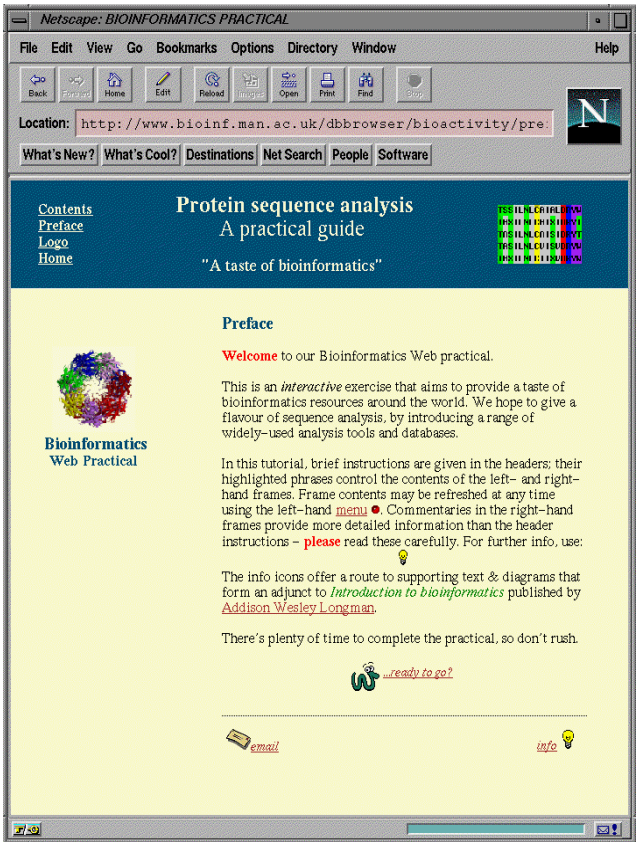


Figure 1. ***BioActivity homepage (http://www.bioinf.man.ac.uk/dbbrowser/bioactivity). BioActivity is an interactive Web tutorial in bioinformatics, used as the basis for the Manchester MSc and other bioinformatics courses around the world.***

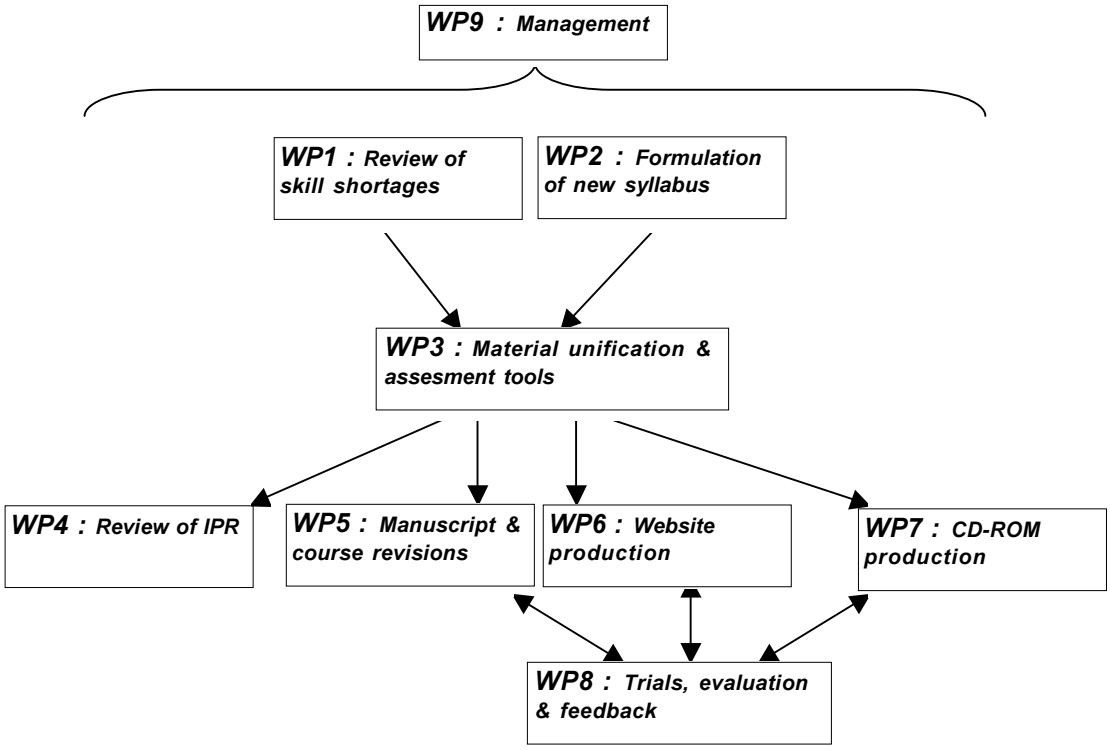


Figure 2. ***Overview of EMBER workpackages (WP) and their inter-relationships. The project is reaching the end of workpackage 3, in preparation for hand-over to ETI for the professional Web-site and CD-ROM production. While this is ongoing, chapters of the book will be written and courses planned to trial the new materials prior to mass production and marketing.***

teaching experience of the consortium partners. Accordingly, the initial phases of the project concern the development of a new curriculum, which will guide the design of the 'new-look' Web- and CD-ROM-based courses and textbook. The courses will be tested by conducting trials across Europe, feedback from which will allow us to make any necessary improvements. The relationship between the project components, or workpackages, is illustrated in Figure 2.

Progress

EMBER began in May 2001. Prior to its formal kick-off, a special meeting of the consortium was held in Lisbon to determine the project strategy. During these discussions, a tentative book outline was agreed, drawing heavily on the course content of the SIB MSc. In the following months, a survey of bioinformatics skill requirements was conducted, by sending out a simple questionnaire to potential employers in a range of academic and industrial organisations. Although feedback from the questionnaire was poor (only 16% of contacts replied), the responses we did get were nevertheless useful, largely supporting the course and book outlines agreed in Lisbon [5].

The principal areas to be covered in the book are illustrated in Table 2. These areas fall into discrete themes – *i.e.*, core and advanced bioinformatics, and supplementary topics. Within the basic scheme shown in Table 2, core topics are those that tend to dominate day-to-day analytical and experimental bioinformatics approaches; advanced topics are those requiring, perhaps, a greater degree of analytical and/or mathematical experience; and supplementary topics are those that, while desirable, are regarded as being more challenging to include within a basic bioinformatics course. Following several months of detailed discussions on the nature of the subjects to be covered within each of these themes, a book proposal was sent to, and has been accepted by, Wiley.

Table 2. Main themes to be covered in the EMBER text-book.

Core Bioinformatics

- Biological databases (*e.g.* sequence and family databases, database technologies)
- Principles of sequence analysis (*e.g.* pairwise sequence analysis, scoring matrices)
- Protein structure (*e.g.* structure classification databases, visualisation)
- The genome (*e.g.* gene prediction, genome annotation, technology platforms)
- The transcriptome (*e.g.* EST data, EST clustering and assembly, microarrays)
- The proteome (*e.g.* 2D gel data, mass spectrometry data, image analysis)

Advanced Bioinformatics

- Molecular evolution and phylogeny (*e.g.* biological foundations, terminology, methodologies)
- Ontologies in bioinformatics (*e.g.* Gene ontology, EcoCyc)
- Principles of protein structure prediction (*e.g.* homology modelling, threading)

Supplementary Topics

- Information theory
- Basic statistics

We are currently busy collating and unifying new and old materials, in readiness to hand over to our electronic publishing colleagues at ETI within the next few weeks. At the same time, we are beginning to develop suitable assessment tools, both to evaluate the product (*e.g.*, look-and-feel, ease of navigation, *etc.*) and to evaluate students' performance on the course (*e.g.*, level of knowledge before and after assessment). During the next critical months, while ETI are generating the Web- & CD-ROM-based materials to a professional standard, chapters of the book will begin to be written and courses to test the materials will be planned. The timeframes and deliverables from each of the project components are illustrated in Figure 3. Meanwhile, we have established a Web-site (http://www.bioinf.man.ac.uk/ember) from which all documents and project progress can be viewed.

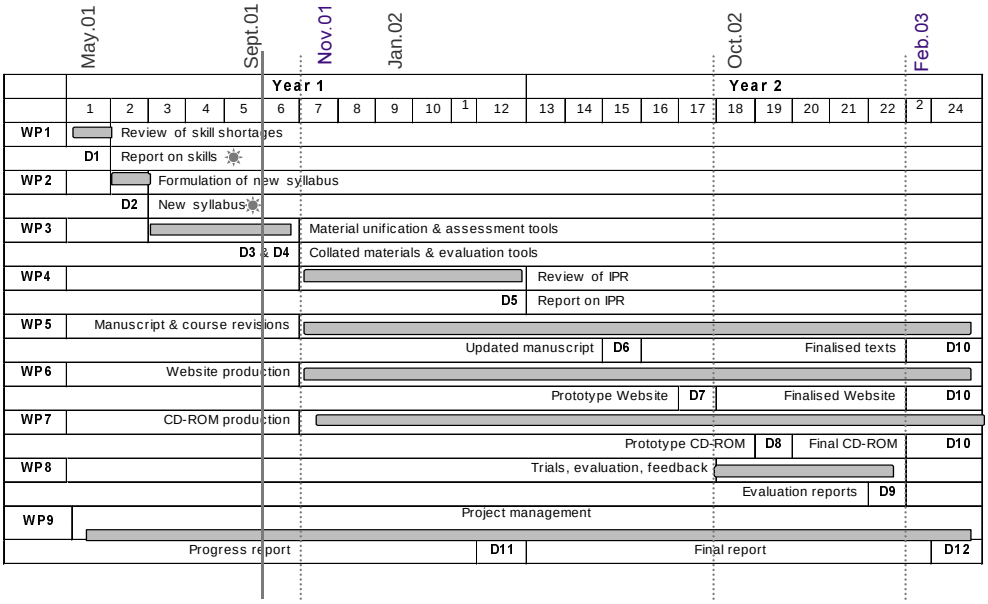


Figure 3. ***Overview of EMBER timeframes and deliverables. The project is in the final month of workpackages 6 and 7. Draft chapters have been produced and an edited manuscript is ready to offer to the publishers. Trials spanning a 5-month window will commence in October this year.***

Conclusion

Although only in the early stages of a two-year EC-funded project, by the end of this period, our goal is that EMBER should provide a standard, self-contained bioinformatics course, suitable for conventional face-to-face delivery or for self-paced tuition, whether at home or within the workplace. Ultimately, we hope that EMBER will help to build strength in bioinformatics by facilitating the transfer of core skills across Europe and making it easier for employers to assess the abilities of future job applicants.

References

1. Attwood, T.K. (2000) An interactive practical at the interface of Web-based and conventional publishing. *CAL-laborate. A collaborative publication on the use of Computer Aided Learning for tertiary level physical sciences*, Vol. 4, pp.1-6.
2. Attwood, T.K. (1999) An interactive bioinformatics practical on the World Wide Web. In *Laboratory Work in Science Education: an Evaluation with Case studies*, (Ed.L.Jervis), pp.42-45. SEED Publications, University of Plymouth.
3. Attwood, T.K. (1999) Introduction to bioinformatics: an interactive practical on the WWW. *Life Sciences Educational Computing*, Vol. 10(2), p.12.
4. Attwood, T.K. and Parry-Smith, D.J. (1999) *Introduction to bioinformatics*. Addison Wesley Longman, Harlow, Essex CM20 2JE, UK. ISBN 0 582 327881
5. Mabey, J.E. and Attwood, T.K. (2001) EMBER: a European Multimedia Bioinformatics Educational Resource. *CAL-laborate. A collaborative publication on the use of Computer Aided Learning for tertiary level physical sciences*, Vol. 6**,** pp.13-16.

Announcement

UPPSALA UNIVERSITET

In connection with the program Human Research Potential & the Socio-economic Knowledge Base : Access to Research Infrastructures (ARI) initiated and funded by the European Commission, the Linnaeus Centre for Bioinformatics (LCB), a recently established highly active research centre located at the Biomedical Centre (BMC), Uppsala, Sweden, is inviting for applications from research teams working in bioinformatics to gain access to its unique research infrastructure. For details, on eligibility, please see http://www.cordis.lu/improving/infrastructure/access.htm.
Access can be granted, in general, for twenty working days and can start any time.

Applications must be written in English and contain a short proposal (approx. 500-1,000 words) describing the project to be carried out at the LCB, a short curriculum vitae (approx. 1 page) and a publication list for each researcher expected to take part in the project, and the expected duration of the stay. They must be sent to Dr Katharina Huber, Linnaeus Centre for Bioinformatics, BMC, Box 598, 751 24 Uppsala, Sweden.

Deadlines for the applications are : September 15th, 2002 (first deadline), March 15th, 2003 (second deadline), and October 15th, 2003 (third deadline).

The EMBnet Nodes

National Nodes

Argentia
Oscar Grau
IBBM, Facultad de Cs. Exactas, Universidad Nacional de La Plata
Email: grau@biol.unlp.edu.ar
Tel: +54-221-4259223 Fax: +54-221-4259223
URL: http://sol.biol.unlp.edu.ar/embnet

Australia
Sonia Cattley
RMC Gunn Building B19, University of Sydney,NSW, 2006
Email: scattley@angis.org.au
Tel: +61-2-9531 2948
URL: http://www.angis.org.au

Austria
Martin Grabner
Vienna Bio Center, University of Vienna
Email: martin.grabner@univie.ac.at
Tel: +43-1-4277/14141
URL: http://www.at.embnet.org

Belgium
Robert Herzog, Marc Colet
BEN ULB Campus Plaine CP 257
Email: rherzog@ulb.ac.be mcolet@ulb.ac.be
Tel: +32 2 6505146 Fax +32 2 6505124
URL: http://www.be.embnet.org

Canada
Simon Mercer
Canadian Bioinformatics Resource, National Research Council Canada, Institute for Marine Biosciences,
Email: Simon.Mercer@nrc.ca
Tel: +1-902-426 7310 Fax: +1-902-426 9413
URL: http://www.cbr.nrc.ca

Chile
Dr. Ricardo Baeza-Yates
Dept. of Computer Science, Santiago,
Email: rbaeza@dcc.uchile.cl
Tel: N/A
URL: http://www.embnet.cl

China
Jingchu Luo
Room 303, Exchange Centre, Peking University
Email: luojc@cbi.pku.edu.cn
Tel: +86-10-6275 9001
URL: http://www.cbi.pku.edu.cn

Cuba
Ricardo Bringas
Centro de Ingeniería Genética y Biotecnolgía, La Habana,
Email: bringas@cigb.edu.cu
Tel: +53 7 218200
URL: http://bio.cigb.edu.cu

Denmark
Hans Ullitz-Moeller
BioBase, University of Aarhus
Email: hum@biobase.dk
Tel: +45-86-13 9788
URL: http://www.biobase.dk

Finland
Kimmo Mattila
CSC, Espoo
Email: Kimmo.Mattila@csc.fi
Tel: +358 9 457 2708
URL: http://www.csc.fi/molbio

France
PLAZA Jean-Marc
INFOBIOGEN, Evry
Email: plaza@infobiogen.fr
Tel: +33 1 60 87 37 11 Fax: +33 1 60 87 37 96
URL: http://www.infobiogen.fr

Germany
Sandor Suhai
EMBnet node at the German Cancer Research Center
Department of Molecular Biophysics (H0200)
Email: genome@dkfz.de
Tel: +49-6221-422 342 Fax: +49-6221-422 333
URL: http://genius.embnet.dkfz-heidelberg.de

Greece
Babis Savakis
Institute of Molecular Biology and Biotechnology
Heraklion, Crete
Email: savakis@nefeli.imbb.forth.gr
Tel: +30-81-391 114 Fax: +30-81-391 104
URL: http://www.imbb.forth.gr

Hungry
Endre Barta
Agricultural Biotechnology Center
Szent-Gyorgyi A. ut 4. Godollo,
Email: barta@abc.hu
Tel: +36 30-2101795
URL: http://www.hu.embnet.org

India
H.A.Nagarajaram
Laboratory of Computational Biology & Bioinformatics facility, Centre for DNA Fingerprinting and Diagnostics (CDFD), Hyderabad
Email: han@www.cdfd.org.in
Tel: +91 40 7155607 / 7151344 ext:1206 Fax : +9140 7155479
URL: http://www.cdfd.org.in

Israel
Leon Esterman
INN (Israeli National Node) Weizmann Institute of Science Department of Biological Services,Biological Computing Unit,Rehovot
Email: leon.esterman@weizmann.ac.il
Tel: +972-8-934 3934
URL: http://il.embnet.org

Italy
Marcella Attimonelli
BARI
Email: marcella@area.ba.cnr.it
Tel: +39(080)5482180 Fax: +39(080)5484467
URL: http://www.area.ba.cnr.it/embnet.htm

Mexico
Cesar Bonavides
Nodo Nacional EMBnet, Centro de Investigación sobre Fijación de Nitrógeno, Cuernavaca, Morelos
Email: embnetmx@cifn.unam.mx
Tel: +52 (7) 3 132063
URL: http://itzamna.cifn.unam.mx/embnet

The Netherlands
Jack A.M. Leunissen
Centre for Molecular and Biomolecular Informatics University of Nijmegen
Email: jackl@cmbi.kun.nl
Tel: +31 (0)24 365 3391
URL: http://www.cmbi.kun.nl

Norway
Rune Groven
The Norwegian EMBnet Node
The Biotechnology Centre of Oslo
Email: admin@embnet.uio.no
Tel: +47 22 84 0535
URL: http://www.no.embnet.org

Poland
Piotr Zielenkiwicz
Institute of Biochemistry and Biophysics Polish Academy of Sciences Warszawa
Email: piotr@pl.embnet.org
Tel: +48-22 86584703
URL: http://www.ibb.waw.pl

Portugal
Pedro Fernandes
Gulbenkian university, OEIRAS
Email: pfern@pen.gulbenkian.pt
Tel: +351 214407900 Fax: +351 2144079070
URL: http://www.pt.embnet.org

Russia
Sergei Spirin
Biocomputing Group, Belozersky Institute Moscow
Email: sas@genebee.msu.su
Tel: +7-095-9328825
URL: http://www.genebee.msu.su

Solvakia
Lubos Klucar
Institute of Molecular Biology SAS Bratislava
Email: klucar@embnet.sk
Tel: +421 7 5941 2284
URL: http://www.embnet.sk

South Africa
Ruediger Braeuning
SANBI, University of the Western Cape, Bellville
Email: ruediger@sanbi.ac.za
Tel: +27 (0)21 9593645
URL: http://www.sanbi.ac.za

Spain
José M. Carazo, José R. Valverde
EMBnet/CNB, Centro Nacional de Biotecnología, Madrid
Email: carazo@es.embnet.org, jrvalverde@es.embnet.org
Tel: +34 915 854 505 Fax: +34 915 854 506
URL: http://www.es.embnet.org

Sweden
Nils-Einar Eriksson, Erik Bongcam-Rudloff
Sweden Biomedical Centre, Computing Department, Uppsala
Email: nils-einar.eriksson@bmc.uu.se erik.bongcam@bmc.uu.se
Tel: +46-(0)18-4714017, +46-(0)18-4714525
URL: http://www.embnet.se

Switzerland
Laurent Falquet
Swiss Institute of Bioinformatics, Epalinges/Lausanne
Email: Laurent.Falquet@isrec.unil.ch
Tel: +41 (21) 692 5954 Fax: +41 (21) 692 5945
URL: http://www.isb-sib.ch

United Kingdom
Alan Bleasby
UK MRC HGMP Resource Centre, Hinxton, Cambridge
Email: ableasby@embnet.org
Tel: +44 (0) 1223 494535
URL: http://www.uk.embnet.org

Specialist Nodes

EBI
Rodrigo López
EBI Embl Outstation, Wellcome trust Genome Campus
Hinxton Hall, Hinxton, Cambridge, United Kingdom
Email: rls@ebi.ac.uk
Phone: +44 (0)1223 494423
URL: http://www.ebi.ac.uk

ETI
P.O. Box 94766
NL-1090 GT Amsterdam, The Netherlands
wouter@eti.uva.nl
Phone: +31-20-5257239 Fax: +31-20-5257238
URL: http://www.eti.uva.nl

EU
Dr Bernard Mulligan
DG Research - European Commission
Brussels BELGIUM
e-mail: bernard.mulligan@cec.eu.int

Hoffmann-La Roche
Martin Ebeling
F. Hoffmann-La Roche Ltd, Basel, Switzerland
Email: martin.ebeling@roche.com
Phone: +41-61-68-71398
URL: http://www.roche.com

ICGEB
Sándor Pongor
International Centre for Genetic Engineering and Biotechnology
AREA Science Park, Trieste, ITALY
Email: pongor@icgeb.trieste.it
Phone: +39 040 3757300
URL: http://www.icgeb.trieste.it

LION Bioscience
Peter Rice
LION Bioscience AG, Heidelberg, Germany
Email: Peter.Rice@uk.lionbioscience.com
Phone: +44 1223 224700
URL: http://www.lionbioscience.com

MIPS
H. Werner Mewes
Email: mewes@mips.embnet.org
Phone: +49-89-8578 2656 Fax: +49-89-8578 2655
URL: http://www.mips.biochem.mpg.de

Pharmacia - Biovitrum
Timoth Wood
Biovitrum, Stockholm, Sweden
Email: timothy.wood@eu.pnu.com
Phone: +46 (8) 695 9134
URL: http://www.pharmacia.se

Sanger Institute
Michelle Clamp
Welcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton Hall, Hinxton CB10 1SD, Cambridge, United Kingdom
Email: michele@sanger.ac.uk
Phone: +44-1223-494 967 Fax: +44-1223-494 919
URL: http://www.sanger.ac.uk

UMBER
Terri Attwood
School of Biological Sciences, The University of Manchester
Oxford Road, Manchester M13 9PT, United Kingdom
Email: attwood@bioinf.man.ac.uk
Phone: +44 (0)61 275 5766 Fax: +44 (0) 61 275 5082
URL: http://www.bioinf.man.ac.uk/dbbrowser

TCH-MGR
Email: tech-mgr@embnet.org
The team gives support to EMBnet nodes and helps them with maintenance and troubleshooting.
The team is formed of experienced system administrators and programmers who ensure the
availability of local services for all EMBnet users.

Dear reader,

If you have any comments or suggestions regarding this newsletter we would be very glad to hear from you. If you have a tip you feel we can print then please let us know. Please send your contributions to one of the editors. You may also submit material by Internet E-mail.

Past issues of embnet.news are available as PostScript or PDF files (ISSN 1023-4144). You can get them by anonymous ftp from:

the EMBnet organisation Web site
www.embnet.org
the Belgian EMBnet node
ftp://ftp.be.embnet.org/pub/embnet.news
the UK EMBnet node
ftp://ftp.uk.embnet.org/pub/embnet.news
the EBI EMBnet node
ftp://ftp.ebi.ac.uk/pub/embnet.news

Publisher:

EMBnet Administration Office.
c/o Jack Leunissen - CMBI
University of Nijmegen
6525 ED Nijmegen
The Netherlands

embnet
embnet.news
Vol.8, No.1
September 2002
ISSN 1023-4144