embnet.news

Volume 5 Nr2 (ISSN1023-4144) June 30, 1998

Editorial

This year EMBnet celebrates its tenth anniversary. In the fast moving field of bioinformatics, you don't have to wait 21 years to come of age. Indeed, after 21 years you are almost certainly obsolete.

Ten years ago, bioinformatics was, for many of us, an amateur occupation. Amateur, however, only in the sense that the cost of making significant contributions making the financial rewards, if any, very low. I well remember being told that "a simple FORTRAN program - It'll take you ten minutes" would provide the analytical tool necessary to crack an unsolved evolutionary problem.

The days when the whole EMBL DNA database could be printed out on form-feed paper and scanned by eye were before my time, but in 1990 the datasets were still tiny and 'missing data' could be extremely frustrating. Now the datasets are enormous: four Encyclopaedia Britannicas' worth of bases in the current EMBL.

There is still a role for those simple FORTRAN programs, by now translated into C++ or perl scripts to feel sufficiently modern, but having available the complete genome sequence of many organisms, you can address really big evolutionary questions. And with the complete genome sequence of an organism you no longer have to worry about any sampling bias.

There have also been big changes as to rewards. All of us have friends and former colleagues who now have uncountable salaries, horrific workloads and company cars.

Contents

Two superlative "amateur" projects are this summer moving into the professional and commercial arenas.

Both SwissProt and SRS were started through the vision, brilliance and high standards of single academic researchers, carried on for many years on something less than a shoestring and suffered from a series of funding crises. The research community including amateur, academic, commercial and multinational came increasingly to appreciate that these were essential tools for their trade.

Now, at least for commercial users, this long free lunch is set to end. There is a report on the development of SIB, which incorporates the Swiss end of SwissProt in this issue. In contrast to bioinformatics, sequencing has always had much higher intrinsic costs. In recent years, partly because automation has brought the cost of each sequenced base down, mega-sequencing projects have become almost standard practice provided the funding was available.

Rumour had it that, before TIGR delivered the Haemophilus influenzae genome into the public domain, this organism had been sequenced at least twice, privately, by rival pharmaceutical companies.

In this issue Peter Rice describes how the Wellcome Trust has committed substantially more than 100 Mecu to the Sanger Centre for its human genome sequencing project. Let us hope that genome data is released as quickly as possible to the public domain so that all of us can use it to continue our research, be it applied academic or wholly theoretical.

We have seen how commercial concerns depend upon pure science in universities and research institutes. If the commercial world fails to nurture the academic community, by making data available, indirectly through tax dollars or by direct funding or collaboration then this resource will shrivel up and blow away on the winds of change.

The embnet.news editorial board:

Alan Bleasby
Rob Harper
Robert Herzog
Andrew Lloyd
Rodrigo Lopez
Peter Rice

Press Release 13th May 1998

Wellcome Trust Announces Major Increase in Human Genome Sequencing

The Wellcome Trust has announced a major increase in its flagship investment in British science in the sequencing of the human genome, the book of life. Previously committed to funding the sequencing of one sixth of the human genome at the Sanger Centre, the Wellcome Trust has today decided to double this to one third. This decision will make available an additional £110 million over seven years, bringing the total Trust investment in the Human Genome Project to £205 million.

The Trust is concerned that commercial entities might file opportunistic patents on DNA sequences. The Trust is conducting an urgent review of the credibility and scope of patents based solely on DNA sequences. It is prepared to challenge such patents.

The Wellcome Trust is the leading European funder of human genome sequencing and has established the Wellcome Trust Genome Campus to help achieve this. Its early entry into this work has enabled Dr John Sulston, Director of the Sanger Centre, and his colleagues, to generate one third of all the sequences which have been produced to date. Todays announcement will ensure the maintenance of this position. The Human Genome Project is one of the most significant projects being carried out in scientific research. The aim of the programme is to identify, to a high degree of accuracy, the complete sequence of the human genome and make this data immediately and freely available to the international research community so as to increase the effectiveness of biomedical research at all levels.

This work is vital to
- future health
- the understanding of biology
- the UK pharmaceutical industry
- support the strong UK biomedical science base

The Human Genome Project is an international project with the most substantial funding contributions from the Wellcome Trust and the US government. The Sanger Centre and Washington University at St Louis in US have led the international collaborative process and through specific meetings co-ordinating the Human Genome Project, established the international policy of free release of data and putting it in the public domain.

This week a commercial venture announced its intention to produce partial sequences of the human genome, to delay release of this information and to have exclusive rights to patent some of these sequences. This venture will not fulfil the aims of the international collaboration of the Human Genome Project, although it will provide complementary information.

The Wellcome Trust believes that the human genome should be sequenced, through an international collaboration, as speedily and accurately as possible, with the results being placed immediately in the public domain. To this end, it is to open discussion with existing members of the Human Genome Project with a view to an international agreement whereby up to 50% of the genome could be sequenced in the UK.

For further information:
Kate Davey/Noorece Ahmed/Catherine Nestor
Wellcome Trust Press Office
Tel: 0171 611 8612/8540/8846
Fax: 0171 611 8416
E-mail: press.office@wellcome.ac.uk

The Wellcome Trust is the world's largest charity spending some £250 million annually on medical research. The Wellcome Trust supports more than 3000 researchers, at 300 locations, in 30 different countries - laying the foundations for the healthcare advances of the next century and helping to maintain the UK's reputation as one of the world's leading scientific nations. As well as funding major initiatives in the public understanding of science, the Wellcome Trust is the country's leading supporter of research into the history of medicine.

ANGIS Node focus

History & structure

ANGIS is Australia's EMBnet node and joined the association in 1996. ANGIS (the Australian National Genomic Information Service; http://www.angis.org.au) is run from AGIC, the Australian Genomic Information Centre at the University of Sydney. ANGIS was established in 1991 when the Australian government called for the formation of a National Genomic Information Service. The service has been run from Sydney ever since, although the service is becoming increasingly
decentralised with ANGIS Nodes being established around the country and wider regions.

A Big Country

Australia is a large and diverse country. With a surface area almost as large as the continental USA Australia has, however, a population density 14 times smaller. With strong

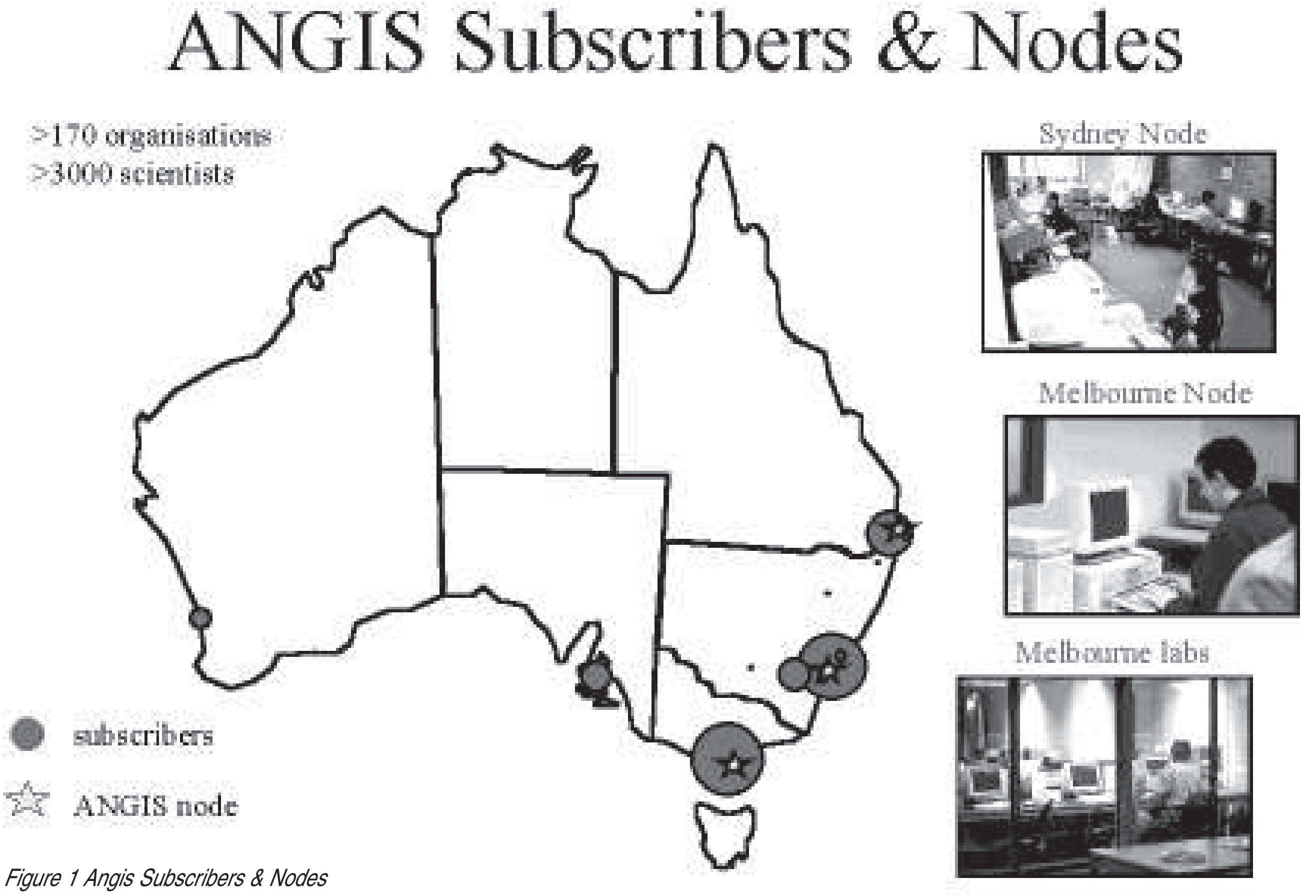


Figure 1 Angis Subscribers & Nodes

historic, cultural and economic ties to Europe then Australia, however, sits comfortably as a member of EMBnet, even if it is located almost directly opposite on the globe (explaining the bleary eyes Australians have as they travel for >24 hours across a dozen time zones to get to Europe!). Australia is, however, also part of Asia, increasingly sharing cultural and economic links with the region. As a consequence, Australia is one of the founding members of the fledgling APBioNet (http://apbionet.angis.org.au), which has broad goals similar to EMBnet.

ANGIS Services

Distribution - ANGIS Nodes

In January 1996 ANGIS was in crisis: Internet traffic bottlenecks between Sydney and outlying regions were preventing effective delivery of ANGIS's Internet services to the nation. As a consequence ANGIS, with the assistance of the Howard Florey Institute (http://www.hfi.unimelb.edu.au), established an ANGIS Node in Melbourne (Figure 1). Engineering the nodes was a significant task as the main server had evolved over 6 years and consequently the system was readily portable. ANGIS took this opportunity to re-engineer the entire system (user discs, database organisation and updates, software engineering practices, etc) to allow increasingly efficient operation. Further adding to the tasks of the ANGIS Team, the Node machines are all administered and maintained from a remote location (Sydney). The Melbourne Node was launched in 1996 and the Brisbane Node in 1997. ANGIS also has built a bioinformatics education facility in Melbourne in collaboration with the Howard Florey Institute which is used for regular courses run in that city. Recently ANGIS has also developed and installed a number of intranet nodes.

Internet

ANGIS, like all EMBnet nodes, offers a suite of databases and software from public and commercial sources. ANGIS has carried out a large amount of R&D into software and database integration and user-interface development. Some of these efforts are described below.

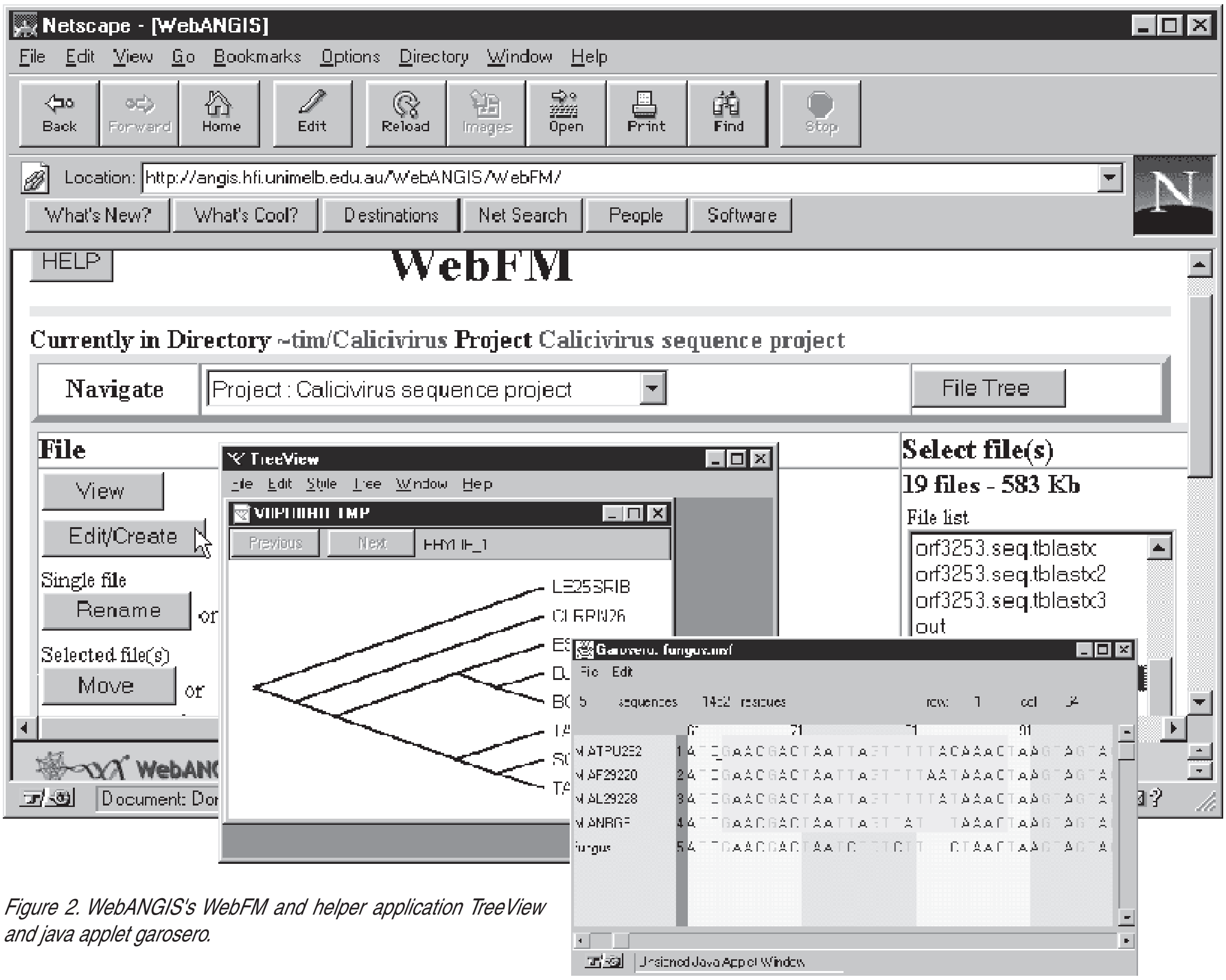


Figure 2. WebANGIS's WebFM and helper application TreeView and java applet garosero.

WebANGIS

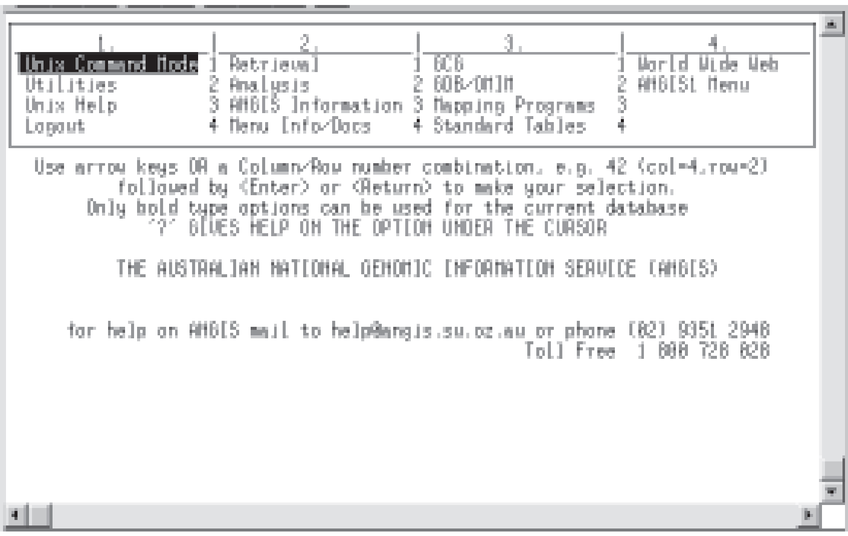
As the main interface to ANGIS's resources, WebANGIS is an integrated WWW interface to most of the software and databases on ANGIS. The heart of WebANGIS is WebFM (file manager), a highly tuned interface to the UNIX file system that is end-user biologist friendly.

WebFM allows navigation, visualisation and editing of data and result files from a large number of applications which support a range of helpers as well as java applets developed at ANGIS (such as pepplotter for visualising PEPTIDESTRUCTURE output and garosero for multiple sequence alignment editing; Figure 2).

2D-ANGIS

The original ANGIS interface was the 2D-ANGIS system. Still in widespread use and popular because of its minimal bandwidth and client-hardware requirements 2D-ANGIS is a telnet interface to ANGIS's services. (Figure 3). Ever increasing popularity of WebANGIS means that 2D-ANGIS development has ceased.

Fig 3. 2D-ANGIS



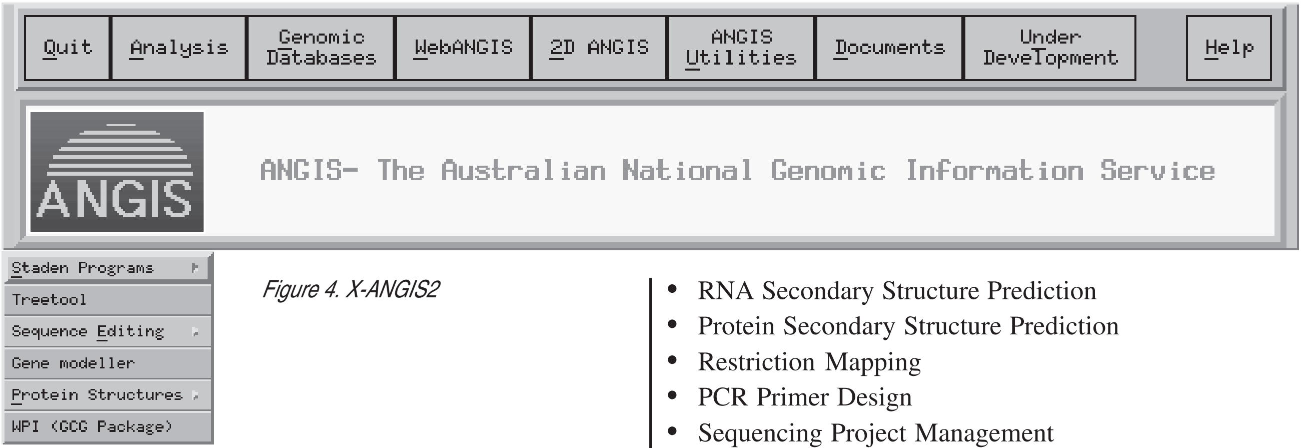


Figure 4. X-ANGIS2

X-ANGIS

ANGIS supports a large number of databases and software packages that need X-Windows (eg AceDB and the Staden package). Developed around 1993 X-ANGIS was an interface to these applications and many others, as well as having a number of other features such as file visualisation and manipulation. With the advent of WebANGIS, X-ANGIS2 was developed which served as a single point of access "launching-pad" for X-Windows applications and databases (Figure 4).

Training

ANGIS has always had a strong commitment to training: in 1998 this equates to approximately 10 one week courses run all around the country and 20 one day workshops. This necessitates a lot of travel by education staff and a lot of logistics by the support team in Sydney, including organising computer labs, taking registrations, etc. ANGIS now has a fully integrated database (implemented in Sybase) of subscriber information that is linked to course enrolments with WWW interfaces allowing on-line enrolment and internal tools for account maintenance.

Curriculum of the one week course is typically:

- Introduction to WebANGIS
- Introduction to 2D ANGIS
- Introduction to XANGIS
- Database Sequence Retrieval
- Sequence Comparison
- Sequence Editing
- Database Searching
- Sequence Editing
- Database Searching
- Multiple Sequence Analysis
- Pattern and Motif Searching
- Phylogeny and the Ribosomal Database
- RNA Secondary Structure Prediction
- Protein Secondary Structure Prediction
- Restriction Mapping
- PCR Primer Design
- Sequencing Project Management
- Gene Discovery
- Genomic Databases

Specialised workshops typically focus on subjects such as "using WebANGIS", molecular modelling, phylogenetics, linkage analysis and other areas.

To support these activities, ANGIS produces a 4 volume bioinformatics handbook. Used for self-paced training, in the ANGIS courses and in undergraduate education programs around the country, this series follows on from previous works such as the "ANGIS Exercise book". See http://www.angis.org.au/Education/Materials/book.html for more details (Figure 5).

Figure 5. The ANGIS Bioinformatics handbooks

Support

ANGIS looks after its subscribers through email and telephone support (including a toll free 1 800 number) as well as through trade displays at conferences, site visits and the ANGIS courses.

Education

In addition to the training courses run by ANGIS (see above), the service is used in over 9 undergraduate university courses around the country and abroad (including SE Asia and the United States). In addition, AGIC (the centre that runs ANGIS) has research students in honours and PhD programmes. Furthermore, AGIC is involved in the new Bachelor of Science (BioInformatics) degree at the University of Sydney.

Research & collaborations

ANGIS conducts a large amount of internal R&D in the area of bioinformatics applications integration and user interface development. Furthermore, AGIC has a number of collaborative research projects with industry, this spans molecular modelling, structure prediction, genome analysis and genome data management systems amongst others. Academic collaborative research projects with groups in Australia and around the globe include population genetics, protein analysis, genome databases, comparative genomics and phylogenetics amongst others. Commercial partners include Sun, SGI, MSI and ForBio research. ANGIS is always seeking out new collaborations with academic and industry groups.

Funding

Service delivery for ANGIS is more than cost-recovery through subscriptions from over 170 organisations around the country. In addition, ANGIS receives grants from the Australian NH&MRC (National Health and Medical Research Council) and in the past from the ARC (Australian Research Council) and rural research bodies, as well through generous equipment donations from Apple and Sun. AGIC's research activities are supported by the ARC and its industry partners, SGI, MSI and ForBio. In addition, in 1997 AGIC received some support directly from the University of Sydney.

Credits

ANGIS's success comes from the tireless work of its team members (currently numbering over 20): programmers, database developers, systems administrators, education and support staff, business and administration personnel, students and volunteers (please consult the ANGIS WWW site for an up to date list). Thanks also to the funding bodies who have and continue to support ANGIS, and for the support of the research and education community in Australia.

Further Information

Please contact Tim Littlejohn for any further information or visit the ANGIS WWW site.

EMBOSS: A European Software Suite

Peter Rice, The Sanger Centre, Hinxton, UK.

Summer 1998 will see the launch of a new initiative in sequence analysis, using the bioinformatics expertise and resources of EMBnet to produce a suite of sequence analysis applications especially suited to the needs of EMBnet's usership of over 30,000 people.

The European Molecular Biology Open Software Suite (EMBOSS) is a sequence analysis development centred at the Sanger Centre, one of the EMBnet specialist members. Using software libraries developed in collaboration with the UK EMBnet node at SEQNET and with contributions from the HGMP Resource Centre, EMBOSS aims to support a broad range of sequence databases and formats to run under most common user interfaces.

EMBOSS at the Sanger Centre is funded through the Wellcome Trust. We aim to develop many new applications and to encourage other developers to integrate their own software with EMBOSS. The specifications for the libraries and interfaces are taking shape and can be seen on the project web site http://www.sanger.ac.uk/Software/EMBOSS/ where we provide a useful list of the latest updated pages.

EMBOSS grew out of the "Extended GCG" project (see embnet.news 3(1):2-4), and the need for a completely free library with full source code and no proprietary claims.

EMBOSS builds on ten years of experience of sequence analysis code development. It designs ways to integrate software into many of the menu and interface systems used around EMBnet. It will make life easier for laboratory biologists who have to work with computers maybe only once or twice a year.

A key feature is the "AJAX Command Definition" (ACD) for each application, which controls the user interface and should be portable to many other systems.

The first EMBOSS application is a very small program called "seqret" which simply reads in a sequence, and writes it out again. This is all that the command definition needs to know and takes up very few lines in the ACD file. There are possibilities to make this file even shorter in a future release.

```
seqret.acd

appl: seqret [
         sequence: sequence [ required: Y prompt:
"Input sequence" ]
       outseq: seqout [ required: Y prompt: "Write
to"]
      ]
```

The seqret program source code is just as simple. The call to "embInit" uses the ACD file to read in a sequence and to open an output file. "ajAcdGetSeq" simply picks up a reference to the sequence which has already been read. "ajAcdGetSeqout" picks up a reference to an open sequence output file. "ajSeqWrite" does have a little work to do. It identifies the output format and writes the sequence. Finally, "ajExit" can close open files and do some tidying up.

```
seqret.c

#include "emboss.h"

      int main (int argc, char * argv[]) {

        AjPSeq seq;
        AjPSeqout seqout;

        embInit ("seqret", argc, argv);

        seq = ajAcdGetSeq ("sequence");
        seqout = ajAcdGetSeqout ("outseq");

        ajSeqWrite (seqout, seq);

        ajExit ();

      }
```

The magic comes from the way the AJAX library works. The "embInit" call can understand many kinds of command line. Some look like Unix, but there are ideas from many other places. Each of the following command lines will read a sequence file "paamir.tfa" in fasta format, starting at base 25, and write it to the terminal.

```
seqret fasta::paamir.tfa -sbegin=25
seqret -sbegin=25 fasta::paamir.tfa
seqret -sbegin=25 paamir.tfa -sformat fasta
seqret -sbeg 25 paamir.tfa -sf=fasta
seqret sbeg=25 -sequence=paamir.tfa sf=fasta
seqret -sbeg 25 -sequence paamir.tfa -sf fasta
seqret -sbeg 25 paamir.tfa -sf fasta
```

EMBOSS can provide a front end to many of the standard applications. More importantly, as can be seen from the example of "seqret", it is quite easy to write new applications using EMBOSS and to make them available to a very large user community. At the Sanger Centre we are committed to replacing the most popular applications in the obsolete "extended GCG" package as soon as we can. We hope that other software developers in EMBnet and beyond will contribute their applications to help EMBOSS grow into a major resource for the biologist and bioinformatician.

Life is not all hard work of course, and there is room for humour in any project. Telling a program what sequence you want is a complicated business. Borrowing from the "Uniform Resource Locator", or URL, familiar to Web users, we have specified our own "Uniform Sequence Address". Many thanks to Rodrigo Lopez for requesting switches to make sure that sequences are in upper or lower case. As all input sequence options start with the letter "s", EMBOSS quickly gained two new qualifiers "-supper" and "-slower", though we wonder whether or not to include them in the documentation.

Talking of documentation, this can be a major problem for academic software developers. For EMBOSS, we plan to make documentation and support available through EMBnet. The "embnet.news team" produces a newsletter which you, gentle reader, are reading now. The "EMBnet Publications and PR committee" produces other goodies such as the new EMBnet "A quick guide to Unix". We hope that they will be able to produce a high standard of user documentation for EMBOSS, and we will do all we can to make their job an easy one.

Support has always been strong in EMBnet. The national node managers will be able to provide support for their users, and they will in turn be assisted by the EMBnet Technical Manager group, who will in turn be guided by the EMBOSS developers within EMBnet.

By working together, we hope that we can make EMBOSS a software package that will help many users and academic developers and will continue to grow for years to come. EMBOSS and EMBnet fit very well together. Their future is looking bright.

A WWW site devoted to protein structural topology

David R. Westhead[1], Daniel C. Hatton[1], David R. Gilbert1,[4] and Janet M. Thornton[1,2,3]

We have recently set up a WWW site[1] devoted to protein structural topology. The most important service available on this site is an atlas of protein topology (TOPS) cartoons, in which most of the structures in the Brookhaven data bank[2] (PDB) are represented. There is also a "server" to which protein structures can be submitted for topology cartoon calculation and HTML pages explaining the topology cartoons and giving information about protein structural topology in general. In addition, users can search them in atlas, or a TOPS version of the PDB, using motifs from a library or by defining their own search patterns.

Protein three dimensional folds can be complicated and difficult to interpret. Protein topology cartoons are a graphical form of simple two dimensional schematic diagrams whose aim is to simplify folds so that they can be more easily understood and compared. They represent a fold as a sequence of secondary structure elements and hold information about their relative orientation and spatial position.

An example cartoon is shown in figure 1. Information about its interpretation is given in the caption. The topology cartoon shows the beta sandwich structure of the fold with the relative spatial and sequential positions of the constituent strands. This is much more clear than the equivalent three dimensional structure.

The TOPS Atlas

The atlas of topology cartoons was generated from the version of the PDB current on the 1st July 1997. In order to avoid the generation of many duplicate cartoons, the chains present in the databank were first clustered at a sequence similarity threshold of 95%. Chains consisting of nucleic acid sequences were removed, as were protein chains of less than 30 residues. From an original total of

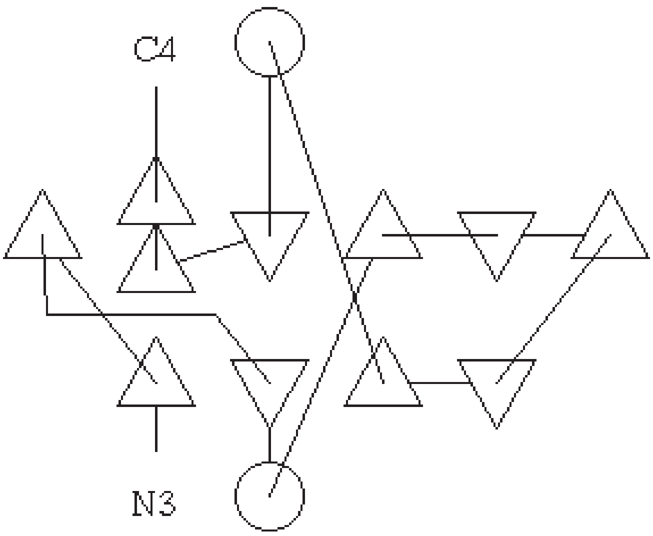


Figure 1. The topology cartoon for the variable domain of the heavy chain of the antibody fragment structure with Brookhaven code 6fab, along with the three dimensional structure for comparison. In the TOPS cartoon beta strands are represented by triangular symbols and helices by circular ones. The peptide chain follows the connecting lines between symbols starting at N3 and ending at C4. The relative direction of beta strands is shown by the orientation of the triangles. Strands are viewed as having one of two directions: "up" strands are shown as upward pointing triangles and should be thought of as representing strands directed out of the plane of the diagram; "down" strands are shown as downward pointing triangles and represent strands directed into the plane of the diagram. Thedirections of the helices can be deduced by studying how connecting lines are drawn: if the N terminal connection is drawn to the centre of the symbol and the C terminal one to the edge then the direction is down, otherwise the N terminal connection is drawn to the edge and the C terminal one to the centre and the direction is up.

1. European Bioinformatics Institute, EMBL outstation, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK, CB10 1SD
2. Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, London, WC1E 6BT, U.K.
3. Laboratory of Molecular Biology, Department of Crystallography, Birkbeck College, University of London, Malet Street, London, WC1E 7HX, U.K.
4. Department of Computer Science, School of Informatics, City University, Northampton Square, London, EC1V 0HB, U.K.

10534 chains this produced 2144 clusters of near identical sequences. From each cluster a representative TOPS diagram was produced from a single structure. This was chosen to be the highest resolution X-ray structure in the cluster, or an N.M.R. structure if no X-ray structures were available. Within a chain, each structural domain was plotted separately using domain definitions taken from the CATH[3] protein structure classification.

The cartoons were generated automatically, in the first instance, using a substantially modified version of the program TOPS [4]. Whilst the original version of TOPS would produce satisfactory cartoons for more simple protein folds, it was found to be unable to do so for many more complicated folds. The modifications were necessary in order to increase the success rate of the program sufficiently to make automatic generation of a large number of cartoons a viable proposition. The generation of the atlas of cartoons was viewed as a test of the new version of the program. Each cartoon in the atlas was checked manually with the 3D structure of the protein, and the success rate in producing satisfactory cartoons was found to be 82%. Amonst the failures were many cartoons which were correct but not aesthetically pleasing, but there were still some complicated folds for which the program failed. The cartoons judged to be failures were corrected by hand editing and included in the atlas.

The atlas can be viewed using an applet (a program written in the Java programming language, delivered over the WWW, and run on the client machine). A basic applet using Java version 1.0 simply allows the user to view the cartoons, whilst users with a WWW browser supporting Java version 1.1 can use a much more functional applet which allows editing and printing of the cartoons. The same applets are used for viewing, editing and printing cartoons generated at the request of the user by the server facility. Some users with older machines and/or browsers have experienced difficulties with the Java technology and for this reason a purely HTML/GIF version of the atlas is also provided

We hope to keep the atlas up to date as new structures arrive in the PDB. However, because updates to the atlas require significant effort, we anticipate that there will always be a time lag between structures arriving in the PDB and cartoons being put into the atlas. In this case users will be able to use the server to generate their own cartoons for the new structures. The software used in the generation of the atlas will be made available in some form, and details will be posted on the WEB site.

Protein topology pattern searching over TOPS databases

We have developed a system which supports fast pattern searching over TOPS protein topology databases. The search engine is based on a string-graph algorithm and uses constraint propagation to prune the search space. Users can search on motifs from a library or define their own search patterns. All 15400 descriptions of protein domains currently in the Brookhaven Protein Data Bank (April 1998) have been converted into a database of TOPS diagrams (11MB). Average match times are 3-5ms per diagram on a Dec-Alpha. For example, it takes 80 secs to find all matches to a jelly-roll type 1 description (500 hits out of 15400) on a Dec-Alpha. Some queries take advantage of precompiled topological information generated by the TOPS program; a search for Tim barrels takes 15 seconds to find 260 matches in the database. Load times are additionally 3 sec for the atlas and 32 sec for the PDB. Users can search on motifs from a library or define their own search patterns. We are now in the process of enhancing the system to permit users to compare the topology of a given domain with all the other domains in a database.

A note on TOPS diagrams and patterns

A TOPS diagram describes more information about a structure than the corresponding cartoon. In addition to the Secondary Structure Elements - their type, position on the backbone, and orientation relative to the plane of the page. The diagram describes the H-bonds (parallel or anti-parallel) between strands, and the chiralities (right or left) between some of the SSEs. For example, the TOPS diagram for 2bop is shown below; it comprises five strands, which form a bifurcated anti-parallel sheet, and three helices. There are two right-handed chirality connections, between strands 1 and 4, and between strands 6 and 8 respectively (see next page, fig.1).

A TOPS pattern is like a TOPS diagram, and may describe several (none or more) TOPS diagrams. This is achieved by permitting insertions of some SSEs into the pattern to obtain a diagram. Each backbone connection between every pair of SSEs is annotated with a pair of integers standing for the minimum and maximum number of SSEs which can be inserted. The values of min and max can range from 0 to a large number (in practice 60) which we denote N. For example, a plait pattern (or motif), which matches amongst other 2bop, is illustrated below(fig.2):

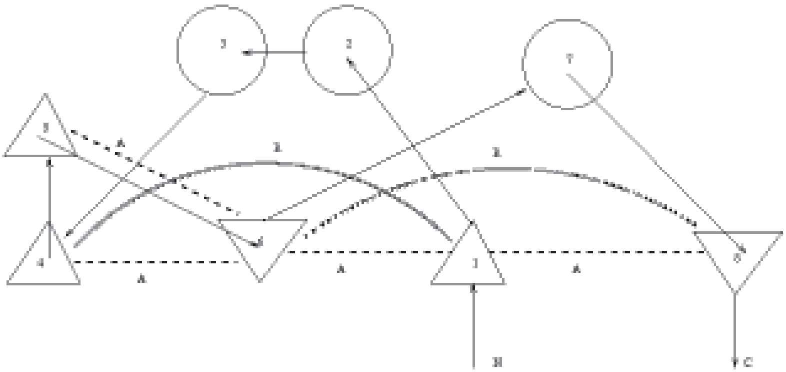


Fig. 1 - TOPS diagram for 2bop

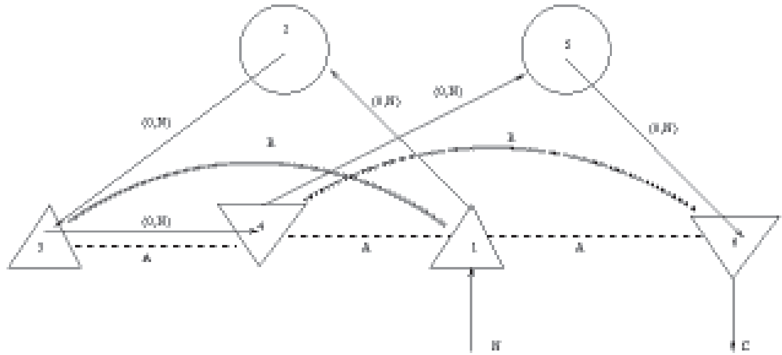


Fig. 2 - Plait pattern

This pattern describes, amongst others, 2bop. In order to illustrate this, consider the plait pattern and the diagram for 2bop to each be `stretched out' and laid side by side, as below (fig3 and 4):

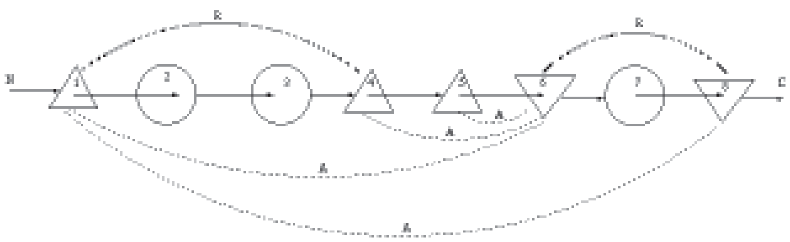


Fig.3 - TOPS diagram for 2bop

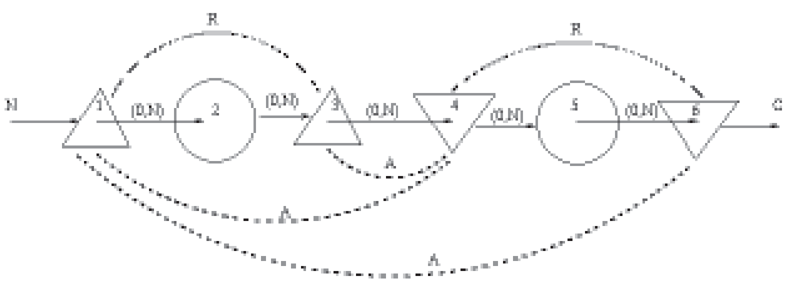


Fig.4 - Plait pattern

Matching a pattern to a diagram

Informally, we match a pattern to a diagram by matching on the SSEs, the H-bonds and the chiralities. When matching on the SSEs we obtain a correspondence. This is a sequence of matching pairs of SSEs, one for each SSE in the pattern, where first member is an SSE from the diagram and the second is an SSE from the pattern. We can describe the result of matching by the the correspondence and also a list of the total number of inserts between adjacent members of the SSE sequence in the diagram. A pattern may match a diagram in none or more ways and may match none or more diagrams in a given database of diagrams.

For example, there are two ways in which the plait motif can match the diagram for 2bop:

Character matches: (1,1),(2,2),(4,3),(6,4),(7,5),(8,6).
Inserts: (0,1,1,0,0)
Character matches: (1,1),(3,2),(4,3),(6,4),(7,5),(8,6).
Inserts: (1,0,1,0,0)

Using the TOPS query system

The TOPS query system is accessed over the Web using an HTML form, and there is on-line help.

Users can select which database (Atlas or PDB) to search, and optionally restrict the search to one given domain. Searches can be speeded up by filtering on TOPS precompiled fixed structure information about targets (e.g. if they contain barrels, sheets of various curvatures or sandwiches).

Users can select several output parameters, the order in which they are output for each successful match, and whether to sort on these. Output is incremental unless sorting is selected. Output parameters include domain name, CATH-number, matching node numbers, and sum of inserts between the nodes. There is an HTML link for each domain found by the search to the TOPS Atlas, which may hold the cartoon (graphical) representation of the domain or a representative domain and also to the CATH database at UCL. Predefined queries can be made for greek keys, jelly rolls, NAD-binding domains, immunoglobulins, plaits, barrels or various types, trefoils and propellers (or all of these). Alternatively, users can construct their own query patterns based on the sequence of SSEs and associated inserts and the associated sets of H-bonds and chiralities. There are facilities to define constraints over the total number of inserts, and parallel/anti parallel H-bonds.

Acknowledgements

We are grateful to Dr. T. P. Flores for giving us the source code for TOPS and allowing us to modify it without restriction. We are also grateful to Dr. C. A. Orengo for providing us with the domain boundary file associated with the CATH3 protein structural domain classification. David Gilbert has been supported by City University, and also by an EPSRC Visiting Fellowship.

References

1. Westhead, D.R., Hatton, D.C., and Thornton, J.M., Trends in Biochemical Sciences 23, 35-36 (1998).
2. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B.,Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol. 112, 535-542 (1977).
3. Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M., Structure 5, 1093-1108 (1997).
4. Flores, T.P., Moss, D.S., and Thornton, J.M., Prot. Eng. 7, 31-37 (1994).

Book review

Biological Sequence Analysis: probabilistic models of proteins and nucleic acids.

R.Durbin, S.Eddy, A.Krogh, G.Mitchison.
Publ. Cambridge University Press
ISBN 0 521 62041 4 - 55.00GBP/80.00USD Hback
ISBN 0 521 62971 3 - 19.95GBP/34.95USD Pback

There would be no genetics without variability and it is hard to think of a sensible approach towards a theory of evolution which did not deal with probability. Evolution has given us a mind-boggling diversity of organisms, genomes and gene sequences to deal with and it would be surprising if simple probabilistic models could adequately explain what has been going on since the primeval soup started bubbling. This book attempts to put an accessible face on some of the necessarily complex models that begin to make sense of such topics as sequence alignment, profiles, phylogenetic trees and RNA secondary structure. Given the authors you can anticipate that hidden Markov models (HMMs) feature prominently. There is a good section entitled "Towards more realistic evolutionary models" which will be heartening to those who realise that some published and even widely used models are simplistic to the point of uselessness.

This is a heavier book than its 0.63 kg, 350 pages and reasonable price would suggest. For those of us whose hearts sink at sentences like "First, let us establish some notation", it is very hard work. When I found a small mass of parentheses, subscripts and greek linked by "From this it is easy to see" to a rather larger mass of symbols, then I almost conceded defeat. Don't give up folks ! It's worth the struggle. Perhaps a good way for the moderately math-anxious to benefit would be to take it chapter by chapter with a small committed support group. It is not recommended for the seriously notationlly anxious.

But what is the point of notation ? It is surely to make ideas clear and concise and unambiguous. Unfortunately, formal notation robust enough to be useful seems to run counter to the need for accessibility. One of the most rewarding pages for me was an explanation, in plain English, of the heuristics that make clustalw a good program. I suppose that, being foreign territory, the authors did not presume to conjure up some notation - to the reader's benefit. In many places, the establishment of the notation is separated by many pages from its continued use, so a case could be made of a notation glossary. On the other hand, the index is good enough to cope with text based expressions such as the Viterbi Algorithm, affine gap scores, CYK, FSA and EVD - just look to the first cited page for an explanation.

In several places the authors recommend a non-linear approach to reading the book or refer the reader forward to a more complete explanation of a topic. They are even kind enough to provide a flow diagram for possible paths from preface to index. This is good because it is important to know which parts can be skipped or taken on trust and which are an essential foundation to later material. This way of viewing the book's structure makes me think that it would work well in a less linear medium, such as the WWW. And while we have the web in mind, I suggest that it is not too late for the answers, with working, to the many exercises to be provided on a suitable web site. This would surely help us strugglers to get more out of the book.

You get a lot of meat for your twenty pounds but you have to chew it well to benefit.

The Genome MOT

The Up-to-Date Status of Major Genome Sequencing Projects

Peter Sterk[1] and Stephan Beck[2]

Introduction

In 1996 the sequencing of the first eukaryotic genome, that of Saccharomyces cerivisiae, was completed.This was around that time that large scale sequencing really started to take off. Since then a number of smaller microbial genomes have been sequenced completely and it is expected that the second eukaryotic genome, that of the worm Caenorhabiditis elegans and 100 Mb in size, will be finished by the end of this year. The list of ongoing genome sequencing projects, currently over 300, is steadily growing (see the MAGPIE World Wide Web site). Systematic efforts to sequence a number of larger genomes, e.g. of the eukaryotes Arabidopsis thaliana (~100 Mb), Drosophila melanogaster (~120 Mb) and Homo sapiens (~3000 Mb), are well underway and completion is anticipated during the first decennium of the next millenium. Anticipated? yes! Take for example the Human Genome Project, with so far only about four percent of the human genome sequenced and funding secured for roughly half of the project, one wonders how a completion date could be predicted at all. Until recently, there wasn't even a suitable system to monitor the actual progress and, because sequencing of these larger genomes is carried out in many different laboratories spread all over the world, this was not a trivial task either. Yet, a year has been predicted and, as it is going to be the year 2005. We felt the need for system which gives up-to-date status reports and have recently established a genome monitoring table (Genome MOT), which allows the progress of a number of projects to be viewed via the World Wide Web (Beck and Sterk (1998).

1. EMBL Hinxton Outstation - The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK
2. Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

The Genome MOT

Genome MOT Genome Monitoring Table
EMBL Nucleotide Sequence Data Library

It is the policy of the Human Genome Organization (HUGO), the body which coordinates the human genome sequencing effort, to put sequence data in the public domain. Many other projects follow this principle too; and it seems to be in the interest of the genome sequencing centres themselves, as visibly high output and efficiency are likely to attract additional funding. Sequence data are submitted to the collaboratively maintained EMBL/GenBank/DDBJ database. Being the only centralised depository for nucleotide sequences, it is the most suitable source for the calculation of progress statistics. The Genome MOT currently lists the total amounts of public and finished genomic DNA in the EMBL database broken down per year for Homo sapiens, Arabidopsis thaliana, Drosophila melanogaster, Mus musculus, Saccharomyces cerevisiae and Schizosaccharomyces pombe, with the human data broken down further according to chromosome number. Data from the HTG division, which only contains unfinished high-throughput genome (HTG) sequences, are also presented separately to give a more complete picture. To indicate the current status of each of the projects as accurately as possible, the total amounts of finished sequence are presented both as absolute values and as percentages of chromosome or genome size with the estimated database redundancy (sequence duplication) into account (see next section). The genome MOT is automatically updated once a week. A cumulative progress plot is provided, too, which clearly shows the progress of these projects (see Figure) and our status report would not be complete without our own prediction of completion dates.

Database redundancy

It is immediately evident from the Genome MOT that the (finished) 12 Mb yeast genome appears to be oversubmitted more than twice. This prompted us to estimate the redundancy for each of the projects. By applying a somewhat arbitrary sequence length cutoff the contribution of smaller database entries can be eliminated, and since the major projects only tend to sequence larger clones, this seems to be an easy way to reduce redundancy. What that cutoff should be is debatable, hence we present our results applying cutoffs of 1000, 10,000, 30,000, 50,000 and 100,000 base pairs. In addition, we have calculated the redundancy in the datasets containing sequences longer than 1000 base pairs using the program CLEANUP by Grillo et al. (1996). CLEANUP considers a sequence to be redundant if it (or its complement) shows a degree of similarity and overlap with a longer sequence in the dataset greater than a certain threshold. As thresholds we chose 85% for the overlap, and 95% identity in the overlapping region. In our experience, CLEANUP is very good at finding sequence matches, and the vast majority of redundant sequences fully overlapped with a longer sequence and the overlapping regions were in most cases more than 99% identical. To give an indication of accuracy, the total redundancy calculated with the aid of CLEANUP was about 48% for yeast, reasonably accurate and for our purpose acceptable. The redundancy values thus obtained are taken into account when the status reports for sequences longer than 1000 base pairs are generated.

Future enhancements

A number of enhancements have been planned, and by the time you read this, some of those will have been implemented.

We are planning to present status reports for other genomes, including smaller genomes, as well, as long as a project submits data regularly and before actual completion. We are about to present tables containing EMBL accession numbers for each of the projects hyperlinked to the corresponding EMBL database record, and possibly present the corresponding clone IDs and originating sequencing centre if this can be automated. This should provide a suitable mechanism to check the validity of each of the database records. We are working on actual genome representation, and will initially present a list of completed genomes with hyperlinks to a database representation of that genome.

And finally ...

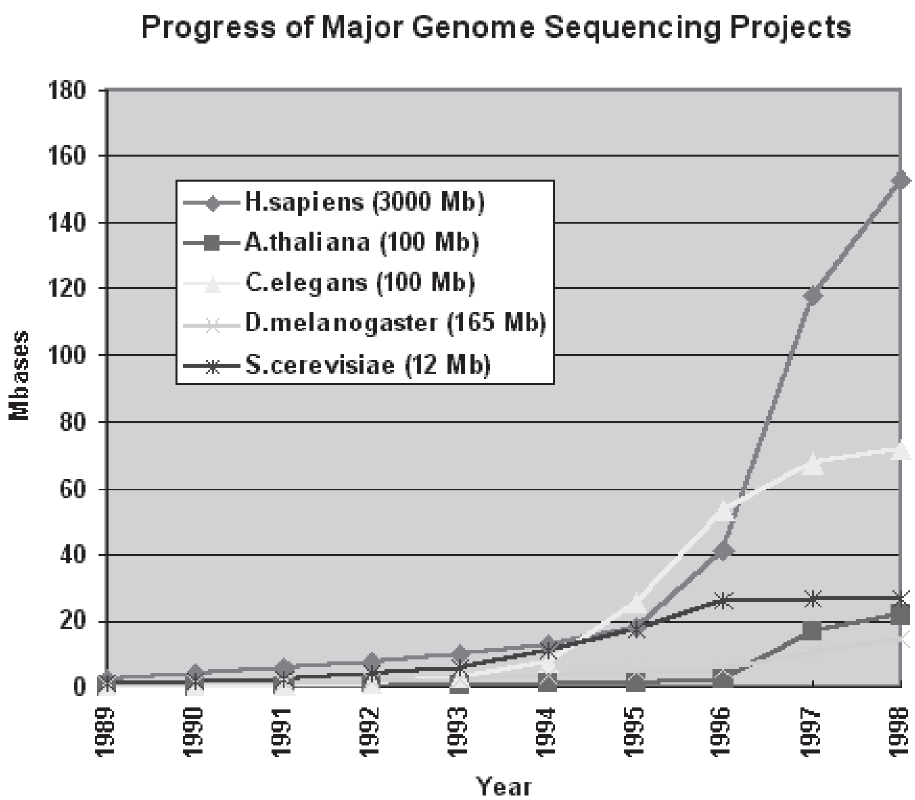
We have established a monitoring system which provides up-to-date status reports for the major genome sequencing projects. We know that the amount of sequence data in the EMBL/Genbank/DDBJ database is currently growing at an exponential rate, but without exponential growth, the year 2005 target for completion of the human genome sequencing project will not be met. Whatever they say the completion dates will be, don't just believe it until you have seen the Genome MOT!

Literature

Beck, S. and Sterk, P. (1998). Genome-scale DNA sequencing: where are we? Curr. Opin. Biotechnol. 9,116-120.

Stoesser, G., Moseley, M.A., Sleep, J., McGowran, M.G., Garcia-Pastor, M.P. and Sterk, P. (1998). The EMBL nucleotide sequence database. Nucl. Acids Res. 26, 8-15.

Grillo, G., Attimonelli, M., Liuni, S., and Pesole G. (1996). CLEANUP: a fast computer program for removing redundancies from nucleotide sequence databases. CABIOS 12, 1-8.



Upcoming Conferences

Objects in Bioinformatics '98

Object-Oriented Technology, Software Components and Distributed Computing for Bioinformatics and Genomics. 3rd and 4th August, 1998. Wellcome Trust Genome Campus, Hinxton, near Cambridge, UK.

URL: http://www.ebi.ac.uk/oib98/

Bioinformatics Workshops at the EBI

General

The objective of these workshops is to provide advanced training in the field of bioinformatics. Their target audience is professionals working in a field in bioinformatics, either at the research level or in the provision of services, and they are open to such professionals in both academia and industry. All workshops will be conducted in English, and be limited to a maximum of twenty participants.

Workshop programme

- SRS (Sequence Retrieval System) 28 September - 2 October, 1998
- Databases in Molecular Biology 12 - 16 October, 1998
- Protein Structure Prediction 28 - 30 October 1998

URL: http://www.ebi.ac.uk/info/eu.html

First International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'98)

Organised by the Institute of Cytology and Genetics, Novosibirsk, Russia , venue Novosibirsk Science Center-Altai Mountains (Russia).
24 August 1998 - 31August 1998

URL: http://bgrs.bionet.nsc.ru/

International Conference on Genes, Proteins & Computers V (GPC-V)

Bioinformatics, Networking and Computing in Molecular Biology Organised by CCP11 Project, venue York University, York, UK.
14 September 1998 - 16 September 1998.

The GPC_V Conference is the fifth in this highly successful series of biannual international conferences. The conference is not intended for computing experts but to help biologists to get the most from using available databases and software. A list of topics is given below.

Plenary Sessions:

Comparative Genomics
Inferring Function From Sequence
Bioinformatics Initiatives
Inferring Function From Structure

The registration fee is 40 GBP exclusive of accomodation. The Conference Organisers have reserved limited accomodation at York University. It is therefore advisable to register early.

The Conference will include a Poster and Commercial Exhibitions. Contributions for the Poster session are welcome on all topics covered by the Conference. The deadline for submitting an abstract is 14 August 1998. Three authors will be selected by the programme committee to give an oral presentation of their work.

URL : http://www.dl.ac.uk/CCP/CCP11/conferences/gpc_v/

2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98)

Organised by University of Nantes, venue University of Nantes, France.
23 September 1998 - 26 September 1998

Data Mining and Knowledge Discovery in Databases (KDD) have emerged from a combination of many research areas: databases, statistics, machine learning, automated scientific discovery, inductive logic programming, artificial intelligence, visualization, decision science, and high performance computing.

URL: http://www.sciences.univ-nantes.fr/pkdd98/

German Conference on Bioinformatics (GCB'98)

Organised by University of Koeln and others, venue Kardinal-Schulte-Haus, Koeln, Germany.
7 October 1998 - 10 October 1998

The conference will present the current research in the theory and application of computational methods to genomic, sequence, and structural information.

Topics include:

- Comprehensive Genome Analysis
- Comparative Genome Analysis
- Protein Structure and Function
- Drug and Protein Design
- Simulation of biological Molecules
- Metabolic Networks

URL: http://www.uni-koeln.de/math-nat-fak/biochemie/gcb98.htm

Functional genomics - the new biotechnology

The Lundberg Institute, Göteborg University, Sweden
27 August 1998 - 28 August 1998

The Organising Committee welcomes you to take part in the symposium about Functional Genomics where this novel concept will be presented to interested parties from academia and industry. Invited international speakers will discuss new technological advances as well as the use of genetic model organisms for elucidation of biological pathways and gene interactions in humans and plants.

URL: http://www.gbg.teknikbro.sshn.se/fg/

An overview of IT resources and methodologies applied to molecular biology.

University of Oxford, Oxford, England
14 July 1998 to 17 July 1998

An intensive course of lectures, practicals and demonstrations giving a sound theorectical and practical overview of IT resources and methodologies applied to molecular biology, especially sequence analysis.

URL: http://www.conted.ox.ac.uk/courses/biosciences.html

Peter Rice interviews Christoph Sensen

Associate Research Officer at the Institute for Marine Biosciences, Halifax, Nova Scotia in Canada.

Peter Rice
Hi Christoph. I remember we were both at EMBL in the early days of automated sequencing. What are you doing now?

Christoph Sensen
Those were the good old days! I am now an Associate Research Officer at the National Research Council of Canada, at the Institute for Marine Biosciences. I am working on genomics and bioinformatics, which are part of our Institute's core mandate. NRC has 18 Institutes in Canada, IMB is one of them. Each Institute has core mandates for the entire country. I am also an Adjunct Professor in the Biochemistry Department of the Medical School at Dalhousie University. I teach bioinformatics (a very low key activity right now, this is still in progress).

Peter Rice
Why did you move to Canada?

Christoph Sensen
I was interested in sequencing entire genomes. Especially in sequencing an archaeal genome. When I got the job offer from Canada, I saw my chance to do what I was interested in. At EMBL I was part of the yeast genome sequencing effort, and that meant to sequence between 25 and 50 kbp/year, something that we do at IMB in a week.

Peter Rice
Do you get to Europe often these days?

Christoph Sensen
Yes indeed, I have just returned from Germany... I travel a lot to give talks and interact with University colleagues and industrial partners. I already made it to Japan this year, and I will return to Europe at least two more times before the end of the year, once for the EMBnet meeting, and once to go to a Bioinformatics Conference in Cologne.

Peter Rice
What are the attractions of EMBnet for centres outside Europe?

Christoph Sensen
EMBnet has an extremely good reputation in Canada. When I came to Canada, there was a bioinformatics "vaccuum". We had to use services in the US or Europe for almost anything. To mirror these services in Canada is one of our goals. We consider EMBnet one of the key bioinformatics organisations worldwide.

Peter Rice
Do you have contacts with other bioinformatics organisations?

Christoph Sensen
Yes, we are a member of APBionet (Asian Pacific Bionet). Canada has strong ties to Singapore and we also have good connections with Tim Littlejohn in Australia (he worked in Montreal before he took over ANGIS).

Peter Rice
How does bioinformatics in Canada compare to other countries?

Christoph Sensen
This question I could probably answer better after next week. Canada is hosting ISMB-98 next week, we have over 450 participants already, almost 100 posters and close to 40 talks. Bioinformatics in Canada is still small, but it is constantly growing. There are many Canadians working abroad, and some of them are now getting a chance to move back to Canada and work in their own country.

Peter Rice
How is the Sulfolobus project going?

Christoph Sensen
The Sulfolobus project, which was my original reason to go to Canada, has now generated more than 80% of the 3 Mbp sequence. We hope to have most of the sequencing done at the end of the year. We have three European partners now, one in Denmark, one in France and one in the Netherlands. That has helped to complete the project tremendously.

Sulfolobus solfataricus P2 is a thermophilic and acidophilic crenarchaeote. Its genome is 3 MBp in size, all in a single circular molecule. The G+C content is only 37%, so the genome is relatively easy to sequence. There are many plasmids and viruses characterized for Sulfolobales, that will help with generating shuttle vectors and such.

We are sequencing Sulfolobus on a cosmid by cosmid and lambda by lambda basis. We have to do this because of the high number of repetitive regions. We are a collaboration of six labs right now, three in Europe and three in Canada. 80% of the sequence is already produced and the rest is coming in fast, so we should have most of the genome at the end of the year.

Peter Rice
Are you working on other genome projects too?

Christoph Sensen
A little bit here and there right now, but soon we will have a new genome initiative in Canada. Then we will enter into more genome projects.

Peter Rice
Do you have many successful international collaborations?

Christoph Sensen
Yes, I do. The foremost one is the MAGPIE collaboration with Terry Gaasterland from Argonne National Laboratory (soon Rockefeller). I am working with Terry since several years on automated genome analysis and annotation and we have published several papers on that.

Peter Rice
If you could make a wish, what bioinformatics development would you most like to see?

Christoph Sensen
Hmm... so many things come to mind 8-) I would like to see more integrated systems being developed and I would like to see this reflecting on the public domain. So many good things are now behind walls because the developers get hired into companies faster than they can generate code... How about going back to the old approach where scientists could use software for free and companies would pay? Lots of good research would come out of that!

Peter Rice
Many thanks Christoph for a most interesting interview, and I think you will like the EMBOSS article in this issue!!

Useful URLs:

Sulfolobus: http://niji.imb.nrc.ca/sulfolobus
MAGPIE: http://www.mcs.anl.gov/home/gaasterl/magpie.html
CBR-RBC: http://www.cbr.nrc.ca
Genomes around the world:
http://www.mcs.anl.gov/home/gaasterl/genomes.html

Node Focus

Amos Bairoch

The Swiss Institute of Bioinformatics

Background

Biomedical research, an information-based discipline, is undergoing a major revolution as novel experimental approaches are yielding unprecedented amounts of data.
Indeed, automation and robotics are becoming integral parts of the experimental process, impacting on the way both academic and industrial research is carried out. In this context it is not surprising that experimental biology and medicine are increasingly dependent on the extensive application of information sciences. Bioinformatics, the interdisciplinary field at the intersection of life and information sciences, provides the necessary tools and resources for this endeavour. Modern fundamental and applied research in the life sciences is critically dependent on this relatively new discipline.

Switzerland, through its academic and industrial skills has always been a key player in biotechnology. In the last decade, bioinformatics research groups in Geneva and Lausanne have developed unique competencies which are ideally suited to support and complement the current and future biotechnological efforts in Switzerland and elsewhere.

It has been emphasised by Swiss scientific authorities that it is now essential and urgent to promote the creation of "centres of excellence" in interdisciplinary domains that are economically important and crucial for tomorrow's society. This is why we created the Swiss Institute of Bioinformatics (SIB).

The goals of the institute are:

• To promote the development of research tools and databases in the field of bioinformatics;
• To offer services to the Swiss scientific user community through the Swiss EMBnet node (which was maintained jointly by the Swiss Institute for Cancer Research and the University of Geneva);
• To provide, in collaboration with academic partners, a curriculum of courses and seminars for the formation of research scientists in the field of bioinformatics.

In summary the goal of the Institute is to promote bioinformatic activities in such a way as to maintain and develop the leadership of Switzerland in this field.

Legal background

The Swiss Institute of Bioinformatics is an academic institution established as a non-profit foundation under the statute of article 80 of Swiss civil law. The bylaws of this foundation are compatible with the provisions regulating institutional funding by the Swiss Federal Government.

Currently, the Institute is tightly associated with the University of Geneva, the University of Lausanne, the Swiss Institute for Cancer Research (ISREC) and the Ludwig Institute for Cancer Research. It also operates in association with industrial partners, in particular Glaxo-Wellcome.

Structure

The activities of the Institute are organised around several research and service areas, each of which is headed by a group leader. Each group leader is an established scientist holding an academic position in one of the partner institutions.

An executive council (Bureau de l'Institut), consisting of the group leaders, has operational responsibility for the Institute. The director of the Institute (currently Victor Jongeneel) is elected for a two-year term among the members of that council.

An international Scientific Advisory Board, consisting of scientists with a well-established international reputation, will guide SIB in its choice of research and service activities.

The Foundation Council represents the interests of the various academic and industrial partners of the Institute, and is the governing body of the Institute.

The institute currently consists of five groups headed by:

• Ron Appel, director of the Molecular Imaging and BioInformatics Laboratory at the Geneva University Hospital. He is responsible for the development of the Melanie software package for 2D-PAGE analysis and the SWISS-2DPAGE database. He is also a co-developer of the ExPASy WWW server and leads the MARVIN research project, a multi-agent information retrieval system.

• Amos Bairoch, group leader at the Department of Medical Biochemistry of the University of Geneva. He is responsible for the development of the SWISS-PROT, PROSITE and ENZYME databases; the PC/Gene sequence analysis software and is also a co-developer of the ExPASy WWW server.

• Philipp Bucher, group leader and head of the Biocomputing Unit at ISREC. He is responsible for the development of the Eukaryotic Promoter Database (EPD), and has pioneered the use of profile-based methods for the description of protein and DNA domains.

• Victor Jongeneel, manager of the Swiss EMBnet node. He is responsible for the provision of services and courses in bioinformatics for academic and industrial users in Switzerland. He maintains the Swiss EMBnet Web server. He is director of the Office of Information Technology for the Ludwig Institute (world-wide).

• Manuel Peitsch, director of Scientific Computing (world-wide), Glaxo-Wellcome Research and Development. He is responsible for the development of the SWISS-MODEL server, as well as the SWISS-3DIMAGE, SWISS-3DMODEL and CD40Lbase databases. He is also leading the development of the Swiss-PDBViewer, a sequence to structure workbench for proteins, and participates in the development of the ExPASy WWW server.

The proposed group leaders are acknowledged scientists in their fields, totalling well over 300 publications in first rate scientific journals. The databases and software they have developed during the last decade are used by most academic and industrial life science laboratories world-wide. The proposed group leaders, and their respective groups, already have a long history of interdisciplinary collaborations exemplified by their collective development of the World Wide Web servers ExPASy and EMBnet-Switzerland.

The expertise and activities of these scientists, and of their current groups, cover a broad range of disciplines necessary to adequately set up a successful bioinformatics institute.

It is expected that, in a later phase, additional groups could be created. The Institute is expected to attract key researchers in the field of computational biology. Several world-class scientists have already expressed their interest in joining the Institute, which would enable them to closely collaborate with the above groups listed.

In addition to the scientific staff of each group, the Institute has a common administrative staff (currently an administrator and two secretaries) and system management personnel.

Teaching activities

The group leaders are the organisers and lecturers of a series of pre- or post-graduate courses given at the Universities of Geneva and Lausanne, as well as at the Swiss Federal Polytechnic School in Lausanne.

In order to attract students to this promising new field, the Institute will develop and teach a "Certificat en bio-informatique", in accordance with the guidelines of the local academic institutions. This pre-graduate course will include theoretical and practical sections and should span a full semester. The goal of this curriculum is to train individuals to become bioinformaticians and should allow students to master the latest bioinformatics tools and at least one recent computer language.

Furthermore, the Institute will provide introductory and advanced courses on selected aspects of bioinformatics to post-graduate students and scientists of Swiss academic institutions. Industrial participants will be expected to defray the costs of the course.

Major activities and research areas

All of the group leaders in SIB are continuing to develop and enhance their current research activities as well as the many collaborative projects in which they are already involved. The grouping of five groups in a single organisation has fostered unprecedented synergy. Thanks to a close collaboration with industrial partners (see section on financial considerations) we expect a scaling up of the activities toward the development of integrated databases and software resources in the field of Proteomics. This will ensure that the prominence of the Swiss research and development bioinformatic activities is strengthened.

Intellectual properties considerations

Databases and software developed by the groups at SIB will remain the property of the Institute. Academic users will continue to have free access to any of these products that are partially or completely funded by public grants. The institute will seek to license its products to commercial users. It should be noted that SWISS-PROT, which has been a joint venture between the group of Amos Bairoch and EMBL/EBI for many years, will continue to be so. EMBL/EBI will share in any profits made from the licensing of SWISS-PROT.

The institute is also expected to carry out research or development under industry mandates. The results of these activities are the property of the mandating parti(es).

Financial considerations

The salaries of the group leaders are provided either by the academic institutions participating in the foundation of SIB or by industrial partners.
The salaries of the other members are financed through a mixture of academic and industrial grants as well as from the licensing of products created by the Institute to Geneva Bioinformatics (GeneBio). This company, which has been created in Geneva to promote technology transfer in bioinformatics, will finance a part of the Institute's activities through the commercialisation to for-profit organisations of the databases and tools developed at the Institute.

Where is the SIB located?

As a bioinformatics institute does not require a centralised infrastructure, the SIB is located in both Geneva and Lausanne. In Geneva, space has been allocated for the SIB within the University Medical Centre (CMU). In Lausanne, the SIB is associated with the Laboratory Centre in Epalinges (where ISREC and the Ludwig Institute are located). Its new office space will be ready in the fall of 1998.
The Institute administrator is currently based in Geneva, but also has an office in Lausanne. Secretarial staff and computer equipment are distributed between both sites.

How about EMBnet?

The SIB, which now regroups the activities of the Swiss national node and the SWISS-PROT special node, has become the sole academic EMBnet node in Switzerland. We expect that the Institute will be a strong and active partner in EMBnet activities, in spite of the isolation of Switzerland from its European neighbours.

Real Servers

Rob Harper

Have Technology will Travel

Remember Gopher? Of course you don't. That is old hat, dead technology, and as soon as it was superceded by the WWW it just faded away and was rarely used again. It seems that biologists are ever mindful of what is going on in the mainstream of computer science, and then latch on to the existing technology and modify it to meet their specific needs. WAIS indexing of GenBank was a first for Don Gilbert wasn't it?

One might ask the question where do we go to after the WWW, and the answer to that question is really a question of bandwidth. If you mention multimedia over the network then everyone throws their hands up in horror, for they envision video where people make jerky puppet like movements, and speak as though they were auditioning for a Dalek on Dr Who. But the adventurous are begining to use RealServers which deliver both audio and video.

Have bandwidth will listen

So what do you need to see an educational RealServer in action? Well you need a RealPlayer. This is an application program that functions as a plugin for your WWW browser. You can get it for free. It is a very simple task to install a RealPlayer as a plugin. Simply follow the next few links and you will be up and running in no time. I would suggest that you go for the audio files first of all to test that everything is working correctly and after that if your bandwidth permits then try a video/audio connection.

Getting up and running

- Dowload the free RealPlayer
- System requirements for RealPlayer
- Install the RealPlayer as a Plugin for your Web Browser
- Try out your RealPlayer of a few sites
  - Audio: BBC Five Live for people this side of the pond
  - Audio: CBC Radio Two Live for people on the other side of the pond
  - Audio: BBC Discovery 30'
  - Audio: BBC SoundByte The Millenium Bug
  - Video : The controversy over silicon breast transplants
  - Video: Missed the first live birth on the internet?

The BBC have invested quite heavily in the use of RealAudio and if you want to keep up with what is happening in science and education then take a look at the following sites.

Science	Education
Science in Discovery The Lab One Planet	Action Plants of power The green World Body of knowledge Major Killers

Distance Learning

It would appear that a few universities in the USA are providing a "Virtual Curriculum". One of the nicer presentations is from Penn State University who are able to deliver a multimedia presentation which involves a video/ audio of a lecture on Ventilation-Perfusion Relationships , which is just a smart way of saying why do we get out of breath when we run, or climb a high mountain.

The RealPlayer will display the video of the lecture as it was recorded and, in addition, the Web Browser will display the slides used during the lecture. Everything is synchronised so that if, on the video, the lecturer changes a slide then the same slide is shown on the Web Browser.

Obviously the intended audience are students who attend Penn State university. I can imagine that on a fast LAN the video/audio service is quite acceptable. From the UK the video is rather patchy but the audio is near perfect. I expect that in the future many universities will have their introductory lectures broadcast via a Webcast, and initially it will be local students who reap the benifits of this technology. However as bandwidth increases, and technology progresses, then RealServers will open up a new window of opportunity in distance learning.

Europe seems to be lagging behind when it comes to providing sites that are investing in Real Servers in biology, but there are no lack of sites thoughout Europe that make use of the technology to provide audio services. Indeed the record industry is becoming worried that in the future the internet will be the main method for thedistribution of music... then how will they make their money.

Classroom servers

It would seem that every scientist who makes a presentation these days uses PowerPoint. They tell about their research, they talk about their Institute and many worthwhile presentations never see the light of day other than at a conference or some internal seminar. Slides are made, transparancies are photocopied but after a while they get thrown to the back of a cupboard and forgotten. It seems such a waste when it is a relatively simple task to convert these presentations and put them on the web.

There is a tool called RealPresenter Plugin which will work with PowerPoint to convert a slide show so it can be viewed with RealPlayer. I could easily envision a set of PowerPoint slides which concentrate on Software Tutorials on the major applications programmes running on a WEB site that would take the novice user through the steps to run a FASTA or BLAST job.

Or if you want to keep it simple without any audio then you and just use PowerPoint itself to generate a presentation that is suitable for the web.

Examples

- EMBL database overview.
- SWISS-PROT and TrEMBL
- Data mining and visualistaion of yeast gene expression data
- Protein Function and Biochemical Pathways

Well I hope that gives someone an insight into what can be done with real servers, and perhaps in the near future we will see more Real Servers dedicated to Biology.

EMBnet Node News

The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

Peter Rice

Through the 100 Megabase barrier:

On June 9th, the Sanger Centre reached 100 million bases of finished sequence data. All our finished data is available as fully annotated EMBL entries.

The unfinished (in progress) sequences are currently running at over 85 million bases, and is available through our FTP server and is submitted to the EMBL HTG division daily.

Details by organism are available on:
http://www.sanger.ac.uk/Info/Statistics/

The Sanger Centre is committed to producing 1000 Mbase of finished human sequence, one third of the human genome, in addition to sequencing major pathogens, the nematode Caenorhabditis elegans and other major projects.

Tuberculosis sequenced:

The genome of Mycobacterium tuberculosis H37Rv has been completed at the Sanger Centre in collaboration with the Institute Pasteur.

Tuberculosis infects about a third of the world's population and kills 3 million people each year. H37Rv is the best characterised laboratory strain of this pathogen. The full paper appeared in Nature on June 11th; Cole et al. (1998) Nature 393:537-544.

The Sanger Centre TB project Web pages are at

http://www.sanger.ac.uk/Projects/M_tuberculosis/

The sequence is available in EMBL/GenBank/DDBJ, under accession number AL123456.

EMBnet China Node, Centre of BioInformatics, Peking University Beijing, China

Jingchu Luo

An EMBnet course was held in early April this year at our node. The aim of this course was to give Chinese molecular biologists a brief introduction to bioinformatics,and to provide selected participants a chance for hands-on practice on software and database tools to analyse DNA and protein sequences. This was the first activity since we were accepted as a member of the EMBnet in November 1996. The course was strongly supported by various EMBnet nodes and local authorities.

Five EMBnet teachers, Alan Bleasby from SEQNET (UK), David Judge from Cambridge University, Thure Etzold from EBI, Jack Leunissen from CAOS/CAMM Center (the Netherlands) and Frank Wright from Biomathematics & Statistics (Scotland, UK) started to prepare the course in February by remote login to our server during the weekends. They started to work as soon as they arrived in Beijing on Saturday, 4th April.

The course was divided into three sessions. The first session was a one-day lecture with some 200 participants. It started with an opening ceremony. Talks started from the introduction to EMBnet by Alan Bleasby, and moved to various topics such as the current progress of bioinformatics by Jack Leunissen, genome infomatics by David Judge, molecular databases by Thure Etzold and molecular evolution by Frank Wright.

Following was a three day hands-on practice session with 45 biologists. Sequence analysis packages GCG and Staden were the major resource of this session. The participants were interested in learning lots of tools. An SRS session for another three days started from the fourth day. The students were lucky to have Thure Etzold, the developer of SRS, to answer their specific questions, to get into the depth of this powerful database query system.

A five day structure based workshop was also held on 2-6 June. Gert Vriend came to Beijing to give hands-on training for 12 students on his WhatIf package. He also gave a lecture on homology modelling for 30 biologists and a talk on biocomputing for 60 undergraduate students during the workshop.

EMBnet visits China

Alan Bleasby UK National EMBnet node

A group of EMBnet people from various countries have recently returned from giving a training course in China (one of EMBnet's more recent nodes.) We didn't know what to expect before we set off but we were excited. You might say that travelling within Europe you see a lot of different cultures; indeed you do. The cultures are not fundamentally inconsistent though. You might say the same about the USA or Australia. We wanted to see an entirely new culture and we were not disappointed. It was wonderful.

As I was one of the speakers in the first of the two weeks we stayed in China my first impressions were limited to the Beijing Campus of peking University and the roads. The campus is delightful. It was originally a private garden in the Qing Dynasty and is now a sight-seeing site with a small, yet elegant lake - the Unnnamed Lake. Walking along the Campus in the early morning is very enjoyable. You listen to the songs of the happy birds in the trees. At night you can find the trees turing green. You can smell the fragrance of the flowers. You can see old women playing Tai Ji boxing on the island and young boys jogging along the lake. There is no problem finding a chance to take photos of pretty and sweet ladies. My film rolls will prove it if you don't believe me :-).

The roads are another story! To give you a flavour I'll put this in a tongue-in-cheek fashion. There have been many poster campaigns in the UK. The "coughs and sneezes spread diseases" one did little to stop the Great Plague in London in 1665; similarly the "wear something white at night" campaign one was less than relished by someone who was subsequently hit by a snow-plough.

With respect to the roads there is a UK rule that you walk facing oncoming traffic. In China it is different as you walk on the right and have cars and bicycles coming up behind you. You have to keep your eyes open at all times and a stiff neck is to be expected in the first few days.

If you decide to ever visit China (and I strongly recommend you do given the chance) then, depending on where you stay, be warned of traffic noise. It is not the sound of the car engines that keeps you awake but the fact that every car sounds its horn every few seconds. Being driven along the roads of Beijing for two weeks makes dodgem cars seem normal. I'm assured there are some rules but I never managed to spot them. The crowdedness of the Beijing roads makes the M25 around London seem like a quiet country lane. There is no lane discipline and you beep your horn and move into the lane that seems to be moving the quickest. It really gets the adrenalin pumping.

The course itself was a great success. The lecturers were:
- David Judge: Cambridge
- Frank Wright: Dundee
- Jack Leunissen: The Netherlands (CAOS/CAMM Nijmegen)
- Thure Etzold: EBI
- Alan Bleasby: Daresbury

A number of people took their family and/or girlfriends along! We all found time to see most of the sights, though the course itself was really a big job.

I have over 200 photos of China some of which are scanned and appear in this edition. We were well looked after by our host (Jingchu Luo) and certainly wish to be invited back.

So, we all saw the sights: The forbidden city, The Summer palace, The Ming Tombs, Beijing itself and many more. We also all have our sweat-shirts saying "I climbed the Great Wall." We did! All 8 million steps! It turns out that you're not a real man or woman unless you've climbed the wall and also eaten Peking Duck. The latter was a problem for our vegetarian contingent but they ate something similar. We obviously didn't do the whole of the Great Wall but it seemed enough at the time.

The China node of EMBnet is a very welcome part of our whole. It was a great honour to visit it.

German EMBNet node, GENIUSnet at the German Cancer Research Centre (DKFZ) in Heidelberg.

Martin Ebeling

Here are the latest news from the German EMBnet node, GENIUSnet, at the German Cancer Research Centre (DKFZ) in Heidelberg.
WWW site - The GENIUSnet homepage has moved and while the old address should still work, we suggest to use the new one at http://genius.embnet.dkfz-heidelberg.de:8080/menu.

Courses - Demand for introductory courses as well as for courses focused on specialised topics is growing among our users. Whilst we offer introductions to our HUSAR package on a regular basis in Heidelberg, several courses were held in other German cities where eight to ten participants could come together.

New co-workers - Two new co-workers have joined our team recently. Dr. Mechthild Falkenhahn, a biologist by training, will mainly be concerned with holding courses and preparing new material for courses and tutorials. Thus we hope to keep up with developments in bioinformatics and to be able to offer our users not only access to new programs, but also to background information and practical help. Peter Ernst, a physicist, had been working in our group already as a graduate student. He is now a full-time employee and will be responsible for web programming and the further development of our WWW2HUSAR interface, in close collaboration with Martin Senger at EBI.

Software - New entries to HUSAR include a GCG-adapted version of Sean Eddy's HMMER package. In addition, we have developed an HMMSCAN program for searching a HMM database with a query sequence. We have extracted on the order of 8000 sequence alignments from the HSSP database and built a database of the correpsonding HMMs which can now be readily searched using HMMSCAN. Also newly introduced into the package was PUZZLE, a program for phylogenetic analysis using the quartet puzzling algorithm developed by Arndt von Haeseler (University of Munich). For a brief introduction, follow the NEWS link on our homepage mentioned above.

Location - The German Cancer Research Centre has grown to eight storeys last year, and the Department of Molecular Biophysics has moved to the top of the building. Thus, we are now enjoying not only our work but a wonderful view over Heidelberg, the Castle and the Neckar valley.

Russian EMBnet node GeneBee

Leonid Brodsky

Leonid Brodsky is leaving GeneBee, or more correctly he is in the process of emigrating from Russia to Israel. As Genebee's founding father he will still participate in our site's activity from there. His duties have been transferred to two persons. Sergie Spirin (sas@ genebee.msu.su) will be responsible for the management of the Russian EMBnet node and Yannis Kalaidzidis for technical management and new software technology development.

The EMBnet Nodes

National nodes:

[AT] EMBnet martin.grabner@cc.univie.ac.at
BioComputing Centre,
Vienna, Austria

[BE] BEN rherzog@ulb.ac.be
Universite Libre de Bruxelles
Sint Genesius Rode, Belgium

[CH] ISREC Victor.Jongeneel@isrec.unil.ch
ISREC Bioinformatics Group
Epalinges, Switzerland

[DE] Genius m.ebeling@dkfz-heidelberg.de
DKFZ
Heidelberg, Germany

[DK] BIOBASE hum@biobase.aau.dk
BioBase
Aarhus, Denmark

[ES] CNB carazo@samba.cnb.uam.es
Centro National de Biotecnologia
Madrid, Spain

[FI] CSC erja.heikkinen@csc.fi
Centre for Scientific Computing
Espoo, Finland

[FR] Infobiogen dessen@infobiogen.fr
Infobiogen
Villejuif, France

[GR] IMBB savakis@nefeli.imbb.forth.gr
Insitute of Molecular Biology
Heraklion, Greece

[HU] HEN embnet@hubi.abc.hu
Agricultural Biotechnology Centre
Godollo, Hungary

[IE] INCBI atlloyd@tcd.ie
Irish National Centre for Bioinformatics
Dublin , Ireland

[IL] INN lsestern@wiezmann.weizmann.ac.il
Weizmann Institute of Science
Rehovot, Israel

[IT] CNR marcella@area.ba.cnr.it
Consiglio Nationale delle Ricerche
Bari, Italy

[NL] CAOS/CAMM embnet@caos.camm.nl
Caos/Camm Centre
Nijmegen, Netherlands

[NO] BiO linda.akselberg@bio.uio.no
Biotechnology Centre of Oslo
Oslo, Norway

[PL] IBB piotr@ibbrain.ibb.waw.pl
Institute of Biochemistry and Biophysics
Warsawa, Poland

[PT] PEN pfern@pen.gulbenkian.pt
Instituto Gulbenkian de Ciencia
Oeiras, Portugal

[SE] EMBnet.se embnetadm@perrier.embnet.se
Biomedical Centre
Uppsala, Sweden

[SU] Genebee libro@brodsky.genebee.msu.su
Belozersky Institute of PhysicoChemical Biology
Moscow, RussiaX

[UK] SEQNET ajb@dl.ac.uk
DRAL Daresbury Laboratory
Daresbury, England

Specialist nodes:

[CH] SwissProt bairoch@cmu.unige.ch
Dept Medical Biochemistry
Geneva, Switzerland

[CH] Roche daniel.doran@roche.com
Hoffman-LaRoche
Basel, Switzerland

[DE] MIPS mewes@mips.embnet.org
Max Planck Institut fur Biochemie
Martinsried, Germany

[IT] ICGEB pongor@genes.icgeb.trieste.it
International Centre for Genetic Engineering
Trieste, Italy

[UK] EBI stoehr@ebi.ac.uk
European Bioinformatics Institute
Hinxton, England

[UK] HGMP-RC mbishop@hgmp.mrc.ac.uk
HGMP Resource Centre
Hinxton, England

[UK] Sanger pmr@sanger.ac.uk
Sanger Centre
Hinxton, England

[UK] UCL attwood@bsm.bioc.ucl.ac.uk
University College London
England

Associate nodes:

[AR] IBBM grauu@biol.unlp.edu.ar
Instituto de Bioquimica y Biologia Molecular
La Plata, Argentina

[AU] ANGIS tim@angis.su.oz.au
Australian National Genomic Information Service
Sydney, Australia

[CN] CCB luojc@lsc.pku.edu.cn
Peking University
Beijing, China

[CU] CIGB bringas@cigb.edu.cu
Centre for Genetic Engineering and Biotechnology
Habana, Cuba

[IN] CDFD cdfddbt@hd1.vsnl.net.in
Centre for DNA Fingerprinting and Diagnostics
Hyderabad, India

[SE] Upjohn mats@inddama.sto.se.pnu.com
Pharmacia-Upjohn AB
Stockholm, Sweden

[ZA] SANBI winhide@techno.sanbi.ac.za
South African National Bioinformatics Institute
Bellville, South Africa

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 in the Tips from the computer room section, please let us know. Submissions for the BITS section are most welcome, but please remember that we cannot extend space beyond two pages per article. Please send your contributions to one of the editors. You may also submit material by Internet E-mail to:

emb-pub@dl.ac.uk

You are invited to contribute to the LETTERS TO THE EDITOR section.

If you had difficulty getting hold of this newsletter, please let us know. We would be only too happy to add your name to our mailing list. This newsletter is also available on-line using any WWW client via the following URLs:

The Online version, (ISSN 1023-4152) :

- ***http://www.uk.embnet.org/embnet.news/vol5_2/contents.html***
- ***http://www.be.embnet.org/embnet.news/vol5_2/contents.html***
- ***http://www2.ebi.ac.uk/embnet.news/vol5_2/contents.html***
- ***http://www.ie.embnet.org/embnet.news/vol5_2/contents.html***

A Postscript version (ISSN 1023-4144) is available. You can get it by anonymous ftp from:

- ***ftp.uk.embnet.org in the directory pub/embnet.news/***
- ***ftp.be.embnet.org in the directory pub/embnet.news/***
- ***ftp.ebi.ac.uk in the directory pub/embnet.news/***
- ***ftp.ie.embnet.org in the directory pub/embnet.news/***

A pdf version (ISSN 1023-4144) in Acrobat 3 format is also available. You can get it by anonymous ftp from:

- ***ftp.uk.embnet.org in the directory pub/embnet.news/***
- ***ftp.be.embnet.org in the directory pub/embnet.news/***
- ***ftp.ebi.ac.uk in the directory pub/embnet.news/***
- ***ftp.ie.embnet.org in the directory pub/embnet.news/***

Back issues are available at most of these sites.

Publisher:

EMBnet Administration Office.
c/o Jan Noordik
CAOS/CAMM Centre
University of Nijmegen
6525 ED Nijmegen
The Netherlands

Editorial Board:

embnet.news
Vol.5, No.2, 1998
30 June 1998
ISSN 1023-4144