embnet.news

Volume 5 Nr4 (ISSN1023-4144)

December 30, 1998

Editorial

It is the end of 1998 and its been an interesting year. EMBnet celebrated its tenth anniversary. The genome of Caenorhabditis elegans was finally finished, more or less on schedule. After several static years, a new gapped version of Blast was released. SRS and Swissprot both went commercial. Although commercialisation is often viewed as a "bad thing" by academics, it can have some very positive fall-out. If carried out sensibly it should guarantee income enough to ensure continuity. It should also provide the salaries of those who will continue to maintain and develop products which are useful to the entire community. Getting paid also puts an obligation on the product or service provider: to write that tedious but necessary manual and documentation, and to respond to users questions and suggestions quickly and effectively.

It is obvious to those of us in the EMBnet community that the infrastructure which EMBnet nodes provide is an essential part of world bioinformatics. Training, documentation, advice, consultancy, specialised and local databases, software development and hardware maintenance are all vital underpinnings of effective research. Unfortunately, those who use and appreciate the services that the nodes provide are rarely those responsible for supporting and funding them. So we have seen several disruptive "rationalisations" across our community in recent months. SEQNET, the United Kingdom National Node, has been united with the HGMP-RC, one of the Specialist EMBnet Nodes located on the Hinxton Genome Campus. Let us hope that this merger will be transparent to users and have all the positive benefits that the responsible funding agencies hope for and expect. A much less agreeable and well-planned change has occurred at the Netherlands EMBnet node. We understand that the manager of the EMBnet grant from the European Union, Jan Noordik, is on "indefinite leave of absence" while the role of the CAOS/CAMM Centre is reconsidered. It is to be sincerely hoped that such local goings-on will not have an adverse effect on EMBnet as a whole; EMBnet cannot and should not interfere in local arrangements. Jan has been a tireless supporter and promoter of EMBnet over the last decade. He deserves and will surely receive the support of the entire EMBnet community.

This issue of embnet.news gives us a seasonal overview of networks in a big country by Christoph Sensen. An interview with Peter Stoehr, Head of Database Operations at the EBI and at the heart of production of the big database. A new WWW-based multiple sequence manipulation tool is described by Michele Clamp and others from the EBI. Iseli and Jongeneel from the SIB introduce an improved blast server and client. Together with Node News, a book review and many of our regular features we have an issue suitable for holiday reading.

A happy new year to all our readers !

The Editorial Board.

Contents

INTERviewNET

Peter Stoehr of the EBI interviewed by Andrew Lloyd

1. I happen to know that you have a British passport but your name is not a common one in Britain. Does your family come from Cornwall?

Actually I don't have a passport at all at the moment - thanks for reminding me. Stoehr is a German surname - my father comes from the Silesia area but I was born in England.

2. How did you come to work at EMBL Heidelberg?

At University I studied Botany/Zoology and then a Masters in biometrical genetics, then ended up as a programmer in the AFRC (now part of BBSRC) first at the Plant Breeding Institute in Cambridge, then at the Computing Centre in Harpenden. It was the early days of sequence databases and

analysis software and I was given the task of implementing this stuff for the growing community of molecular biologists. But they didn't really pay enough, EMBL advertised and I went. I thought I could help the database be better for users and less painful to install. Still trying...

3. Would you consider yourself a "good European"?

I don't know, I've never met one though I do take seriously EMBL's (and EBI's) European status, and in areas like recruitment and collaborations try very hard to look beyond La Manche.

4. About how many times a month do you get head-hunted by the pharmaceutical industry offering you a telephone number for a salary?

Not as often as you, Andrew, I'm sure. I usually give them your name.

*5. As the man who announces the building of a new release of the EMBL DNA database you must be close to the heart of DNA database management and strategy. Perhaps you could answer a daily, monthly and annual question? What *happens* at the nightly interchange between EMBL GB and DDBJ of DNA sequence information? Does it ever require human intervention?*

Well it normally all happens in the dead of night, so noone is really sure what goes on, but next morning there is often some data lying in a pool of blood needing some attention. There has been a sequence-a-minute going into our database for the last couple of years, so the odd hiccup is not surprising. Basically we FTP batches of flat-file data from NCBI and DDBJ, parse it into our ORACLE database and then fix the problem cases somehow. Then, each night, we dump out all new data (from all sources) into daily updates and onto our FTP server. Thats pretty well automated by now. Independently of all that, we synchronise our taxonomy trees every few hours, and each week exchange various lists of accession numbers, identifiers and checksums to inform each other what our current status is.

6. Why does the building take so long?

What do you mean "so long"!? The building (of a release I suppose you mean) takes 7-10 days during which we export all 3 million entries from our ORACLE database into flat-files and check them as many ways as possible. Our ORACLE database is our definitive version of the data, it is here where we make all data changes. We often make global changes to the database (e.g. new line-types or data items, taxonomy, cross-references etc) which do not get propagated in daily updates but only appear when we do a complete flat-file dump at release time. We don't just munge together daily update files with the previous flat-file release. During the release building period we continue to pump out updates, so we are not delaying any data availability.

7. How many people are involved at the EBI? How does that compare with the numbers employed by your Japanese and Merkin competitors/collaborators?

We have about 9 biologists and data submission staff working on the EMBL Nucleotide Sequence Database, a constant number over the years, and a programming group of about 8 is shared with the SWISS-PROT/TREMBL work here and some other EBI database work. I think we have rather less on the job than our collaborators.

8. Will you be working on the 25th?

Yes, our systems for data submission and external services will be all working over Christmas.

*9. What *happens* at those annual meetings between the database providers?*

These are business-like meetings where we try to resolve annotation convention issues and sometimes syntax problems, how to address new classes of data as they arise (ESTs, GSS, patent data, whole genomes), taxonomy issues, interactions with journals so that we can automate data exchange, ensure we capture the same biological data and also develop the biological content. We try to do all these all the time, but a physical meeting helps nail down conclusions and committments.

10. Have you all visited many interesting places with white sand beaches, azure skies and really cold drinks?

We meet in turn at each others places, EBI, DDBJ and NCBI. I've seen azure skies and sand beaches in Japan, and really cold beer in Washington, but EBI is the most interesting.

*11. Is it more than inertia that maintains a very large and exponentially increasing database in two very different formats? Would it not be a simple matter to agree on the best features of each and give the community of users and software developers *one* (better) format?*

That ol' chestnut again... We've been much more successful recently at synchronising our data content to avoid the need for users to install both databases locally. Do you today see data available in GenBank but not in EMBL?

12. You were around at the beginning of EMBnet. How has your view of the organisation changed over the last ten years? Should all EMBnet Nodes maintain a local copy of EMBLorGBorDDBJ despite the excellent homology servers at the EBI and NCBI and despite the cost of distributing, storing and maintaining such an enormous (10GBytes in

size and doubling every 9 months) database?

We had some fun trying to squeeze a few sequences down the wires in the early days, and the project started with an enthusiasm to develop and implement new things together. I think EMBnet got rather bogged down in organisational matters, but there are still good people and nodes involved to realise more of the potential. EMBnet nodes answer to their local users and have to judge for themselves whether they need local databases or can rely on an external provider. I don't think 10Gb is enormous, but there is scope for co-ordination for several established computational services (e.g. FASTA, BLAST, BIC_SW) to establish a load-shared redundant system operated by a smaller number of sites with some compute power to contribute. I feel a project proposal coming on...

13. Do you have a Christmas Message for EMBnet?

We've already announced that release 57 will be available on Christmas Day, 15:00 GMT.

Book review

Molecular Evolution: a phylogenetic approach. (1998) Roderic Page and Edward Holmes.
Publ. Blackwell ISBN 0-86542 889 1
24.95GBP/Pback. 346 pages

Reviewed by Andrew T. Lloyd, EMBnet Ireland

This book has an elegant picture of a Calder mobile on the front cover as an attractive metaphor of a phylogenetic tree, and a sequence trace on the back to illustrate that the authors are not merely theoreticians but know where the basic information comes from. In between is an excellent introduction to one of the central problems in biology: how do we explain the diversity of the living world and the interrelationships between its components.

Chapter 3 is a breathless gallop through the basics of molecular biology, at times rather like a "summarise Proust in two minutes" competition; a great clatter of bold keywords and concepts. Nevertheless, the summary is competent and comprehensive enough that a student could, for most questions, leave Genes V on the shelf. The next chapter is a similar overview of population genetics and it is admittedly handy to have both summaries in the one volume. Certainly, as the price of books is only minimally dictated by the number of pages, there is no loss in having these chapters.

The rest of the book -- chapters entitled Measuring Genetic Change, Inferring Molecular Phylogeny, Models of Molecular Evolution and Applications of Molecular Phylogenetics -- is what people will really be buying. It is quite easy to make the theory of molecular evolution difficult - a few acronyms and a lot of greek letters and Joe Biologist decides that a pint of plain, a packet of crisps and a UPGMA tree will be good enough for him. So, while Page and Holmes do give us the acronyms and the greek (scientific rigour dictates that they must) they are polite and careful enough to explain what they mean, put them in context and, most importantly, explain how ordinary folk might be able to use the theory in their own real world. Each chapter finishes with an executive summary which should not be seen to substitute for a close reading of the full text. This is followed by a page or so of recommendation for further reading.

As a regular biologist who has published phylogenetic trees in anger (and frustration), my confidence has increased immeasurably from reading this book. As one who professes to *teach* phylogenetic inference I am much happier now that there is a text book that I can confidently recommend to graduate students.

A Network Carol

How EMBnet Canada was Born

Christoph W. Sensen National Research Council of Canada Institute for Marine Biosciences 1411 Oxford Street Halifax, NS Canada B3H 3Z1

"Once upon a time in the Great White North, a country so far north of the United States of America that few Americans knew much about it, there lived a people whose scientists were, at best, connected to the Internet by 56 kbit lines." Unfortunately, this is not entirely a fairytale. Canadian scientists have all experienced the frustration of slow lines and many are still in that situation. In Canada, slow networks began to fade into history only as of April 1996 when the world's first GigaPOP (Gigabit Point of Presence) was opened in Halifax, Nova Scotia. Today, all major Canadian research institutions are connected to a network of GigaPOPs. The GigaPOPs are linked via an ATM network with a sustained data transfer rate of 7 Mbit/sec and a burst rate of 45 Mbit/sec. This network, called CA*net2, will serve as Canada's research and development environment for one more year, after which it will be replaced by CA*net3, a fiber optics network that will initially provide connectivity at a speed of 3 Gbit/sec.

For EMBnetters, the most interesting fact about CA*net2 and CA*net3 is probably that the test application and the

showcase demonstration project on these high-speed networks is bioinformatics. This may seem surprising but there is a certain history to this, thus we have to go back about three and a half years to the dreadful times when we were all on 56 kbit lines and start our story there.

"The Great White North is a huge country, spanning four and a half time zones with people living mainly along the coastline of two great oceans and along the shores of the largest inland lakes known to storytellers. Some of the people living in the GWN were scientists working in the 18 Institutes of the National Research Council. The Institutes were located in several cities from Newfoundland to Vancouver, thus people in one of the Institutes would be heading off for lunch when those in another were just showing up for work." Canada has very particular challenges when it comes to communication. The National Research Council of Canada, which is Canada's lead Federal Research organisation, operates 18 Institutes, five of which are focussed on Biotechnology. The NRC Biotechnoloy Institutes are located in Montreal (BRI, the Biotechnology Research Institute), Winnipeg (IBD, the Institute for Biodiagnostics), Ottawa (IBS, the Institute for Biological Sciences), Halifax (IMB, the Institute for Marine Biosciences) and Saskatoon (PBI, the Plant Biotechnology Institute). There are two additional NRC units that indirectly support Biotechnology R&D, (CISTI, the Canada Institute for Scientific and Technological Information, Canada's National Library in Ottawa and the NRC Innovation Center in Vancouver).

In 1995, scientists from BRI, IBS, IMB, PBI and CISTI held a series of meetings to discuss models of how to establish bioinformatics within the Biotechnology Institutes. Very early it was clear that this should be done as an inter-institute collaboration; a duplication of effort at each institute was highly undesirable. The service should be provided from a location that was actively doing R&D in genomics and bioinformatics; however, the scientists insisted that each institute should have the same level of services regardless of location. Having said that, it became evident that the existing 56 kbit lines would not be able to accommodate this requirement. The planned bioinformatics services would provide access to over 800 executables, including GCG, GDE, PHYLIP and the Staden package and more than 70 databases, from the entire EMBL distribution to OWL. X-applications would be launched remotely from servers in Halifax and displayed on machines in the other locations across the country. Ideally, hard discs would be cross-mounted between different cities and OS updates and upgrades would be installed remotely.

The scientists concluded that high performance networking was necessary to make that dream come true. There was no precedent for such a network, so people started dreaming about what could be. The first model was to connect all institutes via T1 lines at a speed of 1 Mbit. This turned out to be a very expensive proposition, each institute would have had to pay of the order of $40,000 Can/year for the connection. The bad news was quite discouraging for a little while, but soon there was new hope. The experimental network developers in Canada heard about our problems and we started to talk to CANARIE about their CA*net2 plans. CANARIE, the organisation that had previously implemented the Internet in Canada, was developing a new high-speed network, called CA*net2, which was planned as an ATM backbone throughout Canada. The ATM backbone would carry TCP/IP protocols at a sustained rate of 7 Mbit/sec with a burst rate of 45 Mbit/sec. CANARIE was looking for a demonstration application that could be used to showcase the new opportunities and at the same time could be used to debug problems in the network.

Until 1995, no scientific group had ever approached CANARIE to get access to high-speed networks, and the only application that was discussed as having a need for CA*net2 was videoconferencing. This situation was frustrating for CANARIE; thus, when the opportunity to help establish a high-speed bioinformatics network materialised, it was as exciting to them as getting high-level bioinformatics connectivity was for th scientists. The bioinformatics network was "adopted" by CANARIE and, from that point, their CA*net2 was developed in close collaboration with the bioinformatics network, which had become dubbed "The Canadian Bioinformatics Resource - Ressource de Bioinformatique Canada", or CBR-RBC. Initially, CBR-RBC was planned and implemented as an Intranet for NRC. Approximately 40 UNIX workstations and servers were purchased, configured and installed at BRI, IBS, IMB, PBI and CISTI and later also at the NRC Innovation Center.

CBR-RBC is managed by a User Committee, headed by David Thomas at BRI, and having members from each of the institutes participating in CBR-RBC. Two full-time computer administrators, Rob Hutten, the UNIX system manager and Marc Boutilier, the Web site, database and applications manager operate CBR-RBC under the supervision of project manager Christoph Sensen. Terry Dalton is the security manager for CBR-RBC. He is also co-ordinating all of NRC's CA*net2 and CA*net3 implementations.

Once the NRC Intranet was in place, users were pleased that the system worked very well indeed. Scientists had access to bioinformatics applications and databases as never before. Nevertheless, the network performance was still not good enough to readily accommodate the high-end operations that had been identified in the initial proposal. Accordingly, in 1998, a new plan was developed to move the entire CBR-RBC network to CA*net3, which will operate at 3 Gbit/sec, approximately 500 times faster than CA*net2. We are looking forward to the summer of 1999, when all of the changes will have been implemented and we can write another report about networking in Canada.

Bioinformatics is an international science, and Canadians are very aware that many of the pioneering efforts in this field are coming from Europe. EMBnet is an excellent model of collaboration among bioinformaticians, and very early on, Canadian bioinformatics experts identified membership in EMBnet as a highly desirable objective for CBR-RBC.

The second phase of CBR-RBC, CBR II was implemented in 1998 to provide bioinformatics services to scientists at not-for-profit organisations in Canada. With strong support from the President and the Vice Presidents of NRC and SUN Microsystems Canada, a high performance SUN Enterprise 4002 with twelve 250 MHz CPUs, 2 Gbyte of main memory, 128 Gbye of hard disk space and 210 Gbyte of DLT tape space were added to CBR-RBC. This system became the official server for EMBnet Canada when membership was granted in September 1998. Accounts on this machine are maintained by CISTI using a Toronto-based, private-sector bioinformatics company, Base4 Bioinformatics Inc, as its agent. There is a flat fee of $195 Can/year for access to CBR II. We anticipate having several hundred users on this machine within a year.

There is still a lot to do within and for CBR-RBC but we are well on our way. We are quite happy to share our unique knowledge of distributed bioinformatics facilities with others who might want to implement a similar model in their country.

"Over time, the people of the Great White North discovered that the distributed bioinformatics facility had not only networked their computers, but had also created new friendships and fostered many new collaborations among scientists, - collaborations that never would have happened without the improved communication among the institutes."

This is the happy ending for our story. A Merry Christmas to all EMBnetters!

AN IMPROVED BLAST SERVER AND CLIENT

Christian Iseli and C. Victor Jongeneel, Swiss Institute of Bioinformatics and Office of Information Technology, Ludwig Institute for Cancer Research Switzerland

Introduction

The comparison of a query sequence to a database of experimentally defined sequences is one of the most commonly used tools in bioinformatics. Among the algorithms available to perform such searches, BLAST (Basic Local Alignment Search Tool; Altshul et al., 1990, 1997) is by far the most popular, due mostly to the quality of its heuristics and its good performance. BLAST software is distributed free of charge to academic users by the NCBI and by Warren Gish at Washington University.

The setup of a server implementing the BLAST programs as currently distributed poses several problems to the system manager. First, there are three versions that use different heuristics, different statistics and different alignment strategies (the "original" BLAST version 1.4, WU-BLAST 2.0 with enhancements by Warren Gish and NCBI BLAST 2.0 with enhancements by Steve Altshul and colleagues). Because of the differences in the algorithms, any one of these three versions can produce optimal results for a given problem. Also, and predictably, each version has its own command-line syntax and set of parameters. Secondly, BLAST searches are usually run in a client-server architecture; the BLAST client and server programs as distributed are minimalist at best. Client-server interactions are also a problem with the GCG program suite; it emulates a BLAST 1.4 client and is incompatible with NCBI BLAST 2.0 servers. Thirdly, the database formats are different and incompatible between BLAST 1.4 and WU-BLAST on the one hand and NCBI BLAST 2 on the other.

Because of these limitations we decided to produce BLAST client and server software that would provide better functionality while retaining, and combining, the features found in the original BLAST programs.

Focus of the project

A complete BLAST server consists of four software components:

- The BLAST program itself, which usually runs on a dedicated machine
- The BLAST server, which is normally installed on the same machine where the database searches are performed and manages requests submitted by clients
- The BLAST client, which takes requests from its local users (e.g. GCG users or CGI scripts called from a Web interface) and passes them to the server
- The Web interface to the BLAST programs, which in most cases is written locally.

The focus of our work was to provide replacements for the second and third components as the functionality provided by these components, in current BLAST distributions, was deemed inadequate.

Design goals

In order to provide adequate functionality to the client/server

pair while maintaining maximal compatibility with existing software, we settled on the following goals:

For the BLAST server

- Control the receipt of BLAST requests.
- Allow only a preset number of BLAST searches to run simultaneously, and keep the other jobs in a FIFO queue, toavoid overloading the CPU and memory of the machine(s) performing the searches (memory swapping can become a major bottleneck during BLAST searches).
- Redirect jobs to other servers, based on what database and what BLAST program (i.e., NCBI blastp, blastn, WU-BLAST blastp, etc.) are requested, to distribute the load in a flexible manner across multiple machines.
- Differentiate between fast (short) and slow (long) jobs. Manage a fast and a slow job queue, giving higher priority to short (interactive) jobs. This is to minimise the time taken by interactive jobs (e.g. submitted from a Web page). * Allow the results to be sent back via e-mail, especially for jobs predicted to take a long time.
- Support database farms, i.e. the searching of multiple databases based on a simple mnemonic (NCBI BLAST 2 only).

For the BLAST client

- Parse the command line switches and adapt them to the syntax of the program called by the server.
- Route the requests to the desired server.
- Collect and display the results.

Software design

We based the new software on the server and client code from the old NCBI BLAST 1.4 distribution. However, these programs provided no support for the new NCBI BLAST 2 programs and none of the extended possibilities stated in the design goals above.

Another drawback of the original BLAST server and client was that they required the NCBI gish and dfa libraries distributed with the old BLAST programs, forcing you to build those libraries on all the systems where you wanted to run the BLAST client. The client used the libraries to build a DFA (Deterministic Finite-state Automaton) to analyze the results returned by the server to decide if there were errors and when the output was complete. We thought that this was overkill and that some simpler string functions from the standard C library could be used for the same purpose. So we rewrote most of the client from scratch.

The client is very straightforward. It is still able to talk to standard BLAST 1.4 and WU-BLAST servers as well as to our enhanced server. The code has been rewritten to depend only on the standard C library. It understands how to start NCBI BLAST 2 jobs, how to specify multiple databases for farming and how to specify an e-mail address for sending the results.

The server uses threading to give a clean task management protocol. The first thread receives jobs through a network socket, parses the input and creates a job descriptor. This first step decides, based on the input received and the content of a local configuration file, if the request is to be processed locally or if it should be re-routed to another BLAST server. The decision is based on the name of the blast program to run (e.g., blastall-blastn would be the NCBI BLAST 2 blastn program, tblastx would be the WU-BLAST or BLAST 1.4 tblastx program) and on the name of the database to search. If the job is to be run on a remote server it is put into a new queue at this point. A pool of threads watches the remote queue and shuttles the requests and results between the remote server and the client.

The server configuration file contains an arbitrary number of lines, each of which is a triplet of the form:

```
program database host
```

where program is something of the form {blastall-}{t}blast[npx], where the prefix blastall- signals a job to be handled by NCBI BLAST 2, database is a database name, which can have the end of its name wildcarded with a *, and host is either local, meaning that the job is handled on the server, or the name of some other host. A request for 'program' run against 'database' is thus routed to host. Comments start with a # character. Anything after the third field is also considered a comment. Badly formatted lines are ignored. If no match to the parameters sent by the client is found in the configuration file, the server assumes that the job should be handled locally.

The next step does farm expansion. This allows the site to offer simple, intuitive names for sets of databases. Please note that WU-BLAST and BLAST 1.4 do not support simultaneous searching of multiple databases, and that therefore the "farm" concept applies only to NCBI BLAST 2. The server translates the farm names into a list of database components which are passed to the search program. The server is configured through two files which list the farms for nucleotide and amino acid databases.

The farm configuration files contain an arbitrary number of lines, each of which has the form:

```
farm-name: component1 component2 ...
componentn;
```

where farm-name is the logical name by which the collection is addressed, while component is the name of an actual

database on the server or a previously defined farm. The # sign starts a comment which terminates at the end of the line.

Once the job descriptor is complete and correct, it is put on the entry FIFO queue and the thread then waits for the next job request. The entry FIFO queue is monitored by N threads, N being the number of concurrent fast BLAST searches allowed on this server (N=4 at EMBnet-CH). N is a variable whose value is set at compile time. Each of these threads waits for a job to be queued on the entry queue and grabs one on a first-come first-served basis. Once the thread has found a job it will start the corresponding BLAST search and monitor its output. If the job completes in the allotted time (5 minutes wall-clock time by default) the results are sent back, either directly through the network socket or via e-mail. In case the job takes longer, it is put on hold and moved into the slow queue. The fast thread then goes on to the next job in the entry queue.

The slow FIFO queue is monitored by M threads, M being the number of concurrent long BLAST searches we want to allow (we have M=2). In a way similar to the fast threads the slow thread resumes the execution of the job it dequeued and allows it to proceed to completion. The results are sent back to the user in the same way as the fast threads.

WWW server

The new client and server do not provide Web services directly. However, it is very simple to implement such services. Normally, a form is used to collect the data pertaining to the BLAST job to be launched (program type, query sequence, database to be searched, search parameters). These data are passed to a CGI script which itself calls the BLAST client with the proper command-line syntax. The results are returned to standard output, parsed by the CGI script (e.g. to provide a graphical output or to "prettify" the presentation) and returned to the browser. The EMBnet-CH Web server includes several pages that allow submission of BLAST jobs and illustrate the utility of database farms.

Comparison to existing clients and servers

Clients

Compared to the BLAST 1.4 client, which is also used by WU-BLAST, our new client is easier to compile (no special libraries required), is able to submit NCBI BLAST 2 jobs and can request that results be returned by E-mail. It can communicate with a standard BLAST 1.4 server but in this case offers no additional functionality.

The NCBI BLAST 2 distribution does not include a command-line client, only an X-Windows graphical client. Our client will not communicate with a standard-distribution NCBI BLAST 2 server.

Servers

Compared with either a BLAST 1.4 server or a NCBI BLAST 2 server, our server adds (1) control of the number of jobs that can run concurrently, (2) redirection of jobs to other servers on the basis of the program and database requested, (3) queuing of both local and remote jobs, (4) rescheduling of jobs from a fast to a slow queue based on execution time, (5) return of search results by E-mail, and (6) expansion of a logical name to a list of databases (farming). All of these functionalities, except for the last, work for all versions of BLAST. It should be noted that the NCBI has implemented a sophisticated job scheduling and control system on its own BLAST servers. This has been specifically tuned to the local hardware configuration at NCBI, where all BLAST jobs run on a series of multiprocessor SGI Origin 2000 machines. To our knowledge the NCBI system is not easily portable to another environment. Our server should be able to handle any environment, from a single high-powered server to a collection of dedicated machines.

Local implementation

At the present time (Dec. 1998), the Swiss EMBnet node operates two physical BLAST servers, one running WU-BLAST 2.0 and the other NCBI BLAST 2.0. One machine, running the server software described above, dispatches the jobs and requeues them as a function of (1) the program suite requested (NCBI or WU-BLAST), (2) the actual comparison program and (3) the time taken for the job to complete. The second machine receives its jobs from the first, does no job control, and runs the standard BLAST 1.4 server.

In the near future we plan to expand the number of servers to four, of which one will still run WU-BLAST while the three others will run NCBI BLAST. The databases will be distributed among these three machines in such a way that the data can be kept in memory at all times; proteins on one machine, non-EST nucleotide sequences on the second, ESTs on the third. This should give us maximal flexibility and performance.

Please note that it would be perfectly possible to distribute the tasks in other ways. For example, one machine could run both NCBI and WU-BLAST on the same database set, keeping in mind the fact that the database formats are different and that they will thus load into memory independently. It would also be possible to reserve different machines for different programs (e.g. run tblastn and tblastx on a machine where they will not interfere with shorter, interactive jobs).

We have taken advantage of the "database farm" concept to offer a wide array of search set possibilities. For example, we keep the EMBL database in separate pieces according to taxonomic classification and update status and define farms to designate the more traditional BLAST databases. This is described in the following configuration file (farms.nuc):

```
# Nucleotide farms
#
# First the ESTs
est_hum: est_hum-56 est_hum-up;
est_mus: LTmouse_brain est_mus-56 est_mus-
up;
est_pln: est_pln-56 est_pln-up;
est_unc: est_unc-56 est_unc-up;
dbest: est_hum est_mus est_pln est_unc;
dbestu: est_hum-up est_mus-up est_pln-up
est_unc-up;
dbest56: est_hum-56 est_mus-56 est_pln-56
est_unc-56;
dbest-56: dbest56;
dbest-up: dbestu;
#
# Now, the STSs
dbsts: sts-56 sts-up;
dbsts-56: sts-56;
dbsts-up: sts-up;
#
# Here is EMBL
fun: fun-56 fun-up;
gss: gss-56 gss-up;
htg: htg-56 htg-up;
hum: hum-56 hum-up;
inv: inv-56 inv-up;
mam: mam-56 mam-up;
org: org-56 org-up;
patent: patent-56;
patent-up: patent;
phg: phg-56 phg-up;
pln: pln-56 pln-up;
pro: pro-56 pro-up;
rod: rod-56 rod-up;
syn: syn-56 syn-up;
unc: unc-56 unc-up;
vrl: vrl-56 vrl-up;
vrt: vrt-56 vrt-up;
embl: fun gss htg hum inv mam org patent
phg pln pro rod syn unc vrl vrt;
embu: fun-up gss-up htg-up hum-up inv-up mam-
up org-up
phg-up pln-up pro-up rod-up syn-up unc-up
vrl-up vrt-up;
emb56: fun-56 gss-56 htg-56 hum-56 inv-56
mam-56 org-56
phg-56 pln-56 pro-56 rod-56 syn-56 unc-56
vrl-56 vrt-56;
embl-56: emb56;
embl-up: embu;
#
# Here comes nr, actually embl + gbex (gbex
# is never updated...)
nr: embl gbex;
```

Future plans

On the client side we hope to patch (or rewrite) the GCG BLAST client so that it will be able to submit NCBI BLAST 2 jobs to our server. Currently, GCG is limited to submitting BLAST 1.4 or WU-BLAST jobs.

On the server side we plan to modify the NCBI BLAST 2 code to allow external filtering of the query sequences. For unknown reasons support for external filters (such as seg, xnu, or cross_match) was dropped with the passage to version 2.0 and the filters provided internally have often proved to be inadequate.

Availability

The software is written in standard ANSI C and will compile with gcc on every platform we have tested so far. The client uses only standard C libraries while the server requires support for POSIX threads. The code for the actual BLAST programs should still be obtained from its original sources.

Please contact Christian Iseli (Christian.Iseli@licr.org) if you are interersted in obtaining the BLAST client and server code. If demand is high we will put it on our anonymous FTP server.

Email Security

Alan Bleasby UK EMbnet National Node

There is a tired old cliche that is always used when discussing security in computer systems. The computer is a castle, its walls and moat the main defences, the drawbridge doubles as its communication point and firewall leaving Postman Pat responsible for the nasty old internet. The author normally avoids tired old cliches like the plague but is at a loss to think of a better simile. This article primarily addresses email security and therefore Postman Pat; or rather how to try and compensate for a postman who is inherently a rather generous chap. Also a certain metaphysical element creeps into email security. It is not so much of a question as "Who am I?", more one of "Who is he?"

Pat is a sharing kind of guy, he doesn't bother keeping his sack closed. It's therefore possible that when he visits a house the owner could grab a handful of letters, some of which don't belong to him. The owner can read the messages and afterwards drop them back in Pat's sack, or maybe keep them. You have to hope that the householder's hobby isn't espionage.

This problem arises as messages are carried by Pat as packets or parcels. Each parcel contains the address of the destination, the address of the sender, a lot of other interesting information beyond the scope of this article and, of course, the message. It is up to each householder to say "That one's for me!"; Pat is a trusting man. You can never be sure, even though the desired recipient gets your message, that it has not been intercepted by some unscrupulous soul. To make matters worse, Pat's eyesight isn't that good. If someone drops a parcel in his sack with a fake sender address he never even notices. In short, not only can someone intercept what you do say but people with ulterior motives can send messages which appear to come from you. By the same token you can't be sure a message you receive was sent by the person claimed. Feeling paranoid?
Who would employ a postman like that? Noone of course, any self respecting employer will insist on one that's also had at least one hernia operation and has a bad sense of direction. Poor old Pat cannot carry heavy parcels and frequently doesn't go straight to the destinations on them. Strangely these turn out to be redeeming features. Every long message sent out by a castle has to be split into lightweight chunks, each marked in a numerical series. The destination castle's job is to collect the pieces and reconstruct the message in the correct order before passing it onto the recipient; there is no guarantee that piece 3 will arrive before piece 4 for example. That does add a little security but will not put off an interceptor who's learned to count. So what can be done? Let me take you back to your childhood, ..childhood, ...childhood, ...childhood.

It should now be clear that the only area in which you can apply security is in the message itself. Most people, as children, will have sent secret messages using codes or cyphers. Codes were usually taken from espionage films of the period e.g.

"Do you have a copy of the Times?"

"No, but I have a red carnation."

Codes tend to be more secure but have a limited range of meaning. Cyphers give greater flexibility. At their most simple they are substitution methods. It doesn't take a genius to crack the message:

BCPYBK GOETBE SIM F TFLLS UTQOEKCFE

Cryptography has developed into a complex science; it would be relatively trivial to crack the WW2 Enigma machine today. Powerful crytographic algorithms are in the public domain.

So, you can turn your messages into gobbledegook and email them. This does, however, require that the person at the other end has been provided with the means to decrypt it. This "key" would have to be transmitted across secure channels. A better way is to use public key cryptography as used by the PGP (Pretty Good Privacy) software. With PGP two keys are used. Each person has a "public" and a "private" key. Only the public key can decrypt messages encrypted by the private key and only the private key can decrypt messages encrypted by the public key. How does this help?

Imagine X (male) and Y (female) want to exchange messages without being eavesdropped on by Z. X emails Y with his public key. It doesn't matter if it is intercepted by Z. Y sends her public key to X. Again, it doesn't matter if Z gets a copy. The position now, providing it was X who sent his key to Y and it was Y who sent her key to X is the following. X can send encrypted messages to Y using her public key. Only she can read the message using her private key. Similarly Y can send messages to X encrypted with his public key so only he can decrypt it using his private key. But, hang on, isn't there a flaw? Yes there is.

There is the possibility that Z has sent one of her keys to X claiming that she was Y and it was Y's key that was being sent. Z could also play the same trick on Y. In order to get around this you do need a secure channel for a while in order to find out whether the public key you have really does belong to that person. That can normally be done by phone. PGP allows you to ask a person for a kind of checksum representing their key. This is done by typing a simple command at each end. If the numbers match then that key is the public key of the person it claims to be from. It is very unlikely that two public keys would have the same checksum and, even if they did it would be no guarantee that they could decrypt the same code.

So, we are now in a position where X and Y can talk securely but Z may have a copy of both their public keys. Z can send messages to either X or Y using these keys and claim to be the other person although she can't decrypt any messages sent by X or Y if they encrypt with each others keys. You're not losing anything by this since Z could always send you mail anyway but it would be nice if there was a way by which you could be sure who really did send the message and solve the "who are you?" problem. PGP allows this by the ability to use signatures.

Signatures are short encrypted addenda to messages. They are produced from a digest of the message itself using your own private key. Only you can produce such messages. Any recipient will know that as they can decrypt the signature since they have a copy of your public key. You now have a secure way of sending messages and know who sent the message. Bingo! We can now return to the subject of just how paranoid you are.

If the message really is hot then you will obviously want to encrypt it. This will turn it into unreadable gobbledegook.

More often though its not important if someone else reads the message but its vital for your peace of mind that people know it's really sent by you. It's also slightly more convenient for other people in that they don't have to decrypt the whole message; they can just read the email as normal. PGP allows you to just sign a message and not encrypt it. The text will appear normal and unadulterated. As your signature is derived, in part, from the message content it is impossible for someone to cut out your signature and use it in another message. That signature would be invalid. It is always good practice to sign your messages even if you don't encrypt them. Most people aren't too paranoid and don't mind if everyone has a copy of their public key

PGP allows much extra flexibility. For example, it keeps all keys in a keyring database, you can get other people to sign your public key as further indication that you are who you say you are etc. With all this functionality you could easily forget how to work PGP from the command line. Life is made very simple though if you have an email reader which is "PGP aware." There are many flavours of such email readers. The author uses "exmh" which is an X-windows interface to the "mh" or "nmh" mail system commonly found on Linux systems. A browse of the net will point you in the direction of PGP aware systems suitable for your machine. These systems allow you to encrypt & sign or just sign messages. They tell you if a signature exists and allow you to verify it. They also automatically decrypt messages before displaying them on your screen, allow you to post your public key etc. All at the click of a mouse button.

By now you may be itching to try it out. One word of caution. You'll see various versions of PGP out there. The one recommended is from the original author; for Europeans this is pgp-2.6.3ia and will work with most PGP aware software, versions with higher numbers may not. The author has bundled up PGP and some related software for mailers and mail interfaces. This is available via anonymous ftp (ftp.uk.embnet.org pub/pgp). If you just want to operate PGP from the command line then simply download the PGP file.

For those of you who are really paranoid don't worry, those noises in the middle of the night will just be Santa coming down the chimney, just make sure he signs the parcels!

A quick guide to Emacs

Alex Finck, European Bioinformatics Insitute, Hinxton Hall, Hinxton UK.

Definitions:

shortcuts

Emacs has perhaps hundreds of them; you can add yours and change the existing ones but the following can be trusted:

C-x stands for "ctrl + x" (at the same time)
M-x stands for "alt + x" (at the same time)

Those shortcuts are also the main modifiers in emacs (consider them as a modifier like "Ctrl" for any other program)

minibuffer

The minibuffer is the small window which you find at the bottom of your emacs. You access it to launch a command. After having typed C-x C-f you're in the minibuffer. Initially you should always quit (C-]) it before executing a new command.

buffer

The buffer contains text and images. If a buffer is a bucket, then a file can be compared with water. Keep in mind that you can put whatever you like in a bucket!

modes

Modes make emacs behave in a particular manner, according to your wishes and to the type of the opened file. The mode-line has not much to do with modes and is not really important. It is the line just
below a buffer, where you find the filename and the line number.

.emacs

This initialisation file is read by emacs on start-up. Unlike other emacs lisp modules .emacs has no .el or .elc extension. "el" stands for emacs lisp and "elc" for emacs lisp compiled. The "-" is the accepted separator in emacs.

What is emacs?

GNU emacs, Xemacs and other emacs-like editors are text editors, all derived from the original distribution (which works on PDP10/ITS or DEC-20/TOPS-20) released in 1975. Popular Emacsen are written in Lisp and contain a

Lisp compiler. That makes emacs configurable in any way you can imagine. It also makes it an addictive tool for programmers.

What is it for?

For Biocomputing, emacs can be used to simplify and unify the use of UNIX, to program with different languages using the same tool. Emacs offers different programming modes (perl, c, c++, java, lisp, prolog, python, sh, csh and many others) which will facilitate your work in many cases. For instance, it can syntax colour your code and auto-indent it so that syntax errors become more obvious. Another use is with languages like perl or python where Integrated (and Visual) Development Environments (IDEs) are not yet available. Emacs can work together with external programs like debuggers, version controlling systems (RCS, CVS, SCCS) and SQL clients. Emacs can use various other utilities like the statistical tool called S-plus.

Paradoxically it can also be embedded in other programs like IDEs. Finally, Latex has a dedicated mode in emacs which can ease the conception of a scientifical paper (using the formula editor to create complex mathematical formulae or the bibtex-mode to include citations).

Where can I find information about it?

On the web, search engines like alta-vista or dejanews are a good starting point but your keywords will need to be precise. Emacs has its web site shared with other gnu programs at http://www.gnu.org/ whereas xemacs has its web site at http://www.xemacs.org/Newsgroups (gnu.emacs.help,comp.emacs, comp.emacs.xemacs...) are active and offer good support.

A recommended book is: Cameron D. et al. Learning Gnu Emacs. O'Reilly & Associates, 1996. Many
other are available just by searching on the web sites of booksellers or editors. When emacs is installed you'll find much more information within it (see "How to use it?") and in the source directories (/lisp for modules and /etc for documentation like refcard.*).

At the European Bioinformatics Institute, Peter Sterk has made a tutorial available under http://www.ebi.ac.uk/~sterk/emacs_tut/tut8.html or at http://www.ebi.ac.uk/~finck/index.html

Is emacs supported on my computer?

Type "where emacs" or "which emacs" (or xemacs) in your UNIX shell (check your PATH environment variable to see which directories are in your search field). If emacs isn't installed, search the above web sites for the nearest ftp mirror or be polite to your system administrator when you ask him or her to install it. As for the availability, emacs has been ported to almost all UNIX machines and NT/95/98 versions are available.

How do I use it?

Emacs is perhaps a proof that programs can be written by an unlimited number of authors, this is also why emacs' use is so tightly linked to emacs lisp code. xemacs tries to overcome this by adding visual functionalities like pop-up windows support and better menus. But even in xemacs, the typical user will launch the editor only once (because of the longer initialisation delay of a customised emacs) and work with only one OS window (because emacs manages specific character based windows). Emacs is well-known for its magic keyboard shortcuts however the menus have a growing support and are useful. Now let's quit theory and dive into the common use of the editor:

1/ Basics

First, type C-x C-f. That means C-x followed by C-f; You will be redirected to the minibuffer unless you're already there (in which case type C-] or C-g to quit). When you get a prompt in the minibuffer you'll have to choose a file (much of what you expect to work in your shell will work there). Choose nothing, just type return. You should see the directory (which was put as a default in the minibuffer) listed in a visible buffer. If not, retry and delete the last "/". To quit a buffer and to kill it type C-x k. If you just want to hide it type C-x 0 that's zero not o because C-x o allows you to go from one window to another window (including the minibuffer) without using the mouse (which is always there to access the menus when you don't remember a shortcut!). Similarly, if you retype C-x C-f to open a file (this time try tab to complete directory and file names), you'll be able to obtain two windows for it by using C-x 2. This shows how you should think about your emacs. Some things are obvious, some others are misleading! If you would like to continue with this style of linear explanations type C-h t, which will lead you to the internal emacs tutorial.

2/ The shortcuts jungle

The previous short introduction should allow you to use the following shortcuts without getting too baffled. The typographic conventions used below are also those used by the emacs documentation. Nevertheless, you will find parts of some shortcuts in parentheses. Parentheses underline the fact that these shortcuts are highly context sensitive, like the very useful M-y (after C-y).

First aid:

C-x C-f open a file
C-x C-b display the buffer-list

In dired-mode, buffer-list:
g update the content of the current buffer (use it systematically)
m mark (select) item
u unmark (unselect) item
x delete selected items
X shell command on selected items
d mark item for deletion
s switch display by time/name (only in dired-mode)

C-x C-s save the current buffer
C-] quit the minibuffer (and recursive minibuffers)
C-g interrupt
C-x k kill a buffer
C-_ undo
C-_ (after a first C-_ followed by any cursor move) redo
C-x 0 hide the current window
C-x o go to the other window
C-x 2 horizontally split the current window into two
C-x C-c quit emacs

Help:

C-u M-x apropos emacs asks for a keyword you want help on; the C-u adds the related shortcut(s) to the results of the query.
C-h b emacs gives the keybindings for the current buffer
C-h k emacs asks you to type a shortcut and explains it
C-h v emacs asks for a variable and gives its value
C-h f emacs asks for a function and explains it
C-h i browse the "info files"; you can move around using return, u, n, p
C-h w emacs asks you for a command name and tells you the associated shortcut.

Searching:

C-s find/search text foward in buffer
C-s (after C-s) find again forward
C-r find/search text backward in buffer
C-r (after C-r) find again backward
M-% search and replace
M-C-s search according to a regular expression (see below)

Regular expressions:
\ escapes the next character
^ beginning of a line
line$ end of a line
. any single character
.* zero or more characters
word\> end of a word
\<word beginning of a word
[] range of characters like [a-z] for a to z or [ac] for a or c
[^a-z] not a range; in this example not [a-z]
i.e. ^GNU.*\<like will find the line beneath "What is emacs?".

Editing:

M-\ erase spaces and tabs near the cursor
C-a go to the beginning of the line
C-e go to the end of the line
C-k delete the rest of the line
C-d delete the character under the cursor
M-q fill or join the words of the current paragraph
M-x set-justification-full emacs fully justifies like a wordprocessor
C-x r t emacs asks for something to insert in front of a selection (supposed to be a rectangle)
M-x goto-line emacs asks you at which line you would like to go (useful for debugging)

Copy/Paste:

C-space (before C-w or M-w) mark the beginning of a region
C-w mark the end of and cut (in emacs say kill) a region
M-w mark the start of and copy a region
C-y paste (in emacs say yank) a region
M-y (after C-y) paste any previously "killed" region

Miscellaneous:

tab emacs completes the file or directory name
M-/ emacs completes the word according to a personal dictionary, to buffers and to the current mode (perfect for long variable names)

C-x C-w save the current buffer in a different file
C-x C-v close the current file and open another in the same buffer instead.
M-; emacs comments out a region according to the current-mode
M-x ispell-buffer emacs verifies the spelling of the current buffer
M-$ emacs verifies the spelling of the current word
C-x C-e evaluates the last Lisp expression (sexp)
C-x C-f /username@systemname:/ open files on a distant system using ftp
M-x telnet start a telnet session
M-! issue a basic shell command
M-x shell start an emulated shell (use M-p and M-n to move in the history)
M-| emacs executes a shell command on a region and returns the results in a user defined or default buffer

C-q escapes emacs interpretation of characters (i.e. in cperl-mode C-q tab will insert a real tab instead of running

perl-indent-command)

When installed:

M-x find-file-at-point emacs opens a file according to the filename under the cursor

Programming modes:

M-x compile emacs asks for your compile command (ie /perl-path/perl -w -c /module-path/module.pl). You can also use the "make". emacs accepts other languages the same way.

M-C-\ emacs indents the selected region
M-x font-lock-fontify-buffer emacs syntax highlights (colours) the current buffer according to the current mode (programming language).

M-x perldb emacs, through an interface named gud, asks for your perl debugger command (ie /perl-path/perl -d/module-path/module.pl). gud accepts other languages and is well interfaced gdb, sdb, dbx, or xdb.

M-x show-paren-mode emacs highlights the matching delimiters of a block

C-x C-q toggles the read-only permission of a file (if the file is version controlled it has the same effect as C-x v v). Read-only permission is useful to avoid syntax errors in a program when consulting it. Emacs is particularly exposed to that risk.

RCS, CVS:

C-x v v emacs asks the VC system to check in/out the current buffer
C-c C-c emacs notifies your comments to the VC system

2/ Customisation

First some more definitions:

commands
commands are directly linked to code but can all be called directly using M-x command-name

keybindings
keybindings are linking shortcuts to "internal" emacs commands

keymaps
keymaps are variables containing the keybindings

hooks
hooks are functions triggered on entering/exiting a mode

modules
modules contain commands and functions; a mode is usually defined in a single module.

When you start to use emacs, some of the above shortcuts will appear strange for a while. However,
much of the complexity of emacs resides in the fact that all the above shortcuts have exact command (M-x) equivalents: C-] <=> M-x abort-recursive-edit; C-x C-f <=> M-x find-file; C-x k <=> M-x kill-buffer. You can find all the exact names (commands or macros) associated with the jungle of shortcuts by using one of them after typing C-h k. The name of a module (i.e. simple.elc) is specified in the results of C-h k. This module (or its source file) will often offer a last chance to understand any perceived strange behaviour of emacs. You can change, add and remove modules enabling you to
upgrade your emacs without waiting for the next version to be released (see the programming and modes sections).

.emacs:

Though different from most, from a users' point of view the most important of the modules is .emacs.
Deciding how you want to fill this file can be a strenuous game but you can get information from various sources. You'll probably find the basics in the "Emacs" or "XEmacs" menu, "Customisation" submenu of the "info files". You can also read existing .emacs files, which are usually well documented. If you use xemacs, open the one provided in the help menu under the "samples" submenu. You may also find one in your home directory and if that is not sufficient, you can search those of your colleagues (many ..emacs files have read access to the world). At first, learn only a minimum of Lisp then you can learn to use C-x C-e, M-x show-paren-mode, ";" for comments, the *Messages* buffer, (load-file), (define-key), (setq), keymaps and a few others. For example, you could put the following in your .emacs:

```
; should make your emacs scroll with less jumps
(setq scroll-step 1)
; should automatically syntax colour buffers
; according to the mode (global-font-lock-mode t)
; should ease the opening of a file.
(global-set-key "\M-o" 'find-file)
; should ease the way you travel between windows
(global-set-key [(meta right)] 'other-window)
(global-set-key [(meta left)] '(lambda()
        (interactive)
        (other-window -1)
        ))
; should avoid you entering recursive edit
; by typing repeated M-x
(define-key minibuffer-local-map "\M-x" 'abort-recursive-edit)
; should link C-x to only one key: pause.
```

; Choose yours!
(define-key global-map [(pause)] ctl-x-map)

You will probably find many other useful configuration tricks and you will also find different modules or parameters to do the same thing; sometimes you will find bogues or side-effects. This is perhaps the main argument for using a short and concentrated .emacs. In particular, try to limit the use of .emacs when you discover the editor.

3/ Modes

Modes are an abstract concept. They have a common structure but their content can be very different (cperl-mode compared with isearch-mode). They group keybindings by using keymaps, hooks and other internal emacs functions. Modes are associated with a particular type of file (a particular extension for instance). Modes can also be associated to other modes. These modes within modes are called minor-modes (isearch-mode is one of them).

Modes are configurable in various ways. You can change the file association (when there is one) as
shown below:

```
; should make sure that files finishing with
; .me and .bak will be opened in text-mode
 (setq auto-mode-alist
       (append '(("\\.bak$" . text-mode)
       ("\\.me$" . text-mode))
       auto-mode-alist))
```

Modes almost always comprehend hooks and keymaps and are also configurable. The quickest way to change the behaviour of a mode is to change the complete associated module. How you change such a module is described in the following programming section using an example of cperl-mode.

4/ Programming

Programmming with emacs is language dependent. For each language there is a corresponding module called by emacs. cperl-mode.el is installed by default in recent emacsen. Even in the latest emacs the module used can be quite old. If you want to have the most recent, or another one, you'll have to download it from the internet. To find where, you have two methods:

- search the original module in emacs source directories and contact the author who almost always gives his on-line references (for cperl-mode the author is Ilya Zakharevich ; http://www.cpan.org/authors/Ilya_Zakharevich)

- use search internet engines to find other authors (see where can I find it?)

When your preferred module is installed (the original one will do for a while), you can open a file with a .pl or .pm extension and you should see "Loading cperl-mode..." in the minibuffer. With the perl file opened, for example /path/my-script.pl, you can compile it (for perl that means check the syntax) using M-x compile. You'll have to enter the full path and syntax for your compiler. For instance:
/other-path/my-perl -w -c /path/my-script.pl. The effective syntax checking will be done by the compile module which will redirect you to any errors it finds.

5/ More

Look for shortcuts by loading some specific modules included in the distribution. Search for pc-mode.el , pc-select.el in emacs; for xemacs use emacs modules or search for motif keybindings.

crypt++, uncompress, tar-mode allow you to automatically decode .a .gz or .Z documents

emacs doesn't complain much if you open documents of 60Mb

emacs is unfortunately not yet threaded.

Sometimes, the only way to understand emacs is to go and see right inside the module called by a buggy function. That might seem horrible but at least emacs code is easy to read!

Introduction to Genesafe

Ewan Birney, Sanger Center
On behalf of the Newton Institute gene prediction week participants.

The genesafe mailing list was created to help the gene predictors to collaborate on training and testing sets. It comes out of discussions at the Newton Institute. This email is meant to summarise the current position of this collaboration and to start the process off of discussing it.

The genesafe mailing list is open to everyone. You can join it by sending an email to majordomo@hgmp.mrc.ac.uk with 'subscribe genesafe' in the body of the message.

The discussions in the Newton Institute were as follows:

a) Chris Burge put forward an idea about making a test set due to only mRNA vs genomic comparisons. Although the mRNAs may not be full length, this still seems like the fastest way of getting a high quality database of known gene structures together. Chris suggested distributing the set in GenBank format with special tags to indicate which introns

were ambiguous or not in this process. The benefits of this were as follows:

- a relatively large dataset
- a completely consistent way of referring to each gene, with confidence of its correctness.
- an automatic procedure for generating the dataset.

b) Later in the week (sadly once Chris had left) the main gene predictor people got together again to discuss the gene prediction problem. In particular there was Soren Brunak, Mark Borodovsky, Anders Krogh, Victor Solovyev and myself (Ewan Birney) as potential gene predictors, and Martin Bishop and people from the Sanger Centre as other people interested in this process. Soren suggested that we focused the discussion on the following areas

- the type of data sets to have
- the practicalities of making the data sets
- the formats to distribute them

This meeting ended up with the following points. We divided data sets into 3 classes

- Class I - Experimentally researched areas, with confirmed 5' ends of genes and nd potential negative data in genomic regions
- Class II - mRNA confirmed genes, probably not full length
- Class III - 'Best guess' gene structures, integrating gene prediction software, homology and est data, usually hand edited.

Chris' data set would be a good example of a Class II dataset, as would the dataset for C.elegans prepared by Steve Jones. This dataset looks as if it would be the first type to be available. Class I data was potentially available from the BRACII region, the XLP locus and Chromosone 22 project from the Sanger centre, as well as probably a number of other regions from other centres.

Class III data is basically the sort of 'standard' annotation that the C.elegans/human/fly projects attempt to make. Because of its reliance on gene prediction programs to 'fill in the gaps' it's use in training and particular testing is questionable.

We thought that there should be a central place to hold the data. There was not a lot of people rushing to host this, but thankfully Martin Bishop suggested that the HGMP hosts a web site and mailing lists. We were not clear at the time how this would work with Chris' plans, and we were very happy to mirror something that Chris has done or is planning to do or to provide a stable single resource for it, taking Chris' dataset and removing the burden of site maintance from him.

For each dataset, we though the following pre-made splitswould be useful

- split into - large genomic regions all genes (including small genomic regions)
- provide pre-purged datasets (Soren Brunak offered to provide scripts/expertise in this)
- provide pre-split datasets to facilitate cross-validation

For anonymous validation test sets, it was thought that the sort of 3rd party evaluation of gene predictions was very valuable, and we would like to encourage Tim Hubbard in particular to maintain his no nonsense evaluation of the gene prediction programs.

The formats suggested were EMBL, GenBank and GFF. The GFF format seems the most focused on this problem and a number of gene prediction groups (Anders Krogh, UCSC and Sanger Centre) are already using it. (a description Of it is at http://www.sanger.ac.uk/Software/GFF/). We thought that if possible the resource should have scripts that interconverted between these formats. Generally it was thought that most groups were savvy enough to handle most formats, so this wasn't a big problem.

Finally we were aware of a number of different resources already trying to address this, or similar problems. These include

- Chris Burge's start on his mRNA dataset
- The Banbury cross contest/web site
- The GFF mailing list
- Steve Jones' C.elegans dataset
- The European Drosophila project dataset.

I think we all felt that it was best to work with all these people to try to provide a single resource. We therefore welcome any suggestions about anything, and at the moment, nothing is fixed. ;)

I would like to outline the following stages for discussion:

encouragement of other interested parties to join this list and post their ideas.
establishment of a web site at the very least to archive this list (I will do so by hand for the moment).
figure out a who-does-what list over the next couple of weeks and time-line for it.

I am looking forward to having nice proper datasets to use and test my programs on. So... please start by forwarding this message onto whoever you think might be interested in it.

Contact birney@sanger.ac.uk or for more information http://www.sanger.ac.uk/Users/birney/

Jalview

Analysis and Manipulation of Multiple Sequence Alignments

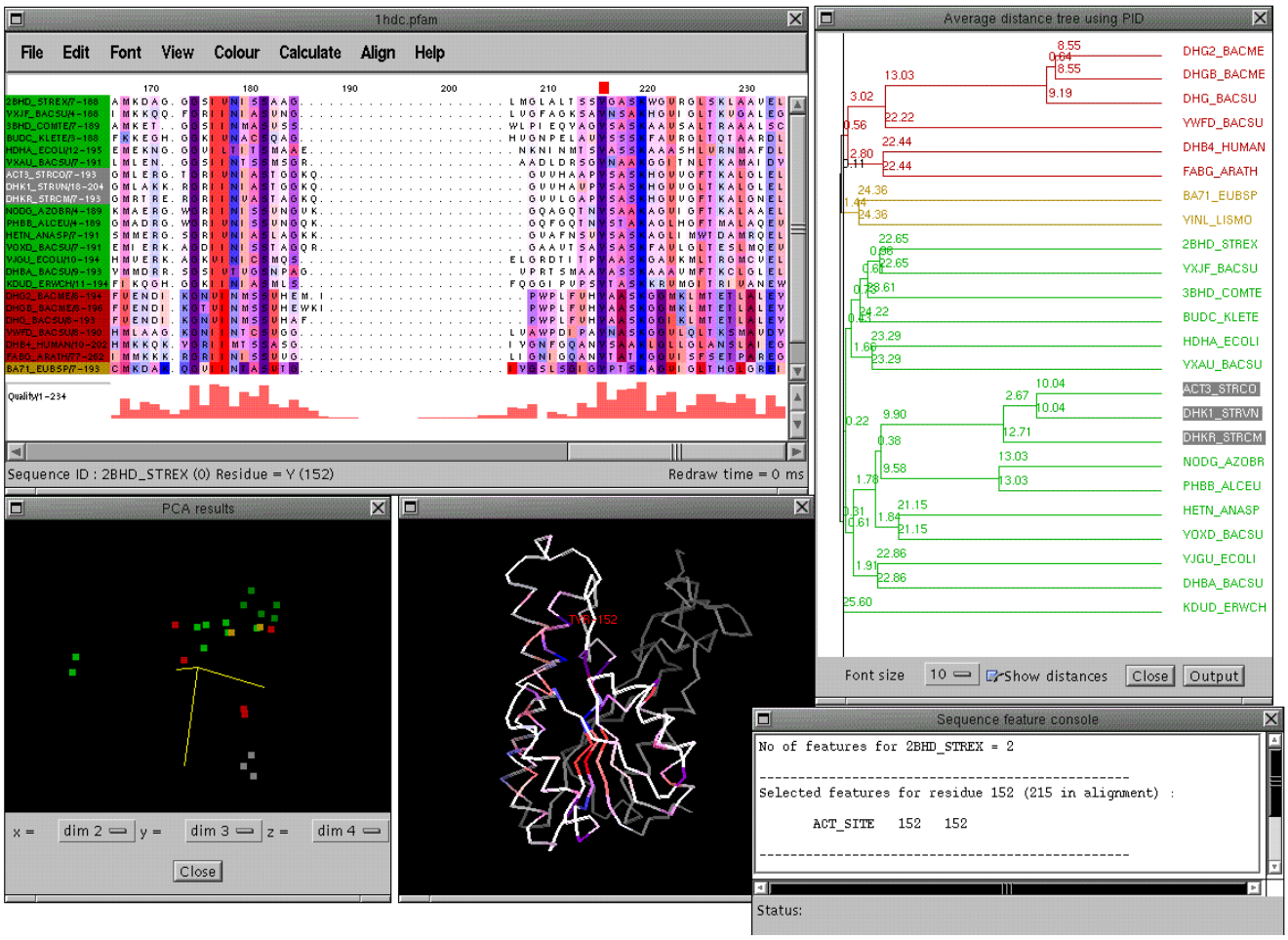
Michele Clamp, James Cuff and Geoff Barton, EMBL-EBI, Hinxton, Cambridge, UK.

Summary

Jalview is a tool written in Java to analyse the residue conservation patterns in a protein multiple alignment as well as being an interactive alignment editor. Unaligned sequences can be aligned either locally or remotely at the EBI with further analysis programs available remotely at the EBI. Access to the database entries for individual sequences is available through SRS. The sequence features can be extracted from the database entries and displayed graphically on the alignment. If three dimensional structures exist for any of the sequences then the structures can be displayed and coloured according to the colour scheme or conservation patterns in the multiple alignment.

Availability:http://circinus.ebi.ac.uk:6543/jalview/ and http://www2.ebi.ac.uk/clustalw

Contact: michele@ebi.ac.uk



A PFAM [14] alignment of short chain alcohol dehydrogenases which have first been grouped using a dendrogram and then the conserved columns in each group have been coloured according to each residue's hydrophobicity. Underneath are shown the 2nd, 3rd and 4th principal components which is an alternative way of clustering the sequences. The PDB code for the structure was obtained from the feature table of one of the sequences in the alignment and has been coloured according to the conservation patterns in the alignment. This has highlighted the hydrophobic (red) core strands in the structure.

Introduction

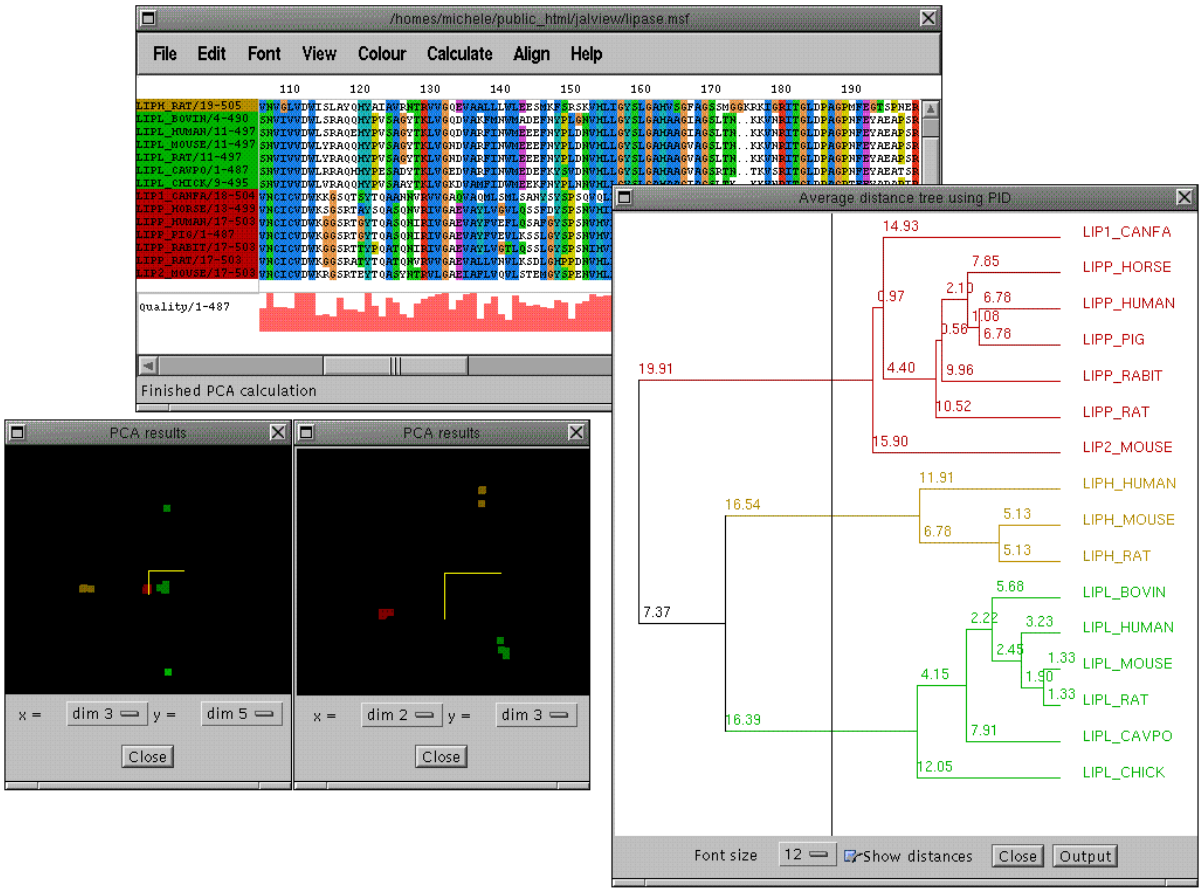
A multiple sequence alignment of a protein and its homologues can be a source of information about their common functional and structural features. Identification of these features requires an accurate alignment from which to extract the common features that may be of interest. Even though there are many excellent multiple alignment programs available (e.g. Clustalw [1] and its front end ClustalX [2] and AMPS [3]) there are unfortunately always cases where these automatic methods fail and the alignment has to be changed by hand. Both automatic and manual methods can be used in Jalview to create an alignment. Sequences can be imported into the program and aligned using ClustalW either locally, if Jalview is being run as an application, or remotely via CGI. The automatic alignment can then be altered by hand using the mouse. Patterns of conservation are displayed by varying the colours and the intensities of the residues.

Other multiple sequence alignment editors do already exist such as seaview [12] and CINEMA [13]. In general a disadvantage with these programs is that when manually editing an alignment the user needs to see immediately the results of their edits and whether they change the patterns of conservation. Jalview allows the user to see the effects of their edits. After each edit the user can immediately recluster the sequences and recalculate the pattern of conservation for each alignment. Other external programs for secondary structure prediction can also be called after each edit allowing the user to see how dependent that prediction is on changes to that part of the alignment.

Description of features

Clustering

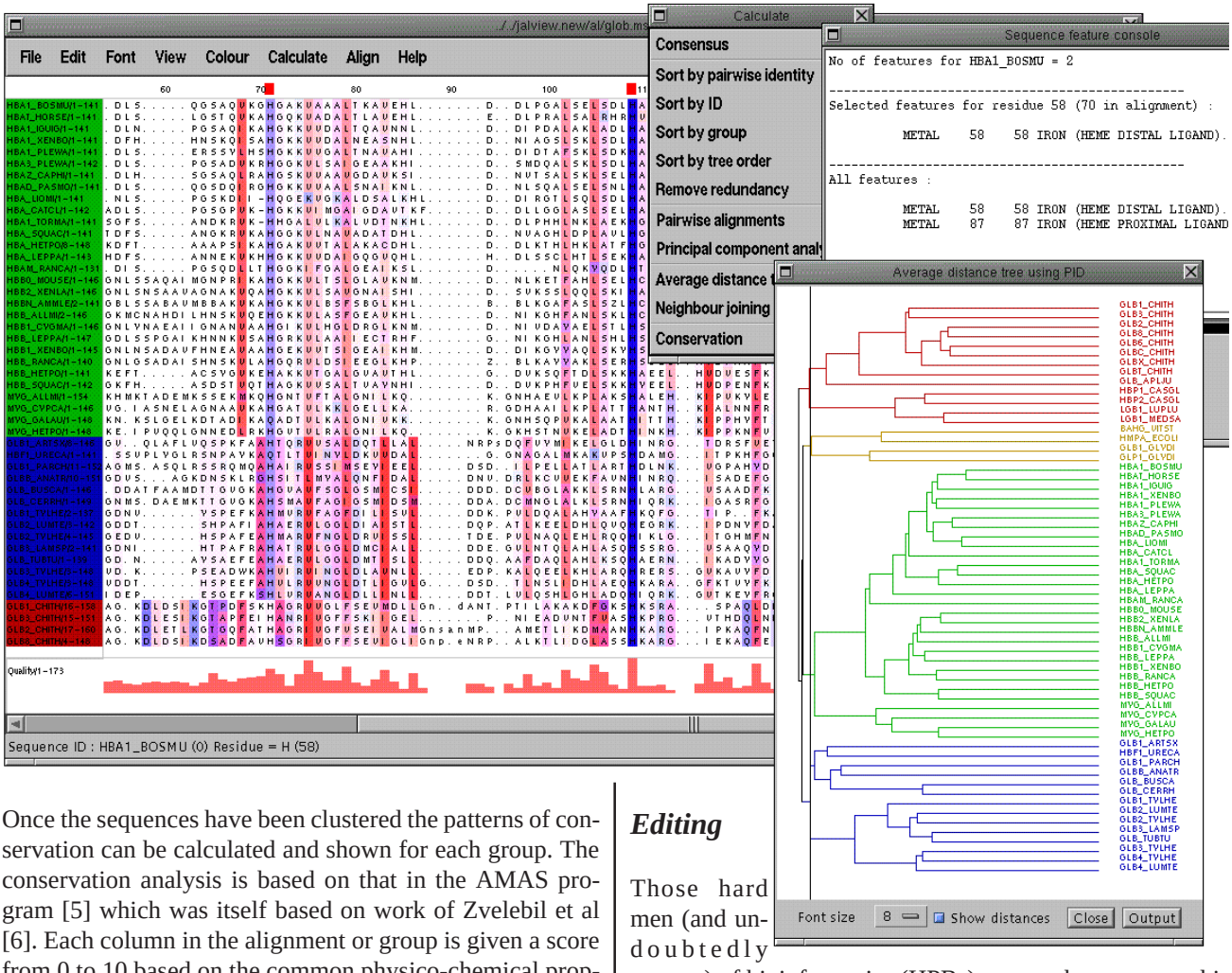
A multiple sequence alignment may consist of a number of subfamilies of sequences that exhibit their own patterns of conservation as well as sharing the common features of the whole alignment. With a large alignment it becomes difficult to spot these subfamilies by eye. Jalview provides two ways of clustering the sequences into subfamilies. A UPGMA dendrogram can be calculated and displayed (Calculate->Average distance tree) either on the whole alignment or on a subset of selected sequences for a large alignment. By selecting a point on the dendrogram with the mouse the maximum distance between any two sequences in a cluster can be defined. The different clusters are then shown in different colours both on the dendrogram and in the main alignment window. In the example below the dendrogram shows there are three obvious subfamilies which have been easily defined by one mouse click.



The other way of grouping sequences in Jalview is by calculating the principal components of the alignment (Calculate->Principal component analysis). This was initially applied to multiple sequence alignments by G. Casari et al and implemented in the program SeqSpace [4]. The PCA window shows 3 of these components at a time in a 3D rotatable view where each axis represents a property of the alignment common to some or all of the sequences. The most informative components to view for clustering sequences are dimensions 2,3 and 4. In the above picture 2 PCA windows are shown. The one on the right shows components 2,3 and 4 which are coloured according to the colours defined in the tree. The window on the left shows components (3,4 and 5) which show a splitting of one of the clusters (in green) showing a subclustering of sequences.

Sequences may also be clustered by hand (Edit->Groups...)

Conservation



Once the sequences have been clustered the patterns of conservation can be calculated and shown for each group. The conservation analysis is based on that in the AMAS program [5] which was itself based on work of Zvelebil et al [6]. Each column in the alignment or group is given a score from 0 to 10 based on the common physico-chemical properties of the residues. The intensity of the colour scheme already present in the alignment is varied according to the score: fully conserved (10) means the most intense colour fading to white for a score of 0. Any colour scheme can be applied before displaying the conservation scores enabling the user to highlight any combination of residues/properties.

In the example PFAM [14] globin alignment below the sequences have been grouped from a dendrogram into 4 groups. The whole alignment has then been coloured according to the hydrophobicity of the residues (Colour->by hydrophobicity) with red being most hydrophobic and blue being hydrophilic. The conservation of each of the groups has then been calculated (Calculate->Conservation) and the intensities of the columns in each of the groups are automatically varied according to the conservation score. The 2 heme binding groups (in blue) can be seen as well as the characteristic 4 or 5 hydrophobic periodicity of the helices e.g. columns 75,79 and 83 and again in columns 98,102 and 105

Editing

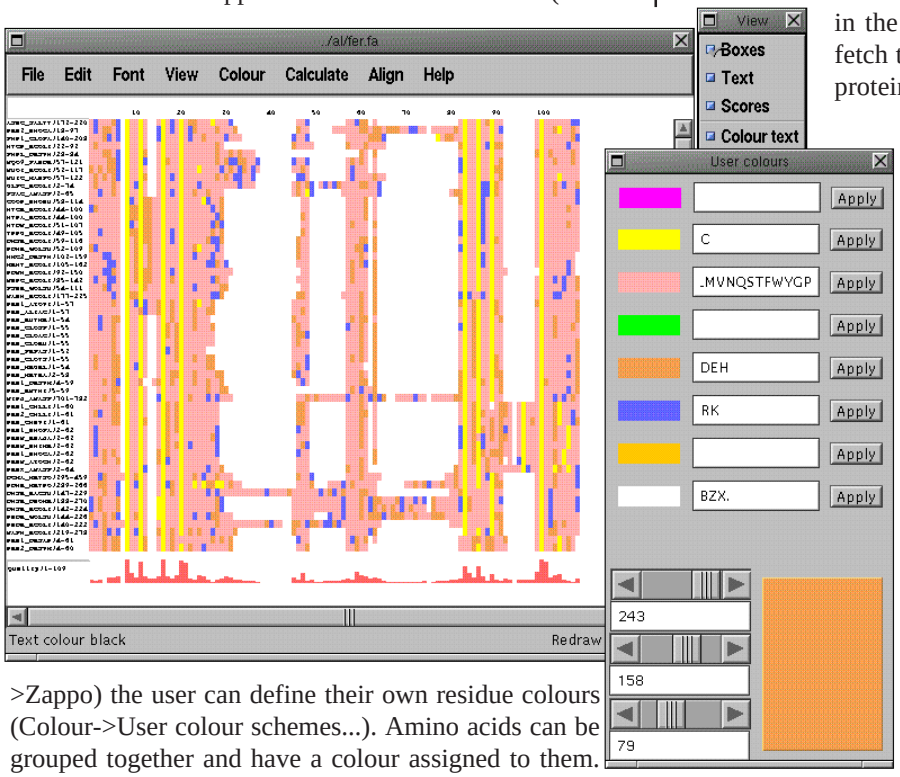
Those hard men (and undoubtedly women) of bioinformatics (HPBs) may eschew any graphical editing of alignments in favour of vi. For us mere mortals having access to the means of quickly recalculating the quality of the alignment and the patterns of conservation within

it without having to go through 3 format conversions, a shell script and the vi beep mode is something to be welcomed. Editing in Jalview is done by selecting a residue with the mouse and dragging left and right to insert or delete gaps. If group editing mode is on (Edit->Group editing mode) all sequences in that group are moved together.

It often happens that a multiple sequence alignment which is based on a database search is 'untidy', i.e. it has ragged edges due to unequal lengths of sequences. Jalview provides the ability (as seen in Belvu by Erik Sonnhammer [10]) to trim the alignment left and right to remove these parts of the alignment. Selecting a column in the top scale panel (where the numbers are) will cause a red box to appear above that column. The alignment can now be trimmed either left or right of this column by choosing the appropriate option in the edit menu. Don't choose the wrong one - There is no undo!!

Colour Schemes

There are a number of pre-formatted colour schemes included in Jalview including Willie Taylor's scheme [7], the ClustalX colour scheme [2] and amino acid hydrophobicity. In addition if the Zappo colour scheme is selected (Colour->Zappo) the user can define their own residue colours (Colour->User colour schemes...). Amino acids can be grouped together and have a colour assigned to them. In the example above is a ferredoxin alignment which has had the font size reduced to 4 (Font->Size=4) to give an overview of the alignment and the text switched off (View->Text) to emphasize the colours. The zappo colour scheme has then been changed to colour only the charges and cysteines (in yellow). This shows up an error in the alignment in the 7th column of cysteines where a gap has been put in the wrong place. The quality profile in pink along the bottom also shows a reduced score in this column compared to the other cysteines.



Sequence feature and structure display

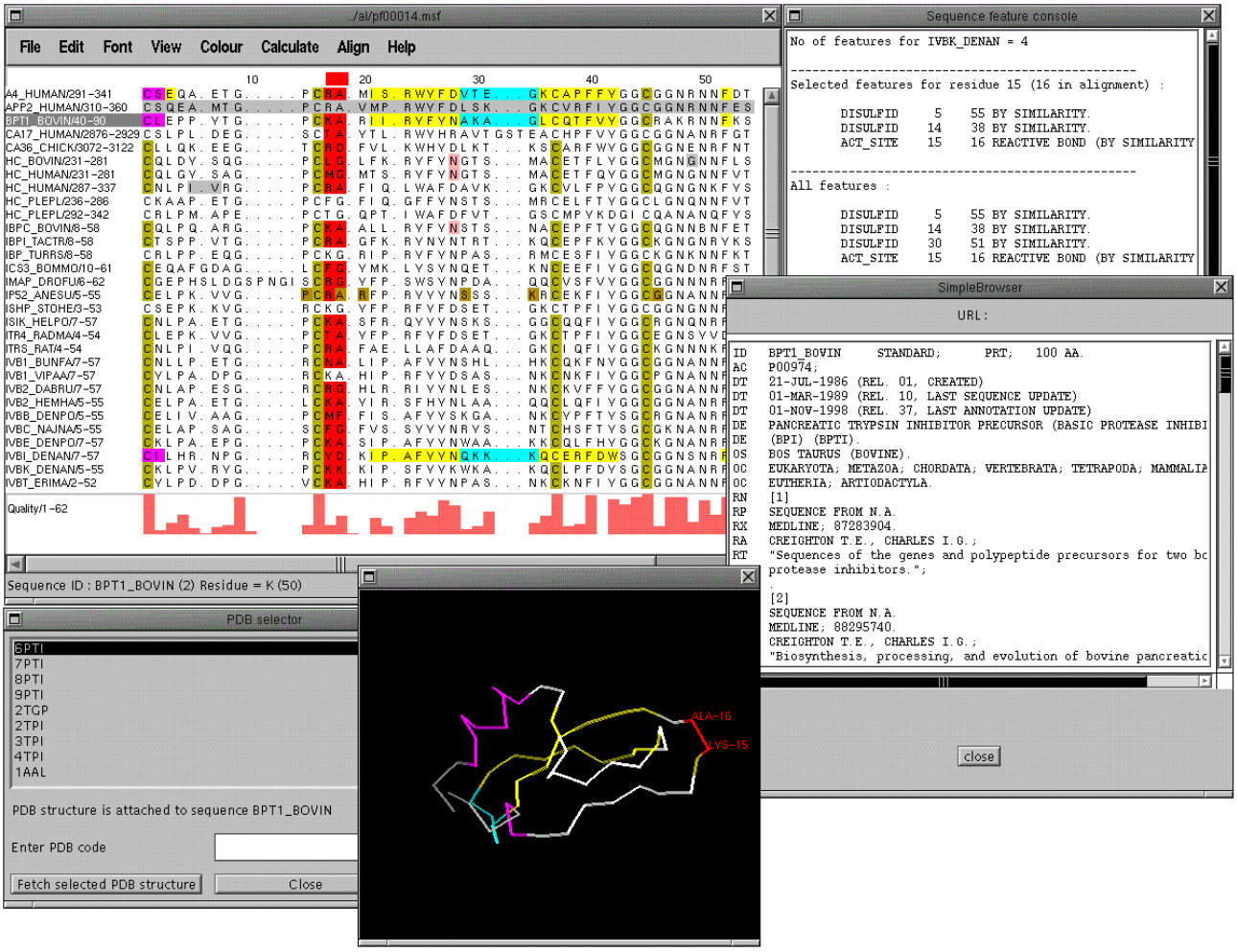
When constructing an alignment or just making sense of a database search browsing through the feature tables of the database entries can give extra insights. The database entries for individual sequences can be retrieved by SRS [8] and displayed either in a new browser window or in the Jalview mini-browser if running as an application. To display the sequence features in colour on the alignment choose Colour->View sequence features. If the sequence IDs are the database IDs the features are attached to that sequence and displayed in the main alignment window. Selecting any feature with the mouse will give details about it and the rest of the features attached to that sequence in a separate window.

The alignment below shows a PFAM [14] pancreatic inhibitor alignment where the active site is coloured red and the cysteines involved in disuphide bonding are in dark yellow. Sequences that have structural features defined show helices as magenta, sheets as yellow and turns as cyan. In addition, if there are any PDB codes present in the database entry SRS is again used to fetch the 3-dimensional coordinates for that protein, dynamically align it to the sequence and display it in a PDB viewer. The colour scheme present in the alignment is also displayed on the structure.

Analysis on remote servers

Of course not everything can be done on the client side. Jalview has the ability to run programs either locally (if running as an application) or remotely using CGI. Below is the result of running Ian Holmes' POSTAL [9] application on an alignment which returns a score for each residue according to how probable that each residue is in the correct place in the alignment. The scores are displayed using a colour scheme where ambiguous portions of the alignment have a dark purple colour underneath them and well-defined regions of the alignment have white.

When a remote or local program runs a console window is displayed showing the length of time the program has been

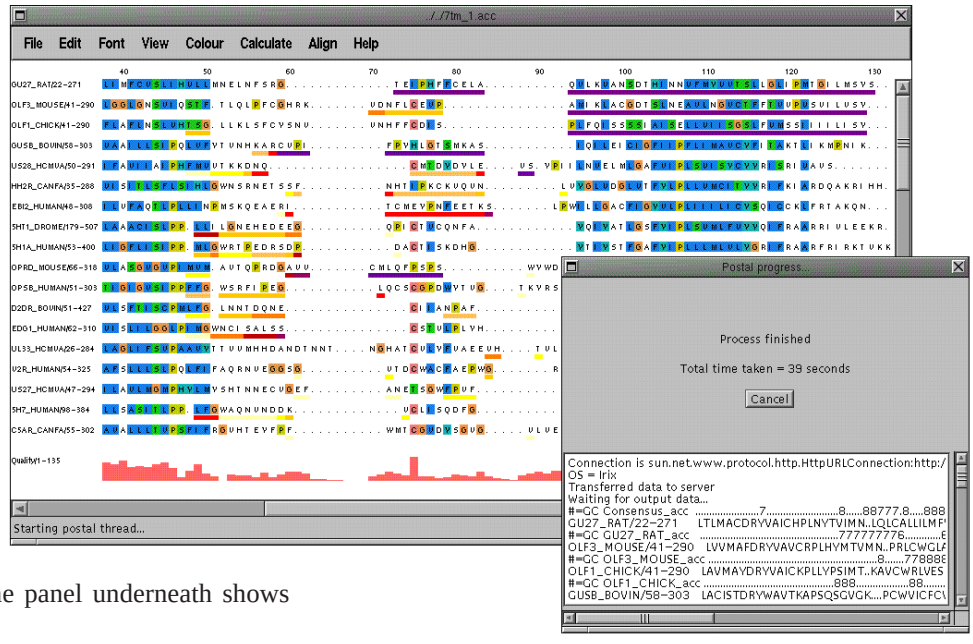


running and any output that may have come back from the server/program. Pressing the cancel button will cancel the job

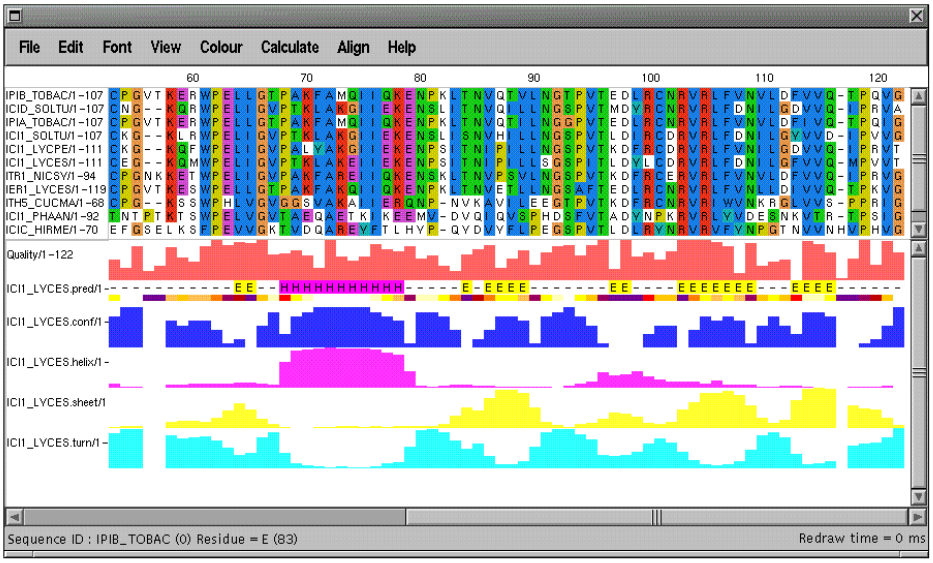
the result when Jalview has taken the full length sequences and realigned them (using clustalw) to the query sequence. The alignment now has far fewer gaps and similarities to the first portion of the query sequence (residues 1-40) have appeared which weren't apparent before.

Alignment of blast results

The results of a blast search can often be only fragments of sequences that have a high enough score to the query to be reported. Extracting the full protein sequence and realigning to the query can give a fuller alignment. Jalview can take as input the output of the blast parser MSPcrunch and extract the full sequences from SRS and realign them to the query. The example below shows in the top panel the individual blast1.4 hits to a protein. The alignment has lots of short sequences and the same protein appears more than once in separate lines. The panel underneath shows



Secondary structure prediction



A multiple sequence alignment is often used for predicting the secondary structure of a protein. Jalview is currently used to view the output of Jpred[11], a consensus secondary structure prediction server at the EBI. In addition, a fast neural network prediction method is available on request an experimental server at the EBI written by James Cuff. This is available directly from the Jalview interface and shows Jalview's ability to display the alignment with a prediction and confidence scores for that prediction. In this case the scoring is for secondary structure but the file format accepted (a variant of AMPS [3] BLC format) could contain scores for any property of the alignment. Applications of this kind where the prediction only takes a second or so (as opposed to Jpred) are ideal for interactive alignment editors. The alignment can easily be changed manually and the structure predicted again to see what, if any, differences occur.

References and links

1. Thompson et al (1994), Nucleic Acids Research, 22, 4673-4680. ftp://ftp.ebi.ac.uk/pub/software/unix/clustalw and ftp://ftp.ebi.ac.uk/pub/software/dos/clustalw
2. Thompson,J.D. et al (1997), Nucleic Acids Research, 24 4876-4882. ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/
3. Barton, G. J. (1990), Methods in Enzymology, 183, 403-428.
4. G. Casari, C. Sander and A. Valencia (1995), Nature: Structural Biology 2. http://industry.ebi.ac.uk/Seqspace
5. Livingstone, C. D. and Barton. G. J. (1993), CABIOS 9, 745-756.http://barton.ebi.ac.uk/servers/amas_server.html
6. Zvelebil, M. J. J. M. et al (1987), J. Mol. Biol.,195 957-961.
7. W. Taylor (1997), Protein Engineering 10 743-746.
8. http://srs.ebi.ac.uk
9. I. Holmes and R. Durbin (1998), Journal Computational Biology 5, 493-198.http://www.sanger.ac.uk/Users/ihh/postal.html
10. http://www.sanger.ac.uk/Users/esr/Belvu.html
11. Cuff, J.A. et al (1998), Bioinformatics (in press). http://circinus.ebi.ac.uk:8081.
12. N. Galtier, M. Gouy and C. Gautier (1996), CABIOS 12 543-548.\ftp://biom3.univ-lyon1.fr/pub/mol_phylogeny/seaview/
13. T. K. Attwood et al EMBnet.news 3 (3).http://www.biochem.ucl.ac.uk/bsm/dbbrowser/CINEMA2.1/
14. E. L. Sonnhammer et al (1998), Nucleic Acids Research 26:320-322 http://www.sanger.ac.uk/Software/Pfam

Node News

Node news from Finland

The Finnish EMBnet node, CSC, has had several changes and rearrangements during the few past months and some of these processes are still going on. As Erja Heikkinen is having one years leave from her position in CSC and EMBnet, new faces have appeared at CSC. During the coming year, Kimmo Mattila will act as the EMBnet node manager together with Timo Kervinen, who will be the main person responsible for the customer support of bioinformatics at CSC. Timo's background is in microbiology but his recent studies also include computer science. Kimmo's previous studies have related to molecular modeling and structural biology.

Another major change at CSC is the replacement of the old SGI Onyx and Power Challenge servers with a new SGI Origin 2000 server. The first setup of the new server, with 24 R10k processors, is already in use and during the coming months the machine will be updated so that the final construc-

tion will have (at least) 86 R12k processors with 107.5 Gb of shared memory. The appearance of the new computing resources gives long awaited relief from the lack of CPU and memory capacity, but unfortunately the new machine is not reserved for the bioscientists but is used for the whole of Finnish academia.

Whilst moving to a new server CSC is also updating its software. CSC has purchased the WWW-based SeqWeb user interface for GCG. As SeqWeb is not designed for nationwide usage on the internet, but rather to be used with intranet connections, some in-house modifications must be made to the code. After negotiations, the Genetics Computing Group gave permission to do the modifications and now the work is underway. The main difficulties have been in setting the balance between the security of the server and the authentication of the users against the flexibility and easy maintenance of the services. We hope to be able to present our solution to these problems, which will became more and more common in other services, in the next issue of embnet.news. Meanwhile, we are more that interested to hear about the WWW related problems and solutions other nodes have figured out.

Node News from the German EMBnet node, GENIUSnet, at the German Cancer Research Centre (DKFZ) in Heidelberg.

Hardware - HUSAR, the "Heidelberg UNIX Sequence Analysis Resources", are currently migrating from two CONVEX SPPs to two SUN "Ultra-Enterprise" machines with six processors each. Both computers will be solely dedicated to HUSAR, one for internal the other for external users. This move allows us to finally upgrade to GCG9.1. We are also looking forward to improved performance and easier access to many Solaris-specific software products.

WWW Interface - Our W2H web interface has been developed further in close collaboration with Martin Senger at EBI. The latest feature is a "simple mode" especially designed for inexperienced users. Many users found the "working list" concept (derived from GCG's WPI) unintuitive, so the new interface optionally replaces it by a simple file browser. Usage of the WWW interface has increased significantly since introduction of the "simple mode".

Co-workers - Several changes in the HUSAR team are about to take place. Dr. Ge Zhang, responsible for web programming, is moving to LION AG here in Heidelberg. Dr. Martin Ebeling, the current EMBnet manager, will join the bioinformatics group of Hoffmann-La Roche in Basel, Switzerland.

Software - An increasing number of users in Germany are asking for automated, tailor-made software solutions to their favourite everyday computational tasks. They gladly accept a variety of scripts we develop especially for this purpose. One of those has grown "out of control" and has become a fully-fledged program built into HUSAR. It is called ESTCLUSTER and it automates iterative BLAST database searches and contig assembly steps. The program does not aim at clustering whole databases but allows the user to interactively perform and monitor a clustering attempt with any of his own sequences. Originally developed for a DKFZ user it is now in wide-spread general use. A web interface to ESTCLUSTER is also accessible from our homepage (but still requires a HUSAR account; test accounts are available upon request).

Node News UK

CONSOLIDATION OF THE UK BIOINFORMATICS SERVICES: HGMP-RC AND SEQNET

BBSRC and MRC are pleased to announce the consolidation of the bioinformatics services provided by SEQNET and the MRC Human Genome Mapping Project Resource Centre (HGMP-RC). This decision has been taken following user consultation, expert consideration, and discussions between the relevant parties including the funding bodies (BBSRC and MRC), employers and grant holders (CCLRC) and the relevant staff located at Daresbury Laboratory and Hinxton.

SEQNET and HGMP-RC both provide access to molecular biology databases and software. An overlap in the scientific focus of the two services currently exists in the area of genomes and protein sequences. SEQNET, which services a broad community of biological scientists, has focused on protein structure and protein structure prediction. The HGMP-RC service has provided data resources and analysis tools for the medical community relevant to genome information.

In the ten years since the start of the UK HGMP there have been a huge number of achievements. Detailed maps of human and mouse have been produced by linkage, radiation hybrid, EST and STS content mapping. The complete sequences of many prokaryotes and a yeast have been determined and the invertebrate nematode sequence is almost complete. Fruit fly, plant (Arabidopsis), mouse and human sequences will all be completed in a few years.

We have entered the era of functional genomics and proteomics that relies on information derived from the study of all these genomes and their transcribed RNA and translated protein products. The combined services of SEQNET and HGMP-RC will enable users a single point of access to this information to help answer such questions as:

- When and where are genes expressed in cells and organisms?
- Which proteins interact with each other in signalling mechanisms?
- Which proteins are related to each other in functional pathways?
- What are the determinants of health and disease?

The consolidated service will be located at the Hinxton site, alongside the Sanger Centre and the European Bioinformatics Institute. The BBSRC sponsored Collaborative Computational Project 11 (CCP11) in Biosequence and Structure Analysis will be relocated to Hinxton in order to assist co-ordination activities for bioinformatics in the UK.

A consolidation of these activities will offer the following advantages:

- It will provide users of these services with a far better quality service than can be individually obtained.
- It will provide a new service, that is both integrated and comprehensive, meeting the needs of the post-genome challenge.
- It will avoid duplication of effort in both software and database support.
- It will combine the expertise of both groups thereby providing fertile ground for capitalising on the potential of such expertise and encouraging the co-operative development of essential software.
- It will promote the formation of a single UK bioinformatics node, incorporating a national UK node of EMBnet and the HGMP specialist node.

The service will be enhanced by the numerous research activities taking place in all three institutes at Hinxton. The transfer of SEQNET will take place in late 1998 and early 1999, with the new integrated service fully operational from 1 April 1999. CCP11 will remain closely associated with the service.

Node News Switzerland

Our grant from the Swiss National Science Fund for EMBnet activities has been extended to Dec 31, 1999 at 23:59:59 (after that, who knows what will happen anyway?). It currently pays for 1.9 salaries and some sundry expenses like the GCG license.

The other bit of news is that we moved into new offices in early October, located in the ex-cafeteria space of ISREC (the Swiss Institute for Experimental Cancer Research). We have a beautiful view over Lake Geneva and the winter sun right in our eyes. We are working on matching the Irish Node by having a pub downstairs - for the time being we make do with an occasional keg of Guinness.

Our computer infrastructure has also been significantly upgraded (courtesy of the Ludwig Institute) and we now have two dual-processor Sun Enterprise servers (one 250 and one 450), a dedicated

BLAST server (a 4-processor PentiumPro) and Very Soon Now increased BLASTing capacity through two dual-processor Pentium IIs. The BLAST jobs will be despatched to one of the three machines according to the database(s) being searched, so that we can keep the databases in RAM at all times. We have contributed an article to this issue of embnet.news about our Better BLAST server and client.

Node News Ireland

As noted in the Swiss Node News, we are indeed located over a pub. It is not just any pub however, but that in which large parts of the development of clustal were carried out by Des Higgins in the mid 1980s.

The funding source for the Irish EMBnet Node has changed yet again, so that we now are supported almost exclusively by a strategic research grant from the Higher Education Authority via the host institution Trinity College Dublin. Reasonable stability is thus promised into the next century; midnight on St. Andrew's Day 2000 to be precise.

The job description has been significantly changed from providing software and databases to providing courses and training.

Node News from the Sanger Centre

The first sequence of an animal genome is essentially complete .See http://www.sanger.ac.uk/WhatsNew/index.shtml#finwrm

Funded by the Medical Research Council and America's National Institutes of Health, the Sanger Centre and the Genome Sequencing Centre at St Louis have completed a fifteen year project to sequence the complete genome of the nematode worm Caenorhabditis elegans.

Containing less than a thousand cells and 1mm in length, C.elegans seems very different from us but is actually built using remarkably similar principles. Like us it develops from embryo to adult, has a gut, nerves, muscles, skin and around 40 per cent of its genes are closely related to ours.

By comparing worm and human sequences it is possible to identify the related genes, and one can then use the worm to examine their function. From these studies conclusions can be drawn about genetic causes of disease and disorders.

This completed gene sequence gives scientists and health practitioners world-wide valuable information to aid the study of the human body in health as well as in illness and may for example lead to new treatments for disease.

Introducing the press conference at the Royal Society, Professor George Radda, MRC Chief Executive said:

"This is an exciting day for British science. The first complete genomic sequence of a complex organism - an animal, with which the human body can be compared, promises to open a new chapter in the understanding of human health and disease."

Lord Sainsbury, Minister for Science said:

"The completion of this project is a terrific scientific achievement. Not only is it an example of international partnership and co-operation with strong British involvement, but a world scientific first - the first multicellular animal to be completely sequenced. This research will ultimately contribute towards interpretation of other genomes, including the human, and help to ensure that we revolutionise healthcare."

The C. elegans papers appear in a special edition of Science, published 11th December 1998

The above is taken from the Medical Research Council Press Release. For more information please call the MRC Press Office on +44 171 637 6011

The EMBnet Nodes

National Nodes

Argentina
Dr Oscar Grau
IBBM Facultad de Ciencas Exactas Universidad Nacional de LaPlata Argentina
Email: grau@biol.unlp.edu.ar
Tel:+54-21-250497 Fax:+54-21-259223

Australia
Dr Tim Littlejohn
ANGIS Electrical Engineering Building J03 University of Sydney
Sydney NSW 2006 Australia
Email: tim@angis.org.au
Tel:+61 2 9351 2948 Fax:+61-2-9351 5694

Austria
Dr Martin Grabner
BioComputing Centre Vienna University
Computing Centre Dr Bohr Gasse 9 Vienna.
Email: martin.grabner@cc.univie.ac.at
Tel: +43-1-4277-14141 Fax: +43-1-7986224

Belgium
Dr Robert Herzog
BEN Université Libre de Bruxelles CP300 Paardenstraat 67 1640 Sint Genesius Rode Belgium
Email rherzog@ulb.ac.be
Tel: +32-2-6509762 Fax:+32-2-6509767

Canada
Dr Christoph Sensen
National Research Council of Canada Institute of Marine Biosciences 1411 Oxford St Halifax Nova Scotia Canada B3H 2Z1
Email: sensencw@niji.imb.nrc.ca
Tel:+1-902-4267310 Fax:+1-902-4269413

China
Professor Jingchu Luo
College of Life Sciences Peking University Beijing 100871 China
Email: luojc@lsc.pku.edu.cn
Tel:+86-10-6275 5206 Fax:+86-10-6275 1843

Cuba
Dr Ricardo Bringas
Centre for Genetic Engineering PO Box 6162 Havana Cuba
Email: bringas@cigb.edu.cu
Tel:+53-7 218200 Fax:+53-7 218070

Denmark
Mr Hans Ullitz-Moller
BioBase - Danish Human Genome Centre Aarhus Universitet Ole Worms Alle 170-171 DK-8000 Aarhus C Denmark
Email: hum@biobase.dk
Tel:+45-86139788 Fax:+45-86131160

Finland
Dr Kimmo Mattila
CSC Center for Scientific Computing PL 405 (Tietotie 6) 02101 Espoo Finland
Email: erja.heikkinen@csc.fi
Tel:+358-9-4572433 Fax:+358-9-4572302

France
Dr Philippe Dessen
Infobiogen 7 rue Guy Moquet - BP8 94801 Villejuif Cedex France
Email: dessen@infobiogen.fr
Tel:+33-1-45595241 Fax:+33-1-45595250

Germany
Dr Martin Ebeling
Department of Molecular Biophysics (0810) German Cancer Research Centre Im Neuenheimer Feld 280 69120 Heidelberg Germany
Email: m.ebeling@dkfz-heidelberg.de
Tel:+49/6221-42-2342 Fax:+49/6221-42-2333

Greece
Dr Babis Savakis
FORTH Insitute of Molecular Biology PO Box 1527 711 10 Heraklion Crete Greece
Email: savakis@nefeli.imbb.forth.gr
Tel:+30-81-212647 Fax:+30-81-231308

Hungary
Dr Endre Barta
Agricultural Biotechnology Centre Szent-Gyorgyi u. 4 PO Box 410 2100 Godollo Hungary
Email: barta@abc.hu
Tel:+36-28-430127 Fax:+36-28-420096

India
Prof MW Pandit
Centre for DNA Fingerprinting (CDFD) CCMB Campus Uppal Road Hyderabad 500 007 India
Email: cdfddbt@hd1.vsnl.net.in
Tel: +91-40-7150008

Ireland
Dr Andrew Lloyd
INCBI Dept Genetics Trinity College Dublin 2 Ireland
Email: atlloyd@tcd.ie
Tel:+353-1-608-1969 Fax:+353-1-679-8558

Israel
Dr Leon Esterman
Biological Computing Division Weizmann Institute of Science Rehovot 76100 Israel
Email: lsestern@weizmann.weizmann.ac.il
Tel:+972-8-9343934 Fax:+972-8-9466269

Italy
Dr Marcella Attimonelli
Area di Ricerca CNR-BARI Via Amendoal 166/5 70126 - Bari Italy
Email: marcella@area.ba.cnr.it
Tel:+39-80-5482130 Fax:+39-80-5484467

Netherlands
Dr Jack Leunissen
Caos/Camm Centre University of Nijmegen Toernooiveld 6525 ED Nijmegen Netherlands
Email: jackl@caos.kun.nl
Tel:+31 24 365 22 48 Fax:+31 24 365 29 77

Norway
Ms Karin Lagesen
Biotechnology Centre of Oslo University of Oslo
Gaustadalleen 21 0317 Oslo Norway
Email: karin.lagesen@biotek.uio.no
Tel:+47-22958756 Fax:+47-22694130

Poland
Dr Piotr Zielenkiewicz
Institute of Biochemistry and Biophysics Polish Academy of Sciences Pawinskiego 5a 02-106 Warsawa Poland
Email: piotr@ibbrain.ibb.waw.pl
Tel:+48-2-6584703 Fax:+48-39-121623

Portugal
Dr Pedro Fernandes
Instituto Gulbenkian de Ciencia Rua da Quinta Grande Apt. 14 2781 Oeiras Codex Portugal
Email: pfern@pen.gulbenkian.pt
Tel:+351-1-443 1408 Fax:+351-1-443 5625

Russia
Professor Sergei Spirin
Belozersky Institute of PhysicoChemical Biology Moscow State University Laboratory Korpus A - Room 612 119899 Vorobyevy Gory - MOSCOW Russia
Email: sas@brodsky.genebee.msu.su
Tel:+7 (095) 932 8825 Fax:+7 (095) 939 3181

South Africa
Dr Win Hide
SANBI Private Bag X17 Bellville 7535 University of the Western Cape South Africa
Email: winhide@techno.sanbi.ac.za
Tel:+27 21 959 3645 Fax:+27 21 959 2512

Spain
Dr JoseRamon Valverde
CNB Universidad Autonoma de Madrid Campus de Canto Blanco 28049 Madrid Spain
Email: jvalverde@cnb.uam.es
Tel:+341-5854543 Fax:+341-5854506

Sweden
Mr Nils-Einar Eriksson
Uppsala Biomedical Centre Box 570 S-721 23 Uppsala Sweden
Email: Nils-Einar.Eriksson@bmc.uu.se
Tel:+46-18-471 40 17 Fax:+46-18-55 17 59

Switzerland
Dr Victor Jongeneel
ISREC Bioinformatics Group Chemin des Boveresses 155 CH-1066 Epalinges Switzerland
Email: victor.jongeneel@isrec.unil.ch
Tel:+41-21-692-5994 Fax:+41-21-653-4474

Turkey
Dr Zehra Sayers
National Bioinformatics Node (NBN) MAM GMBAE NBN POB 21 41470 Gebze-Kocaeli Turkey
Email: zehra@bioinfo.rigeb.gov.tr
Tel:+90-262-6412300 ext 4007 Fax:+90-262-6412309

United Kingdom
Dr Alan Bleasby
SEQNET DRAL Daresbury Laboratory Daresbury Warrington WA4 4AD England
Email: ajb@dl.ac.uk
Tel:+44 1925 603351 Fax:+44 1925 603100

Specialist Nodes

EMBL-EBI
Dr Rodrigo Lopez
EBI Hinxton Hall Hinxton Cambridge CB10 1SD England
Email: rls@ebi.ac.uk
Tel: 1223 494438 Fax:+44 1223 494 468

ETI
Dr Peter Schalk
ETI biodiversity Center Universiteit van Amsterdam Mauritskade 61 1092 AD Amsterdam theNetherlands
Email: pschalk@eti.uva.nl
Tel:+31-20-5257239 Fax:+31-20-5257238

HGMP-RC
Dr Martin Bishop
HGMP Resource Centre Hinxton Cambridge CB10 1SB UK
Email: mbishop@hgmp.mrc.ac.uk
Tel:+44 1223 494500Fax: +44 1223 494512

Hoffman-LaRoche
Dr Daniel Doran
Pharma Preclinical Research Hoffman-LaRoche CH-4002 Basel Switzerland
Email: daniel.doran@roche.com
Tel:+41 61 688 8270 Fax:+41 61 688 1075

ICGEB
Dr Sandor Pongor
ICGEB Padriciano 99 34012 Trieste Italy
Email: pongor@genes.icgeb.trieste.it
Tel:+39 40 3757300 Fax:+39 40 226555

MIPS
Dr Werner Mewes
MIPS Max Planck Institut fur Biochemie Am Klopferspitz 18a D-82152 Martinsried Germany
Email: mewes@mips.biochem.mpg.de
Tel:+49 89 8578 2656 Fax:+49 89 8578 2655

Pharmacia
Dr Staffan Bergh
Pharmacia-Upjohn AB Strandbergsgatan 49 112 87 Stockholm Sweden
Email: staffan.bergh@eu.pnu.se
Tel:+46-8-6959884

Sanger Centre
Mr Peter Rice
The Sanger Centre Wellcome Trust Genome Campus Hinxton Cambridge CB10 1SA England
Email: pmr@sanger.ac.uk
Tel:+44 1223 494967 Fax:+44 1223 494919

UCL-BCM
Dr Terri Attwood
Biomolecular Modelling Unit University College London WC1E 6BT England
Email: attwood@bsm.bioc.ucl.ac.uk
Tel:+44 171 419 3879 Fax:+44 171 380 7193

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_4/contents.html***
- ***http://www.be.embnet.org/embnet.news/vol5_4/contents.html***
- ***http://www2.ebi.ac.uk/embnet.news/vol5_4/contents.html***
- ***http://www.ie.embnet.org/embnet.news/vol5_4/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

embnet.news
Vol.5, No.4, 1998
30 December 1998 ISSN 1023-4144