embnet.news

Volume 4 Nr 3 (ISSN1023-4144)

Editorial

As well as being the Christmas Bumper issue, this is also the after EMBnet AGM Issue. The 11th Annual Business Meeting was organised this year by the Italian Node (CNR-Bari) and took place up in the hills at Selva di Fasano at the end of September. For us northerners, the concept of "O for a beaker full of the warm south" so affected one of the delegates that he jumped (or was he pushed ?) fully clothed into the hotel swimming pool. Despite the balmy weather and the excellent food and wine, we did manage to get a solid day and a half of work done.

EMBnet is having to make some difficult choices about what its future direction and purpose should be. Our major source of funds is from the EU, but pretty much all countries which are eligible for EU funding have already joined the organisation. Nevertheless, as the only international organisation comprised of bioinformatic support and service centres, there is an obvious community of interest with similar groups from outside the EU and indeed from outside of Europe. Last year in Helsinki, we were joined by Associate Nodes from China and Australia. This year, the precedent having been established, we welcomed nodes from Argentina, Cuba, India and South Africa. We believe that there will be a valuable two-way traffic of ideas, people, software and expertise between our European core and our colleagues in Greater EMBnet. However, EMBnet must debate and finally agree about how big it wishes to get; there is by no means a consensus of 'expansionists'.

The other side of this coin is a debate about what it is that EMBnet nodes should do and what constituency EMBnet should represent. This can also be cast as a variant of an on-going disagreement about the definition of "bioinformatics". Is it (narrowly) the storage, retrieval and analysis of sequence data or (broadly) most of computational biology ? There is a long established concept of EMBnet Specialist Nodes, which provide a special expertise or service which cannot normally be maintained or supported by National Nodes. This year we accepted the Biomolecular Structure and Modelling Unit of University College London as Specialist Node, joining such institutions as the EBI, SwissProt, the Sanger Centre and the ICGEB. However, we were unable to achieve agreement about how to incorporate expertise in Taxonomy (and later, presumably, ecological genetics, molecular epidemiology etc. etc.). So there needs to be some active debate between 'diversificationists' and those who believe that it is time to retrench to and build upon our core expertise in sequence analysis.

Such debates can only be construed as healthy. Stasis can all too easily become stagnation. After some cliff-hanging recounts and reballots at the AGM there have been changes in all of EMBnet's committees. It is to be hoped that new committee members will help galvanise us all into a more active phase after a relatively quiet 1997. The fact that the financial status of EMBnet is presently very healthy, will certainly not impede this drive. Everyone agrees that bioinformatics is one of science's growth areas and there is nobody better equipped than EMBnet to make solid contributions to the field.

The clustalWWW server at EBI

Andrew Lloyd

Within the next couple of weeks, I will be moving to a new office immediately above Kennedy's Bar in Dublin where many of the early bugs in the Clustal multiple alignment program were ironed out. I have been a regular user of Clustal in its 1, 2, 3, 4, V, and W incarnations since about 1990. I've been running courses in bioinformatics for about the same length of time and always exhort students "don't

Contents

take the defaults". We all agree that the defaults are chosen, not arbitrarily, but to give meaningful results in a majority of cases.

However, a full, comprehensive and publishable bioinformatics analysis should have investigated the robustness of the results given a variety of input parameters. Blast searches, for example, should be tried with different substitution matrices, and with and without low complexity masking. For software developers there is always a trade-off between making the program user-friendly and accessible and making it powerful and flexible. GCG works well because it is internally consistent: if you've run one GCG program you can guess at how to run any other.

The menus of ClustalW also show this sort of consistency, so that with very little experience, it becomes mechanical to shuffle a file full of sequences into a multiple alignment and even generate a phylogenetic tree. But you have to tell naive users that there is a wealth of options, alternatives and parameters down there in the bowels of the program and encourage their use.

So what comes after ClustalW ? Well one strand of development has produced ClustalX and another has given us http://www2.ebi.ac.uk/clustalw/

Also known, with the wit that is so often devoted to choosing a clever name, as ClustalWWW. And very good it is too. On my first trial of the service, I uploaded a file of 23 recA sequences, played with the options a little, remembered how slow things had been on the PHD server in Heidelberg and chose to have the the results sent back to me by e-mail.

After getting a message saying "Clustalw Mail server is not ready", I switched to "interactive" results, submitted the job and started to go for coffee. Before I'd picked up my empty cup, the results were delivered ! Even in the middle of the day, ClustalWWW works faster than my DEC alpha by almost an order of magnitude.

This is obviously a strong reason for recommending it. Another one is that all the available options are "in your face" on the top page. There is no excuse for my naive users to know nothing about gap penalties, gap extension penalties or alternative substitution matrices. There is even a nice innovation offering to display the alignment colour coded.

So it's pretty and pretty useful too for beginners. Web tools are particularly attractive for training courses because "everyone" is familiar with the conventions of browsers while they may not, yet, be familiar with the conventions of GCG. In its present state though, ClustalWWW does not replace clustalW as a research tool. There is no option for drawing a final Neighbor-joining tree, there is no option for uploading your own substitution matrix, there is far less flexibility in choosing gap (extension) penalties.

I am sure that all these issues are already on a "Things To Do" list in Hinxton, and I happen to know that the interface is in a process of revision. Let's hope that the EBI can devote the time and money to continue the development of this tool so that it becomes as rich and famous as its immediate ancestor ClustalW.

MIME types and ClustalW

What is a MIME type? MIME stands of 'Multipurpose Internet Mail Extensions'. These extensions allow an e-mail program or a WWW browser to react in a user configurable way when a file is sent using a MIME type. This means that the browser will launch an application that knows how to deal with the file (*).

The clustalw sequence alignment service at http://www2.ebi.ac.uk/clustalw/ is capable of returning the clustalw result files (the dendrogram tree (.dnd) and the sequence alignment files (.aln)) using MIME types.

These are very useful extensions to your WWW browser if you know and understand how to use them. Lets first assume that you have a browser that can correctly interpret MIME types such as netscape3, netscape communicator or IE3 or IE4.

Assume the browser is correctly installed on a MAC, Win95/NT or UNIX machine. When you click on the the .dnd and/or .aln file hypertext links of the results page of a clustalw run, your browser should open a dialog asking you to save the file or open it.

If you decide to open it the browser should ask which application it should attempt to launch to view the file with. Simply browse/locate the application you want to use. You have to install one of following programs on your computer:

belvu - UNIX Multiple sequence alignment viewer
ftp://ncbi.nlm.nih.gov/pub/esr/belvu/

njplot - UNIX, Mac & PC - Tree viewer
http://acnuc.univ-lyon1.fr/phylogeny/njplot

GeneDoc - GCG MSF file viewer (NB - Use GCG format in the OUTPUT option)
http://www.cris.com/~Ketchup/genedoc.shtml

(*) There is an article about MIME in a previous issue of embnet.news Vol.2.2, so if the reader wants a more etailed explanation then point the browser to: http://www2.ebi.ac.uk/embnet.news/vol2_2/mime.html

TreeView - Tree viewer for PC running Windows
http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

clustalx - UNIX, MAC & PC MSA program:
ftp://ftp.ebi.ac.uk/pub/software/unix/clustalw/clustalx
ftp://ftp.ebi.ac.uk/pub/software/mac/clustalw/clustalx
ftp://ftp.ebi.ac.uk/pub/software/dos/clustalw/clustalx

If you are a UNIX user and if you feel adventurous you can edit either your .mailcap or .mime.types files and add to it the following two lines: application/x-tree njplot %s (if you have choosen njplot to handle the MIME type) application/x-align clustalx %s (if you have choosen clustalx to handle the MIME type). Thats it! Nothing else is required to make your browser react to a MIME type. Enjoy!

INSECT and MOLLUSCS supercomputing on the cheap

Victor Jongeneel, Thomas Junier, Christian Iseli, Kay Hofmann and Philipp Bucher
ISREC Biocomputing Group and Swiss EMBnet node, 1066 Epalinges, Switzerland

The problem

Much of sequence analysis involves comparing a query sequence or pattern to a reference database. In general, the time required to complete such a search is directly proportional to the size of the database. With the relentless growth of the databases, and in spite of almost equally rapid progress in computer performance, exhaustive searches are becoming more and more time-consuming, to the point of making some promising analytical approaches impractical.

There are basically three ways to tackle the problem, which are not mutually exclusive:

1 Develop more efficient algorithms, and fine-tune them to require as little computation as possible. This approach was taken by the developers of the BLAST and FASTA algorithms, and explains their popularity to a large extent. Unfortunately, there is always a price to pay in terms of the sensitivity of the comparison and the quality of the statistics. These limitations make heuristic algorithms unsuitable for exploring distant evolutionary relationships.

2 Buy more powerful general-purpose computers (high-end servers, "supercomputers"). While this guarantees flexibility in the choice of algorithms, it is usually a very expensive proposition. Also, very few centres can afford to dedicate a high-end server to the tasks of database searching alone: usually, these jobs have to compete for CPU time with housekeeping tasks (e.g. reformatting databases), user interactive sessions, general-purpose sequence analysis (GCG), and sometimes unrelated activities such as molecular dynamics calculations from the Chemistry Department. Additional speedup can sometimes be achieved with a multiprocessor architecture, provided the program code supports threading. This is far from always the case. Also, multiprocessor machines are more complex, and thus more expensive, than single-processor models.

3 Buy a dedicated sequence comparison processor (Bioccelerator, FDF), whose hardware has been designed for this purpose. This is an increasingly popular option, as the performance of these machines far outstrips that of a similarly priced workstation. A serious drawback, however, is that the machine architecture limits the flexibility of the algorithm. While this could in some cases be alleviated by reprogramming the microcode of the FPGA chips, there is no accelerator available at the present time that can accommodate a full complement of useful algorithms.

Our approach has been to develop an Inexpensive Networked SEquence Comparison Technology (INSECT) based on the "pile of PCs" concept. Standard desktop computers have been equipped with increasingly powerful processors, mostly to keep up with increasingly bloated software products and fancy user interfaces. Being mass-produced commodity items, their prices have steadily plunged. As the raw processing power of a cheap Intel (or compatible) or PowerPC processor differs by much less than an order of magnitude from top-of-the-line DEC, Sun or MIPS processors, one can naively assume that 10cheap processors should handily outperform a single expensive one. Our experience amply proves this point.

Hardware

In principle, any cheap hardware from your favourite clone vendor will do. You can even recycle old machines that were deemed too slow by some local "power user", provided you balance database sizes according to CPU performance (see below). In our case, since we had a little money to spend, we bought the following hardware:

- Basic clones with a Cyrix/IBM 6x86 PR200+ processor, 512KB pipelined burst cache, 16MB EDO RAM, 2GB EIDE hard disk, basic video card, no keyboard or mouse, no monitor. Such a system currently costs about CHF 950 in Switzerland (your mileage may vary).
- Cheap Addtron AN16CT ISA network cards (CHF 35 a shot) and a basic 16-port 10BaseT hub
- A manual keyboard and monitor switch (CHF 285 for a 12-port model)
- Miscellaneous cables (network, keyboard, monitor) to link

it all together (about CHF 500 total). We also scrounged up an old monitor and keyboard from our spare parts reserve.

For a system with one master and 8 slaves, total cost was about CHF 10000 (about £4000 or $6500 at current rates). The main problem was to find space for the equipment, which landed on an old utility cart now decorated with spaghetti wiring. In a next incarnation, we may try to use an industrial rack with structured cabling, central power supply, etc. etc.

The master machine (which also sports a CD-ROM drive to simplify software installation) received two network cards, one to communicate with our in-house network and the Net, and the other to talk to the slaves. Linux 2.0.30 was installed on all the machines (an easy duplication from the Master), with kernel patches to power down the processor when idle. The entire hard disk of each slave is exported to the master. The INSECT network is invisible to the outside, and has its own (fake) IP domain, "beehive.org".

Software

The basic concept is simple: each slave receives a portion of each searchable database on its local hard disk, proportional to its processor power (equal in our case). It accesses the executables and configuration files for the analysis software in a shared directory structure exported by the master. When it receives instructions to start a job, it performs a search on its chunk of the database and returns the results to the master. The master is responsible for scheduling the jobs, post-processing the results, and keeping order among the slaves.

We decided against using PVM or some other sophisticated task scheduler, mostly because of stability concerns (our experience with PVM had been mixed at best). Instead we developed a MOdular Low-cost Linux-based Unified Sequence Comparison System (MOLLUSCS). The role of MOLLUSCS is to provide the user with a Unix command line as similar as possible to the one used on a traditional system and to handle the details of dispatching, process creation, data collection, post-processing and cleanup.

The current [05/08/97] version of MOLLUSCS consists of one core script (mollusc.pl), a Perl module for each of the biological programs (e.g., Pfscan.pm), and auxiliary modules. Features include:

- The biological program's syntax is conserved as is, and can be invoked by preceding it with 'mollusc' and mollusc options (if necessary)
- Each program module is self-loading, so only the code for the relevant program gets compiled
- The job of each child is specified as a short Perl script which is constructed in the program module and gets executed in the main module by a call to eval - this ensures maximal flexibility
- A set of auxiliary subroutines (CLAux.pm), designed for use in the biological program modules, provides functions for manipulating the command line, for example to force some options and reject others
- Interrupt signals are trapped and a cleanup is attempted before exiting.
- mollusc jobs can be sent into background by & or ^Z and invoked from a remote machine via rsh/remsh

Modules exist currently for ssearch3 (Smith-Waterman searches, W. Pearson), pfsearch and pfscan (generalised profile against sequence database and sequence against profile database, Ph. Bucher), and pattern_find (search database with extended regexp, K. Hofmann); an additional module will be developed for searchwise (framesearch with differential gap scores, E. Birney). We have also provided an easy way for MOLLUSCS to be invoked from CGI scripts, and have thus been able to incorporate the INSECT into our Web-based services.

An additional utility takes care of splitting the databases into chunks and distributing them to the slaves. The percentage of the database given to each slave can be specified, to account for possible differences in performance.

Performance

We have done some rough testing of the INSECT's performance. The following table gives some preliminary data on performance with the pfsearch (search the yeast protein database, 6141 entries, with a 53-aa profile) and ssearch (search the same database with a 72-aa peptide, using the Smith-Waterman algorithm) programs. Times are in seconds. More precise results may be obtained by using longer queries and larger databases.

Hardware (compile options)	pfsearch	ssearch3
Sun Sparc 20	786	29
HP 735/125 (+O4)	92	21
DEC3000 (gcc -O4)	164	25
Pentium 166 (Linux, gcc -O6 -pent)	210	18
Dual PentiumPro 200 (Solaris, gcc -O6)	112	22
SGI Origin2000 server	36	10
Single slave	172	22

It is already obvious from these preliminary data that the INSECT performs extremely well compared to other hardware we have available in Lausanne, including a brand-new SGI Origin2000 that cost almost 20 times more. We have observed another 2-fold increase in speed when we moved to the "full" configuration of one master and fifteen slaves.

We have also run the INSECT through the test devised at the EBI (see http://industry.ebi.ac.uk/~thanaraj/seqassess/repedit3.html) to compare the Bioccelerator, the FDF and the MasPar sequence comparison accelerators. On the INSECT, we ran the ssearch3 program using the blosum62 scoring matrix, and with default gap opening and extension penalties. We repeated the same search on the SGI Origin2000 (4 processors), using ssearch3 compiled either as a standard or as a threaded application. The results show that the INSECT with 15 slaves outperforms to the MasPar for all but the longest queries. The search time for the INSECT increases roughly proportionally to the query length, while for the dedicated machines performance improves. Nevertheless, it is clear that the price/performance ratio still strongly favours the INSECT (cost of about 15000 CHF for the configuration tested here, as compared to 60000 CHF and up for the dedicated processors and about CHF 150'000 for the SGI server).
NB: the queries are ranked by length. Details of the test can be found on the EBI Web site. The benchmarks on the SGI Origin2000 were not done in single-user (max. performance) mode.

Raw Search Time (seconds)

Query	Bioccelerator	FDF	MPsrch_pp	MPsrch_ppa	INSECT (15)	SGI (single)	SGI (threaded)
plasto	30.0	11.3	37.88	47.86	24.69	128.9	34.17
calmod	33.0	12.3	42.03	59.02	36.20	187.08	48.72
histone	38.0	14.3	46.17	71.20	45.72	239.39	57.92
riboS3	40.0	14.3	49.89	75.06	48.05	285.65	68.54
vmat	45.0	17.4	56.10	88.72	59.64	366.09	88.67
coat	52.0	20.6	65.61	107.17	76.63	480.09	119.88
amid	60.0	28.6	76.40	131.02	100.72	629.38	159.31
dnak	67.0	32.7	87.38	158.32	155.21	798.34	191.87
efg	76.0	36.3	95.40	174.16	167.49	907.34	240.62
ski	80.0	42.0	99.95	185.44	163.74	964.94	217.75
amdm	85.0	80.0	105.59	194.33	159.49	1031.35	243.44
phsg	92.0	79.9	114.03	212.71	191.71	1157.95	288.07
abl	141.0	118.0	181.77	386.38	426.81	2083.54	523.41
cin2	179.0	146.6	218.59	444.95	556.09	2549.9	646.71

Conclusions

In environments with limited financial resources, such as many of the EMBnet nodes or modestly endowed academic institutions, INSECT technology may provide an attractive alternative to expensive dedicated hardware. Scaling up is easy, in that additional units can be added at any time, but storage, wiring and maintenance may become problematic when too many units are connected. This may be circumvented (at a price...) by substituting industrial motherboards for the PCs, or by introducing structured wiring cabinets and autosensing KVM (keyboard, video, mouse) switches. Using PCs has the added bonus of allowing "recycling" of units between the biocomputing lab and regular office users.

Besides performance at an attractive price, the INSECTs also offer total flexibility in the choice of the most appropriate algorithm and parameters. This is by far not the case when using hardware-accelerated sequence comparison machines. In principle, any Unix-based sequence comparison program can be adapted to run in this environment, by adding appropriate modules to the MOLLUSCS. For example, there is no hardware-accelerated version of the pfscan program, which scans a protein sequence against a database of profiles. We have recently developed a Web interface to pfscan and a PROSITE regexp scanner running on the INSECTs, with access to our own profile collection as well as to a reformatted version of Eddy & Sonnhammer's PfamA HMM collection.

It can be found at http://ulrec3.unil.ch/software/PFSCAN_form.html. To our knowledge, this is the fastest available Web server for protein motif searches.

The code for the MOLLUSCS is available from Thomas Junier (Thomas.Junier@isrec.unil.ch)

Help for setting up INSECTs can be obtained from Christian Iseli (chris@cmpteam4.unil.ch)

GQserver Automatic Annotation of Protein Sequences

Miguel Andrade, Nigel Brown, Angelo Franchini, Sebastian Hoersch, Christophe Leroy, Christian Reich and Chris Sander.

From Dec. 9, 1997, the GeneQuiz Team at EMBL-EBI is making available a new WWW service to the molecular biology research community.

http://www.sander.ebi.ac.uk/gqsrv/submit

Example of results:

http://columba.ebi.ac.uk:8765/GeneQuizServer/00044765023/argi_human/frames.argi_human.html

The researcher submits to the server the amino acid sequence of a single protein (or open reading frame). After the GeneQuiz analysis has finished it returns the WWW address of a report summarising the automatically assigned functional annotation.

The annotation is either retrieved directly from public protein databases or derived indirectly using an expert system module (GQreason). Sequence similarity searches are performed in a non-redundant database kept up to date nightly (GQupdate). In addition, GQserver provides protein family information, including colour-coded multiple sequence alignments, species distribution, keyword digests, and other supporting evidence (GQbrowse).

Users are asked to consider these limitations:

- limited hardware resources: Use this service wisely and do not use it just to run a BLAST search (otherwise available from the NCBI at http://www.ncbi.nlm.nih.gov/BLAST/ or the EBI http://www2.ebi.ac.uk/blast2/)

- limited human resources: No user support is given, but please report evidence of bugs to genequiz@embl-ebi.ac.uk

- limited features: The server uses the software GeneQuiz 3.0 (August 1997), which lacks several important features known to be desirable (see documentation).

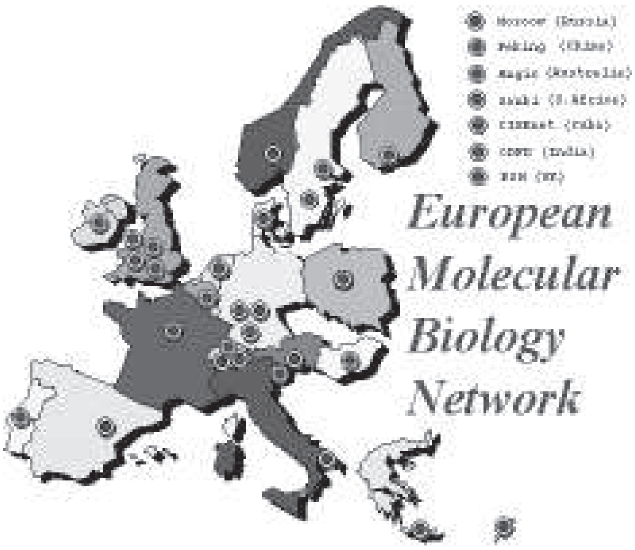
- limited accuracy: The error rate of functional annotation, is estimated to be in the range 1-3 % (2.4 % for Helicobacter pylori as assessed by the team's human quality control) In addition, we continue to make available, we have for the past two years, GeneQuiz tables of functional annotations and supporting evidence for completely sequence genomes, such as Synechocystis and several archaebacteria.

http://www.sander.ebi.ac.uk/genequiz/

Network Performance Monitoring in EMBnet

Jan H. Noordik, J.A.M. Leunissen and K. Cuelenaere

Dutch National EMBnet Node - CAOS/CAMM Center, University of Nijmegen, The Netherlands



Introduction

The European Molecular Biology network EMBnet was established in 1988 to link European laboratories where biocomputing and bioinformatics were used in molecular biology research. The initiators saw the network as a way of bringing a fast growing stream of information to users throughout Europe and to and from, at that time, the EMBL Data Library in Heidelberg. But EMBnet was also seen as much more.

Bioinformatics, and equally the rapidly developing science, required extensive user training and support and on occasion users/researchers would require specialised hardware and software that could not be economically duplicated throughout a country or a group of nations. It was thought that such needs could best be handled by providing national language help and regionally tailored services.
Thus EMBnet swiftly evolved into a series of collaborating national and specialised nodes, spread throughout Europe

and cooperating for their and the users common good. Today EMBnet, as an "Institute without Walls", not only complements Europe's central facilities such as the EBI (European Bioinformatics Institute) in Hinxton (UK), but it is also the defacto collaboration forum for bioinformatics worldwide. Currently EMBnet consists of over thirty partner institutes or EMBnet nodes. The network is organised as a "Stichting" under Dutch law. Funds are mainly obtained from node fees and from a Concerted Action Program grant from the European Commission; (ERBBIO4-CT96-0030).

Network requirements

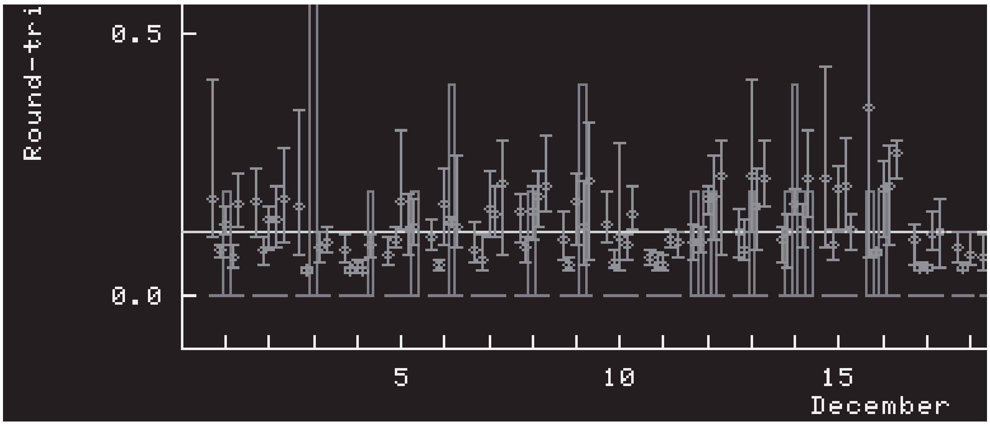
EMBnet nodes maintain daily updated DNA sequence databases and provide login and/or Web biocomputing services for hundreds of researchers throughout Europe, who wish to use their national node's infrastructure and facilities for database searching and biocomputing. These operations involve the daily transport of hundreds of Mb's of data between nodes and between nodes and their end users. Considerable bandwidth availability is a condition for smooth operation. On-line services, as provided by many EMBnet nodes, require fast response times. In its operational strategy, EMBnet's success is critically dependent on both these network parameters; i.e. fast and reliable network connections with moderately high throughput capacity.
Spurred by the rapidly increasing amount of data to be handled in EMBnet, and the as rapidly increasing network traffic in general, by the end of 1994 EMBnet decided to start an ongoing network performance monitoring program. The data to be produced by this program were intended to provide individual EMBnet nodes with objective and reliable figures on their network accessibility. Many nodes expressed an urgent need for this information, to be used in discussions with their local network authorities. In addition the data could be used to verify claims of international e.g. DANTE and national data network providers like SURFNET in the Netherlands, that (international) network performance and the quality of service (QoS), as experienced by the end-user in the network, are constantly improving and that bottlenecks in network traffic are gradually disappearing.

Methods

In the monitoring program described here, network performance data have been collected during 1995, 1996 and 1997 and measurements still continue. For the data collection two simple tools for network testing were used; ping and traceroute. The IRIX Man Pages describe Ping as follows:

"a tool for network testing, measurement and management. It utilizes the ICMP protocol's ECHO_REQUEST datagram to elicit an ICMP ECHO_RESPONSE from a host or gateway. ECHO_REQUEST datagrams (`pings') have an IP and ICMP header, followed by an 8-byte time stamp, and then an arbitrary number of `pad' bytes used to fill out the packet."

For our monitoring facility, we used packet sizes of 64 bits to mimic average traffic in EMBnet, which is usually a mix of telnet for on-line sessions and FTP for data transfer. Ping requests are sent to all EMBnet nodes several times a day on a daily basis from different locations within the network. At 01:00, 09:00, 12:00, 15:00 and 18:00 hrs. a fixed number of five requests is sent, of which round-trip times, RTT's and packet loss statistics are collected. From these data, mimimum, maximum and average values per day are calculated and the daily average RTT values were used (averaged) to an overall monthly indicator, the quality of network accessibility (QNA) for a particular EMBnet node. The variation of QNA's (the average RTT in a period of one month), shows general and node specific improvement or deterioration of network performance as a function of the time. Packet loss data give an indication of node reachability. A high monthly average percentage packet loss in practice means that a node is almost cut off from the network. A 100% packet loss in the daily data indicates the impossibility to reach the node at that specific day. Traceroute data were used occasionally to track specific network hops causing extreme delays.

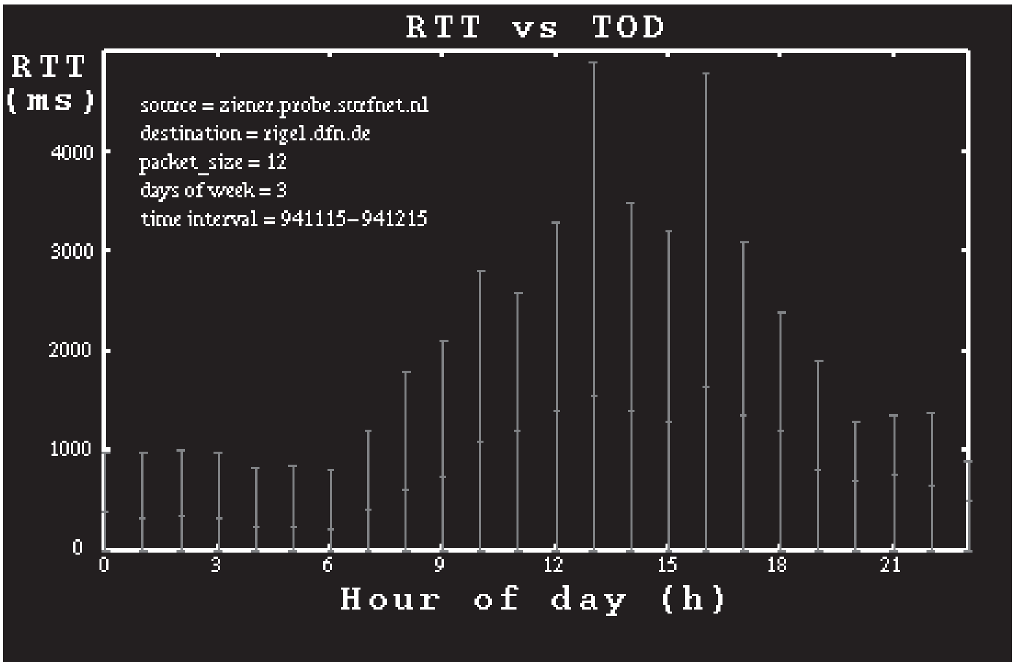


Minimum, average and maximum round-trip time as measured on a specific day in five sets of five ping requests. Round-trip times are given in seconds. Relative packet loss, the number of `lost' packets divided by the number of packets sent, ranging from 0.0 to 1.0 (100% packet loss). Blue bars of 100% indicate that the remote node was down at the time of the measurement. In that case, obviously no RTT's are reported. Packet-loss statistics can be used as a measure for node availability. Average round-trip time level (QNA) in a given month, in seconds.

Results

With the Dutch node as originating node, data have been collected since November 1994 and QNA numbers have been calculated for all EMBnet nodes since then. For verification, also data sets from Spain and Norway were collected for several months in 1995 and 1996. These daily data for each month are graphically accessible from a QNA request (hypertext link at the bottom of this document), in gnuplot graphs as shown in the following picture fragment:

To interpret these data correctly, one should be aware that RTT delays are due to both network congestion and destination occupancy. Since both are generally strongly dependent on the time of the day, RTT data show a time-off-the-day dependence as is exemplified in the picture below. This example shows measurements to a German destination from The Netherlands. These daily fluctuations are only visible as the difference between minimum and maximum RTT in the representation of the daily measurements and disappear completely after averaging to monthly QNA numbers.



Plot of round-trip times versus the time of the day

It must also be emphasised that the ping data collected in this project do not give an indication of specific bottlenecks along the data paths. We only try to collect and present factual data on the accessibility of the different EMBnet nodes, in order to determine if network problems hamper node services. By having performed these measurements over longer periods (some years), improvement or deterioration for specific nodes could be traced. For less global measurements of this kind, network and routing topology is required but that would result in a project far beyond the scope of the current one. However, to help track down some extreme delays for specific nodes, traceroute measurements have been performed on an incidental basis.

Conclusions

In 1995, a Network Usage and Quality Advisory Group of the Dutch Network organization SURFnet, defined "an upper RTT limit of 125 msec. without packet loss" as a minimum QoS level for interactive on-line work. For data transport only, a slightly less stringent criterium could be used. If one accepts this value of 125 msec. as an upper QNA limit for on-line work, the network performance results collected sofar in this project, are rather disappointing.

At this very moment, the majority of EMBnet Nodes for which data are being collected (15 out of 24 or 63%) show QNA numbers at or above this maximal acceptable value. Greece, Poland, Portugal and Israel perform very badly with QNA's of >500 msec. In Europe, only the Nordic countries, the UK, Holland and surprisingly Hungary, meet the QNA limit. An additional concern is that for about half of the EMBnet nodes the situation has not significantly changed or improved and in some cases even has deteriorated since the beginning of our measurements in 1994. For some countries a significant improvement is observed somewhere in the time path since 1995, but often immediately followed by a gradual deterioration. Packet loss statistics show that a few nodes have been unreachable for longer periods of time (at_biocenter, fr_inserm, gr_imbb, it_cnr). Reasons for this observation are unclear but are probably more of a local nature than that is has to do with global network capacity.

As a reader and/or EMBnet node manager you are invited to draw your own conclusions about the accessibility of your own node.

A plot of QNA vs. time for your node of choice will be generated "on-the fly" here. QNA numbers and daily data for each month are accessible here.

This document is intended to be permanently usable as a monitor for the control of network performance to or from any specific EMBnet node. Therefore we, the Dutch node, are ready to expand the number of nodes taking part in the measurements. A copy of the monitoring protocol is available on request. Data resulting from these measurements will be sent to the Dutch node where they will be added automatically to the files which can be queried from this publication. For further information and a local copy of the monitoring protocol, contact Koen Cuelenaere, CAOS/CAMM Center Nijmegen

Acknowledgement

We thank Hans Engelkamp and Wim Janssen of the CAOS/ CAMM Center for their help in the development of the monitoring utility and the data collection and the data processing code. The Network Performance Monitoring Project was partially funded by European Commission under grant ERBBIO4-CT96-0030.

An Interactive Bioinformatics Practical on the WWW

Life beyond GCG

T.K.Attwood
Department of Biochemistry and Molecular Biology
University College London, London WC1E 6BT, UK.

http://www.biochem.ucl.ac.uk/bsm/dbbrowser/jj/prefacefrm.html

Introduction

Bioinformatics is one of the fastest growing disciplines in the biological sciences, its growth fuelled by the quest to uncover the functional gems latent in the human genome. But the rate of emergence of the field has created a skills vacuum; postgraduate and mainstream undergraduate courses are still rare, and the pool of trained bioinformaticians from which to draw new recruits remains small. In this highly-advanced, computer-literate age, persuading biology students not to be afraid of computers is surprisingly difficult.

At UCL, we run a final year bioinformatics course, which includes an interactive Web-based practical. This is important for various reasons: many students are familiar with Web interfaces, so the practical is, in principle at least, easy to use; armed with the URL, study may continue away from formal sessions; and, by exploiting interactive Web technologies, practicals can be made visually stimulating, and sometimes even fun! With such approaches, we hope to demystify computers sufficiently to appeal to new generations of students, and encourage them to progress to those highly prized postgraduate qualifications.

Why teach bioinformatics?

In the context of genome initiatives, the term bioinformatics strictly applies to the computational manipulation of biological sequences; yet, increasingly, it is also used to embrace protein structures. However, it is instructive to bear in mind the difference in scale of handling sequence and structural data: today, in public, non-redundant data repositories: there are >250,000 protein sequences, but still <2000 structures. A central challenge of bioinformatics thus lies in the rationalisation of the flood of sequence information. In other words, if we are to derive the maximum benefit from the wealth of sequence data, we must deal with it in a concerted way: this means establishing and maintaining databases; designing powerful new analysis software; and providing tools to help interpret the results of those analyses in biologically meaningful ways. To achieve this, scientists skilled in relevant areas of both biology and computing are required now.

How best to teach bioinformatics?

Teaching computer applications to biologists can be remarkably difficult. In spite of today's technological revolution, many students (and their teachers) are still `silicophobic' - they may be used to dealing with systems in vitro or in vivo, but not yet in silico. Nevertheless, the emergence of the WWW, coupled with advances in Web technology, has provided interesting new opportunities for teaching. There are several advantages to Web-based approaches: `point & click' interfaces are easy to use, even by the computer-wary student; data from different sources can more easily be brought together, so teachers need not be expert in all areas, but may direct students to more appropriate expertise on the Internet; the use of interactive programs can render data more tangible and visually appealing; and, armed with a URL, study may continue away from formal classes, allowing students to work at their own pace, whether at home or on computers elsewhere on the College campus.

An interactive Web practical

Conventional teaching of bioinformatics has tended to rely on commercial analysis packages, e.g. GCG. Such programs can be difficult and tedious to use for all but the computer enthusiast, and are undoubtedly opaque and off-putting for the bench biologist. In an attempt to address some of these problems, we have made an interactive bioinformatics practical (BioActivity) available on the Web. BioActivity provides an interface to sequence and structure analysis resources around the world. In the frames version, each page includes: a short set of header instructions; a detailed commentary, giving the rationale for the relevant part of the practical (this is augmented with pictures, to help illustrate more explicitly what is meant in the text); and additional information and/or `Help', providing further explanations, background information, and so on. The practical includes some Java software, allowing students to interact directly with data downloaded from the Internet.

The practical is self-contained and falls into distinct sections: nucleic acid sequence analysis (translation and reading frame identification); protein sequence analysis (primary and secondary database searching); and protein structure analysis (querying the structure classification resources). A set of 30-base nucleic acid fragments is provided, from which the student selects one example to take through the investigation. BioActivity begins with translation of the fragment, and identification of the correct reading frame by rapid identity searching of the OWL database. Once the protein has been identified, and the full sequence extracted from the database, a search of the primary databases is made for known homologues using the BLAST search tool. The secondary `pattern' databases are then searched to discover if the protein contains any characteristic functional sites or motifs: this includes searches of PROSITE, the BLOCKS databases, the PRINTS fingerprint database, and Pfam. The final stage of the practical is to examine the structure classification resources, such as scop and CATH: the aim here is to relate functionally important motifs to the 3D fold and to try to understand their relevance in structural terms. References are given to the various databases and search tools used in the practical, and an extensive glossary of terms is also provided.

Conclusion

BioActivity is a versatile, interactive bioinformatics practical. By exploiting user-friendly, Web-based approaches, we hope to make bioinformatics a more accessible subject for biologists, and show that there is life beyond the traditional commercial analysis package!

Interview: Des Higgins reveals all about ClustalW

Andrew Lloyd interviews Des Higgins

AL-1: ClustalWWW was recently launched at the EBI as the latest incarnation of your original multiple sequence alignment program. I guess it's more than a decade since you first started working on the problem. Can you tell us how and why you came to do so ?

DH: There have been some WWW versions/incarnations of clustalw for a while. I have never seen most of them. I was getting some very confusing help requests from users that asked me why their sequences disappeared when they clicked "reload" and so forth. It took a while before it clicked with me. It has been a strange journey.

It started in Dublin in the mid 1980s when there were still no reliable multiple alignment methods available. This was in the days when "homology searches" had just been cracked by Lipman and Pearson and there were still plenty of relatively simple unsolved problems. I was sick of doing multiple alignment by hand (as were many other people). As your predecessor, I got endless requests from users for multiple alignments. One day I had a brain wave; why not do this using a phylogenetic tree as a guide for bigger and bigger alignments. I tried it out and it seemed to be fast and relatively accurate. Sadly, the idea had also occurred to Willie Taylor in Mill Hill (London) and Da-Fei Feng and Russ Doolittle in San Diego. They published papers in 1987 and 1988. Our lab (Paul Sharp's group in the Genetics Department, Trinity College, Dublin) was run on a couple of PCs with 8086 processors (one fast one had an 80286) and I had to spend ages shoe horning my program into 640kb of memory and make it fast enough for the PC.

This paid off as I now had a niche: multiple alignments on a PC. After a few hours in Kennedy's pub at the top of Westland Row, we even had a name for the program: clustal was born. It is not a great name but the only alternatives we could think of were rude.

It came in 4 programs which you had to clunkily run one after the next. In the meantime (mid 1988), I saw papers by Geoff Barton and Florence Corpet with their related programs. To my horror, Florence's program also worked on a PC and was more or less as fast as mine.

Clustal V came a few years later (the V did stand for 5) when I was at EMBL in Heidelberg and was based on a C version of older clustals by Alan Bleasby and was done in collaboration also with Rainer Fuchs, who tried to teach me some C.

It was a proper Unix and VAX/VMS program as well as PC and we had a clunky MAC version. It took off. GCG brought out PILEUP, using the same basic method at around the same time and I assumed that that was curtains for clustal but many people did not have access to GCG and we had some extra features like bootstraped NJ phylogenetic trees.

Clustal V stayed as it was with unimplimented improvements and unfixed bugs until I started collaborating with Toby Gibson (still at EMBL) and his amazing programmer: Julie Thompson. I had run out of energy/ability/ideas for Clustal V and Toby started playing with sequence weights and positions specific parameters. Poor Julie had to impliment all his wild hand waving but the result was a souped up high powered version which was called Clustal W (W for "weighted") in 1994.

Clustal W is updated about once a year and is now at version 1.7. It has become a complicated bag of tricks and I no longer know what half the code does (this was probably true from the start with the other clustals anyway). Julie is now in Strasbourg and the program is still under development.

Finally, there is Clustal X which is a beautiful X Windows version. This was done by Julie in collaboration with Francois Jeanmougin in Strasbourg. I say it is beautiful because apart from it being my suggestion (I think), I had very little to do with it and I think it looks fantastic. What comes next is that I have to write a course on introductory biochemistry for medical students here in Cork.

AL-2: Would it be fair to say that it is the most widely cited single program in bioinformatics ?

DH: Sadly, I doubt it! Blast (Altschul et al) will surely have an order of magnitude more citations. We do get lots though which is nice. It is a direct measure of usage that we can wave at people when we need jobs or money. I would be interested to know what the top papers are by the way if anyone is counting :-).

AL-3: Several years ago (round about the first Genes, Proteins Computers meeting in Chester I remember) there was some active gossip that clustal had been sold to Intelligenetics and that we'd all have to pay to use it in future. Do you ever regret keeping the program in the public domain and turning down all those royalty cheques ?

DH: Ironically we do now look for some royalties now to help pay for upkeep. We only ask people who look like they are making money from the program and it helps us keep it alive. In the past though, making it completely free was a stroke of luck. I know this from talking to some of the "competitors" who made their software too hard to get at. It helped clustal to become so popular. Basically, for a long time, I just asked people to cite the papers but otherwise do what they liked with the code.

AL-4: You've been in some controversial phylogenetic waters in your time - I think particularly of your "whales are artiodactyls" paper with Dan Graur. Do you think that molecular phylogenetics is in a position to tell traditional taxonomy to move over or even roll over and die ? Or do you feel that they are still complementary and both required?

DH: That is a leading question. Species are dying out faster than they can be stuffed into bottles of alcohol and stored, never mind classified or described so DNA still has a bit to go before it replaces that. For dating evolutionary events and trying to figure out difficult phylogenetic patterns, however, there has been a revolution. There are still many unsolved technical issues but these are exciting times if you are interested in trees.

AL-5: It is not widely known that you are amongst the foremost experts on the taxonomy of Irish spiders. Do you still work as a "real biologist" ?

DH: Blush. I used to collect spiders and I did discover some that were new to Ireland but that is a relatively easy thing to do when there is only one other person in the country doing it. I have no time now (because of children rather than work although we do collect snails and slugs in our vegetable patch).

AL-6: You moved back to Ireland recently after several years at such hubs of European bioinformatics as Heidelberg and Hinxton. Do you find Cork isolated or are we all in a global village now ?

DH: It is a bit of both. Cork is a small city in a small country (I still have not found a source for Italian Pancetta) and yes it can be quiet here. The Internet, however, has changed what we do almost beyond recognition. Yes it is a global village and I wish
the neighbours would not slam their car doors so often. It is a seething mass of activity which Cork is connected to as much as California.

AL-7: How important is it that undergraduates in the Biochemistry Department should be given formal training in bioinformatics as well as the Krebs Cycle ?

DH: Two possible answers:
1) What's the Krebs Cycle? or
2) It is not just important, it has become unavoidable.

They combine well anyway (the Krebs Cycle and bioinformatics) and it is cheap and safe to teach. Fortunately for me, the powers that be here had heard of bioinformatics and thought it was worth trying out. In some biochemistry departments, I would not have been so lucky.

AL-8 What do you see Des Higgins doing in the future ?

DH: I am still trying to buy a house. After that, the immediate plans are all about aligning ribosomal RNA sequences. After that, I have absolutely no idea :-).

Book Review:

Molecular Evolution

Wen-Hsiung Li
publ. Sinauer Assocs 1997.
ISBN- 0-87893-463-4 Hbk
32.95GBP. No softcover alternative.

Reviewed by Andrew Lloyd, EMBnet Ireland.

It is now more than five years since Li and Graur gave us "Fundamentals of Molecular Evolution". Li's latest book might be called "Molecular Evolution for Grown-Ups". It is altogether a weightier tome and pulls few punches about the mathematics and molecular biology that are necessary to deal with the subject properly. If, after two consecutive Σ's, your eyes glaze over and hunt wildly for the next block of prose, don't worry: there is enough of the latter to make it a perfectly readable book. The tables and illustrations are generally relevant and helpful. There is no glossary, but the index is good enough so that if you go to the first page referenced you should find the term both defined and in bold. On the other hand, authors are not generally indexed, but can be tracked down fairly quickly through sensible use of the subject index. As befits a book aimed at American graduate students the first few chapters end with a few problems, the answers to which are given at the end of the book.

So much for the structure; is the material any good ? Those who are familiar with Li and Graur 1991 will not be disappointed with it. It is obviously a bit biased: towards the works of Li and his co-workers for example. But as Li has made substantive contributions to the field this does not particularly obtrude. Being a single author work gives it a much desired cohesion and drive, compared to compilation volumes. Li claims to have been choosy rather than comprehensive but covers most of the topics that you would expect to find.

The chapter on molecular phylogeny could have been a little more prescriptive than a review of the literature. I was surprised, for example, to see Lake's invariants given the same billing as much more widely used and accepted methods as parsimony, neighbor-joining and maximum likelihood. The discussion of introns (early/late) could have had a bit more background than a referral to Genes V but is otherwise fair. There are excellent chapters on horizontal transfer and the evolution of genomes. However, the final chapter where Li deals with the selectionist- neutralist controversy requires more fleshing out to convince the other side. Earlier, Li unfortunately takes the selectionist explanation of mammalian isochores at its own estimation by accepting the concept of warm and cold blooded vertebrates. Let one of those selectionists take the rectal temperature of a lizard basking in the sun and see who's warm blooded. In other words, theoretical biologists sometimes seem to treat the living world as an array of cardboard tokens with a tiny number of attributes which require explanation. Thus, Li stoutly maintains (three times) that there was only one relevant environmental change in the evolution of colour vision in primates because he seems satisfied that "apples are red, leaves are green and some primates eat fruit" is a sufficient explanation. But, in fairness, one of Li's central chapters giving some real world case histories is as good an exposition of the value of molecular evolution studies as you're likely to find in one place.

I think that the shorter, cheaper and more accessible Li and Graur will still sell copies but that the present book is better value. Indeed it is a valuable contribution to the education of the next generation of evolutionary biologists.

Andrew Lloyd, Dublin September 1997

Node Focus: University College London joins EMBnet

T.K.Attwood

Department of Biochemistry and Molecular Biology, University College London, London WC1E 6BT, UK

In September, at the EMBnet Annual General Meeting in Bari, the Biomolecular Structure and Modelling (BSM) unit at University College London (UCL) was accepted as one of the new Specialist EMBnet Nodes. The UCL node provides access to a range of in-house databases and analysis tools via the DbBrowser bioinformatics Web server at:

http://www.biochem.ucl.ac.uk/bsm/dbbrowser/embnet.html

The BSM unit is a biocomputing centre with expertise in two central areas of bioinformatics, specifically in protein sequence and structure analysis. The unit currently hosts groups responsible for the curation of the PRINTS protein motif fingerprint database; the CATH protein structure classification resource; the PDBsum database (which provides summaries and structural analyses of PDB data files); the Enzyme Structure Database (which links the E.C. numbers to the structural data in the PDB); and the KabatMan antibody database.

In addition to the databases, the unit makes a variety of analysis tools available from its anonymous ftp site (ftp.biochem.ucl.ac.uk in /pub): these include LIGPLOT, a program to plot schematic diagrams of protein-ligand interactions; BPLUS, which calculates hydrogen- and non-bonded interactions; PROMOTIF, which analyses protein structural motifs; and PROCHECK, a suite of programs to check stereochemical quality of protein structures.

Other programs are available for use directly via the Web, including the fingerPRINTScan suite for searching the PRINTS database (either with individual query sequences or with complete genomes); the CINEMA Colour INteractive Editor for Multiple Alignments, which allows interactive manipulation of existing protein sequence alignments on the Internet and/or custom creation of alignments locally; the GPCR pattern-recognition resource, for rapid diagnosis of G-protein-coupled receptor sequences; the SAS package, for annotating protein sequence alignments with structural information; and many others.

DbBrowser also hosts the BioActivity interactive bioinformatics Web practical. This provides an interface both to our own in-house protein sequence and structure analysis facilities and to resources worldwide. BioActivity is a self-contained tutorial that leads the user step-by-step from unknown fragments of DNA, through translation and reading frame identification, to searches of the primary sequence and secondary `pattern' databases, and on to queries of the protein structure classification resources.

The practical offers an alternative approach to the use of commercial packages such as GCG, and provides a good basis for student and staff training. We have run a number of introductory bioinformatics courses, with BioActivity providing the core, at UCL and Birkbeck College; our hope is to run similar courses for EMBnet. UCL's computing system includes around 40 Silicon Graphics workstations (from personal Iris through to Octane), 3 four-processor Origin 200 servers and a 4 processor challenge L server. Of these, one R10K O2 is designated as "external services server" handling anonymous ftp and web requests, and one of the Origin 200 machines accepts requests for offline processing of more complex web queries, as well as being used to generate and maintain the databases created within the unit.

Node manager:
Terri Attwood - E-mail: attwood@biochemistry.ucl.ac.uk
Tel: +44 171 419 3879 Fax: +44 171 380 7193

Sysadmin:
Martin Jones - E-mail: mlj@biochemistry.ucl.ac.uk
Tel: +44 171 419 3896

Protein sequence enquiries:
Julian Selley - E-mail: selley@biochemistry.ucl.ac.uk
Tel: +44 171 419 3896

Protein structure enquiries:
Sue Jones - E-mail: sue@biochemistry.ucl.ac.uk
Tel: +44 171 419 3890

The UCL node is not making accounts available - this is the province of our National Node, SEQNET. Nevertheless, we will provide databases, related services and network tools, and hopefully training. Positioned within a unit that integrates protein sequence and structure groups, we can also offer expert advice on sequence and structure analysis. The main contact people are set out above. Please call us if you have any problems, or if we can be of help in any way.

The Informant

Back in the good old days when there was quality control on Bionet and Reinhard Doelz was at the helm of bionet.software.www there used to be a regular input of WWW sites that dealt with molecular biology. Nowadays where do you go to to find interesting sites. The classic lists can be found at Bio-wURLd at EBI and the WWW VL Biosciences Index at Harvard run by Keith Robison.

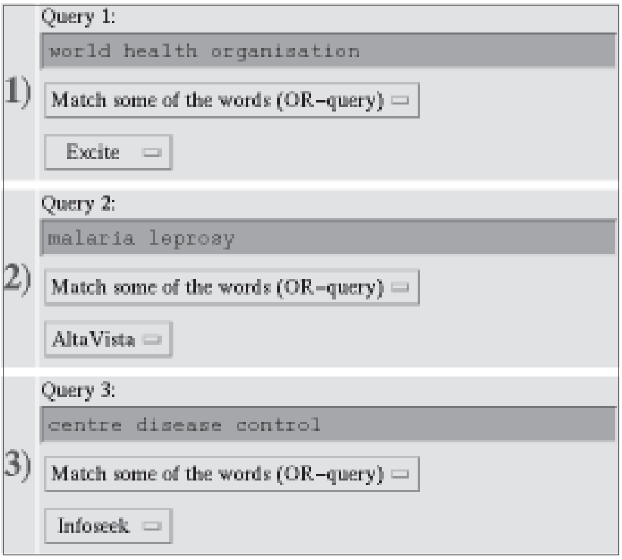
However a valuable new service can be found from the Informant. It is very simple to use.

Option 1

The first thing you do once you have registered is to fill in General Preferences, which is basically just your Email address and how often you want information sent to you.

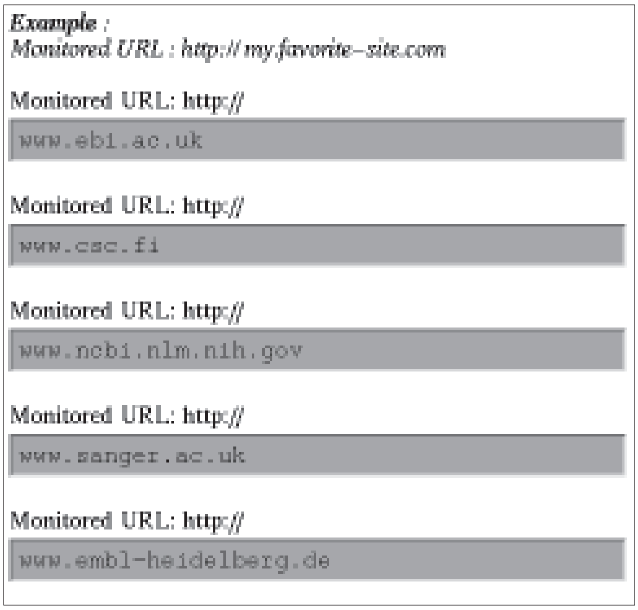
Option 2

The second form you have to fill in is Keyword Preferences. You are allowed to make three different queries and select from a variety of search engines like Alta Vista, Excite, InfoSeek and Lycos.



Option 3

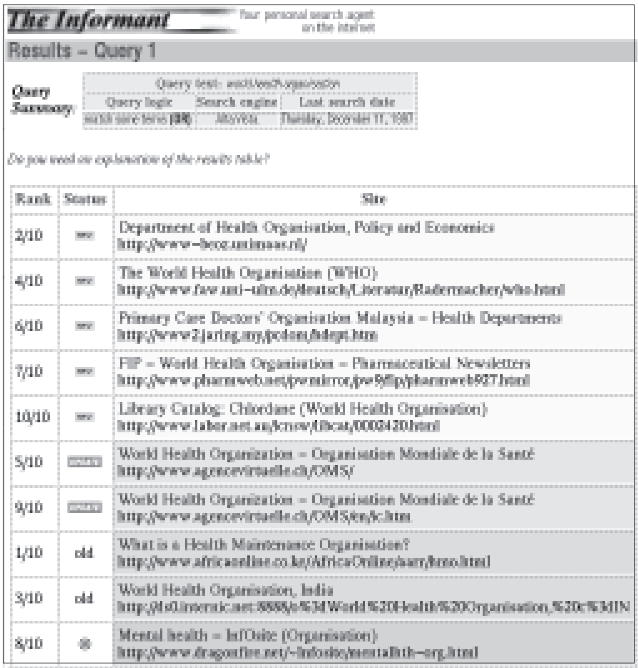
The third option is to keep a check on sites that you would like to monitor to see if they are updated.



Then all you need to do is sit back and wait for the results. You will be informed by Email once your keywords have been processed. and when you go to the Informant site you can get a general summary of the results of your search neatly laid out in a table.

By clicking on the Query1 button you can then obtain the complete results of your search. See next column for such a result.

That's it. If you would like to take Informant for a spin then just click on the icon


The Informant
Your personal search agent on the internet

Results – Query 1

Query Summary

Query logic | Search engine | Last search date

Do you need an explanation of the results table?

Rank	Status	Site
2/10	new	Department of Health Organisation, Policy and Economics http://www–beoz.unimaas.nl/
4/10	new	The World Health Organisation (WHO) http://www.faw.uni–ulm.de/deutsch/Literatur/Radermacher/who.html
6/10	new	Primary Care Doctors' Organisation Malaysia – Health Departments http://www2.jaring.my/pcdom/hdept.htm
7/10	new	FIP – World Health Organisation – Pharmaceutical Newsletters http://www.pharmweb.net/pwmirror/pw9/fip/pharmweb927.html
10/10	new	Library Catalog: Chlordane (World Health Organisation) http://www.labor.net.au/know/libcat/0002420.html
5/10		World Health Organization – Organisation Mondiale de la Santé http://www.agencevirtuelle.ch/OMS/
9/10		World Health Organization – Organisation Mondiale de la Santé http://www.agencevirtuelle.ch/OMS/en/ic.htm
1/10	old	What is a Health Maintenance Organisation? http://www.africaonline.co.ke/AfricaOnline/aarr/hmo.html
3/10	old	World Health Organisation, India http://ds0.internic.net:8888/o%3dWorld%20Health%20Organisation,%20c%3dIN
8/10		Mental health – InfOsite (Organisation) http://www.dragonfire.net/~Infosite/mentalhth–org.html

Node News

Spain

Staff

The Spanish node, EMBnet/CNB has expanded their human resources with the addition of a new person to the team: Sonia de Diego is a young Computer Scientist graduated in the University Autonoma of Madrid, who has joined the Service as a Systems Manager.

Bioinformatic Services

During this year, we have expanded our computing systems and services, increasing our storage capacity. Recently we have ordered an upgrade of our main server, a PowerChallenge computer from six (6) to ten (10) R10000 CPUs, plus 256 MB additional main memory, raising it to 1GB total memory. In this time, we have been able to add a panoply of new software tools and services, including tools for DNA fingerprint analysis, evolution, sequence analysis and molecular structure analysis, as well as new documents and manuals. All the new tools and services are described on our web server (http://www.es.embnet.org/Services/).

New Network services

Users can use a new dial-up service, providing six 33.6K modems and PPP access to our network. We expect to increase the number to 8 and then 16 phone lines in the next few months. We have installed three stratum 3 XNTP time servers integrated in the NTP project of RedIRIS, the Spanish Academic Network. Next we are adding cryptographic authentication to certify the time served. The FTP server is now also included in the archie indexing infrastructure, identified as ftp.cnb.uam.es -an alias of ftp.es.embnet.org- to facilitate localization of its contents. We now have Kerberos v4 and v5 servers for user authentication, and we are using them to increase security in our network. We have a Certification Authority able to issue SSL certificates which will allow us to certify our web servers and issue personal certificates for EMBnet/CNB users. This CA is going to be an integral part of the upcoming Certification Hierarchy of RedIRIS, the Spanish Academic Network.

GeneBee - Russian EMBnet Node

This year activity of the Russian Node team was directed to three main fields.

1). Free service of domestic and international users by the own sequence analysis programs through the Web and e-mail. We are also improving and developing this set of programs.

2). Establishment of the GCG package service for the Russian community. The package has been purchased and now it is in the process of installation.

3). The development of the object-oriented biocomputing server based on a relational database (to store databanks). We plan that for experienced users, this system will be accessible via a special client to make possible the inclusion of user-defined computational classes, and it will have a CORBA interface too.

The most popular program of our present service is MULTIPLE SEQUENCE ALIGNMENT program. Recently it has been improved and is considered to be one of the best at this field. This software will be the subject of a BITS piece in the forthcoming embnet.news issue.

ICGEB Italy

We have just completed the EMBnet course:
http://www.icgeb.trieste.it/net/courses/bioinfo97.html

Shortly before we moved from GCG version 8 to GCG 9; now we are waiting for the EGCG suite v. 9. We also added some html documentation on our WEB (for clustal and phylip programmes).

On the hardware front, we have installed a new backup system. We had problems with our UPS because of a thunderstorm, and it is still under repair. Now we are waiting for the next lightning.

ICGEB bioinformatics courses 1998:
http://www.icgeb.trieste.it/net/netcourse.html

Structure of Biological Macromolecules March 16-27, 1998
http://www.icgeb.trieste.it/net/courses/miramare.html

Structural Biology and Functional Genomics , May 4-8, 1998
http://www.icgeb.trieste.it/net/courses/nato.html

"Bioinformatics: Computer Methods in Molecular Biology", July 3-10, 1998 http://www.icgeb.trieste.it/net/courses/bioinfo98.html

China

Node Manager Jingchu LUO. Email:
luojc@lsc.pku.edu.cn,

WWW Site: http://www.cbi.pku.edu.cn/

It has been a year since the National Laboratory of Protein Engineering and Plant Genetic Engineering at Peking University was accepted as a national node of EMBNet in November 1996. We started our web server early this year with the EMBL and SwissProt Databases installed under SRS. The hardware was a SUN 1000e with 8GB of disc space. It was upgraded to an SGI Origin 200 in September. A better service has been provided since then by this two CPU (R10000) machine. Disc space was expanded to 31GB in November.

Mirrors of UNIDO Biosafety regulation, EMBNet Biocomputing Tutorial, GBF Transfac database, UK MRC HGMP Genome Information and the EMBNet newsletter have been installed. Databases of protein loop classification and protein domain assignment developed collaboratively with colleagues at Imperial Cancer Research Fund were also set up at our site. In addition to the sequence and sequence related databases some protein structure, genome, mapping, mutation and other useful databases have also been put onto the server. All these databases are also available via SRS.

A mirror of our web server has been set up at three local institutions where a full network connection is under construction. This service interests more and more potential users. A 9GB disk was purchased to server as a "mother disk". The essential databases were downloaded together with SRS and an apache httpd server. Copies of this disc are easy to install on the local systems at other sites.

An introduction to bioinformatics, EMBnet and CBI was given in the annual meeting of the national High-tech programme of biology; this is a symposium of industrial genome work and bioinformatics, the first symposium of young Chinese geneticists. Talks and demos about the service we are providing were given to lots of users.

The development of a new algorithm for database mining has just started collaboratively with mathematicians of our university. Projects for protein and peptide structure determination by NMR, protein and peptide design and prediction plus DNA and protein sequences analysis are being carried on in our lab.

An EMBnet advanced course is planned for next April. A WhatIf workshop is to be held next June along with Gert Vriend.

CAOS/CAMM Center

The CAOS/CAMM Center recently acquired a new Bioccelerator model XL/G, containing 24 processors, and 512 Mb RAM. This machine will take over most database searching tasks, that still run on the Center's main server, a 10-processor SGI Challenge. Currently only hosting the ever-so-popular FASTA suite, in time the XL/G will also be able to run the BLAST programs and Smith & Waterman searches (e.g. profile- and framesearching). The latter type of database applications is performed by the "old" Bioccelerator, which will continue to be used for this task in the future.

ASCII interface for SRS 5.0.5

The ASCII interface to SRS (Schaftenaar, 1996) has been adapted in such a way that it now can be used with SRS 5 (release 5.0.5). Only the "getz" calls have been changed, so the internal (using the SRS API calls) and the Hassle versions can no longer be used. The Xwindows version is currently unavailable, but will most probably be finalised after the introduction of SRS 5.1.

This version is has only been tested on SGI's Irix 5.3 and 6.2, using SRS version 5.0.5. It will probably run on any other UNIX OS that is capable of running the previous ASCII interface. Unless the "getz" flags do change dramatically in SRS v5.1, it will also work with the new release.

The ASCII interface of SRS 4 was written by Gijs Schaftenaar. The current port to SRS 5 were made by Koen Cuelenaere. If you would like to install a copy of the new version, please contact post@caos.kun.nl

The EMBnet Nodes

National nodes:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Special nodes:

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

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

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

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

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

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

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

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

Associate nodes:

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

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

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

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

[IN] CDFD India

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

[AR] IBBM Argentina

[CB] CGEB Cuba

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/vol4_3/contents.html***
- ***http://www.be.embnet.org/embnet.news/vol4_3/contents.html***
- ***http://www.no.embnet.org/embnet.news/vol4_3/contents.html***
- ***http://www.ie.embnet.org/embnet.news/vol4_3/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.no.embnet.org 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.no.embnet.org in the directory pub/embnet.news/***
- ***ftp.ie.embnet.org in the directory pub/embnet.news/***

Back issues are available at most of these sites.

Publisher:

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

Editorial Board:

embnet.news

Vol.4, No.3, 1997
December 25, 1997
ISSN 1023-4144