embnet.news

Volume 5 Nr3 (ISSN1023-4144) September 30, 1998

Editorial

EMBnet celebrated its tenth anniversary at its AGM in Hinxton two weeks ago. Those who were at the founding meeting in Heidelberg in 1988 doubtless expected that the organisation would last at least a decade and I am sure that they anticipated some growth. What is perhaps surprising is the scale of that growth which, in some respects, has tracked the growth of the sequence databases that are still at the core of the service EMBnet provides.

Counting the number of nodes has this year been a three steps upwards one step down process, bringing the current total to 37. EMBnet has welcomed new National Nodes from Turkey and Canada and a Specialist Node in the Expert Centre for Taxonomic Identification (ETI) from Amsterdam. The merger of the Swiss National Node with the SwissProt Specialist Node in the new SIB should certainly not be seen as a retrogressive step even if the EMBnet Node count is effectively one less.

Formalising the relationships within the multi- institute Swiss Node will surely make for better research collaborations, more efficient use of resources and a better service provision. EMBnet has now burst beyond its European bounds and 20% of the Nodes come from (all) other parts of the world. We'd like to hope that a bigger group is a richer group, with more opportunities for collaborations, more sharing of assets, and more interchange of ideas. In many ways, notwithstanding its increasingly global stance, EMBnet is the epitome of what an EU Concerted Action grant hoped to achieve.

The EMBnet CA grant now has a year to run. It is important that each month of that year is actively used by everyone in EMBnet, from the Executive Board who must seek out new sources of future funding, through the Project Committees which must more pro-actively fulfil their obligations, to individual node managers who should use this year to forge new relationships, run more EMBnet-sponsored training courses and promote EMBnet both nationally and internationally. There is no doubt that the next ten years will see significant changes both in the size of EMBnet and the nature of its activities.

Bioinformatics is now big business, employing thousands of people world-wide. Nevertheless demand far exceeds the supply of adequately trained computational biologists in both the academic and commercial sectors. EMBnet should no longer concentrate on database provision - outsiders are surprised, in this age of the internet, that so many copies of large and identical databases are maintained across Europe. Of far more immediate importance is the provision of bioinformaticians who know what to do with the data.

Training and education is almost certainly the most valuable resource that EMBnet nodes provide. Building on this existing strength, by developing more teaching resources and running more courses, will allow EMBnet to further promote itself as an absolutely essential cog in the enormous machine that is world bioinformatics today.

The embnet.news editorial board

EMBnet Annual General Meeting

A report on the 12th Business Meeting

Andrew Lloyd, Irish EMBnet Node

In many ways it was appropriate to hold the 10th Anniversary meeting at the Hinxton Genome Campus. The resources clustered in such an attractive English parkland setting are a monument to the changes that have occurred in bioinformatics in the last decade.

Three of EMBnet's best resourced Specialist Nodes are located in Hinxton and they got together to jointly organise the AGM and the associated historical and scientific meeting. With the addition of nodes from Turkey, Canada and the ETI in the Netherlands, the EMBnet constituency now consists of 37 nodes. This year we agreed to discard the distinction between European National Nodes and those from elsewhere in the world. There are therefore 28 National Nodes, 2 Industrial and, with the merger of SwissProt into the Swiss National Node, 7 Specialist Nodes.

Two guest speakers were invited to give a presentation to the Board. Chris Rawlings, vice-president of the International Society for Computational Biology (ISCB), made the case that, as the only international society for bioinformatics, ISCB and EMBnet had a lot of interests in common. He suggested that an EMBnet Board member should serve, ex-officio, on the ISCB Board.

Graham Cameron from the EBI was also anxious to establish the shared interests and mutualism of the EBI and EMBnet. While ISCB can offer EMBnet a chair on the Board, the EBI is putting a whole office at the disposal of EMBnet. The EBI would like to see how EMBnet's expertise in training and education can complement the EBI's role in database development and provision.

From the other side of the world JingChu Luo presented a report from APBionet, a confederation of bioinformatics support nodes from Asia and the Pacific rim. APBionet is a smaller and much looser organisation than EMBnet which believes that it has much to learn from us. Already we have three nodes in common - Angis in Australia, CBR-RBC in Canada and CBI in China - and, disconcertingly, in two countries (Russia and India) the sister organisations are represented by different institutions.

A number of changes have occurred among the officers of EMBnet as a result of elections at this meeting. Sandor Pongor has stepped down from the Executive Board and been replaced by Robert Herzog. The Chairmanship has rotated from Peter Rice to Andrew Lloyd, who hands the Secretary's portfolio to Peter Rice, and the stable core is provided by Jack Leunissen who continues to look after the Treasury. There have also been six changes in the composition of the Project Committees.

Concern about the future is inevitable as we enter the last year of the current EU Concerted Action grant. While the Stichting Finances are buoyant and have a significant income from node subscriptions, EMBnet still requires outside funds if it is to achieve anything beyond mere existence. The Executive Board will have its work cut out to identify, apply for and secure funding for EMBnet projects and initiatives that will mobilise the expertise, both latent and expressed, within the organisation.

Hopefully, at the next AGM in Brussels next year the future of European bioinformatics will be secure.

EMBnet at the Wellcome Trust Genome Campus

EBI, Hinxton, 18-9-1998

R. Herzog, Belgian EMBnet Node

On the occasion of the EMBnet AGM meeting being held this year at the Wellcome Trust Genome Campus in Hinxton, Southern England, representatives of the three important institutions who share this site presented introductory talks. Graham Cameron spoke on behalf of the EBI, Martin Bishop presented the HGMP-RC and Richard Durbin reported on the achievements of the Sanger Centre.

Graham Cameron, head of the European Bioinformatics Institute, has been with the EMBL data library since its creation in the early 80's. He was in the privileged position to talk about what he called the `prehistory' and the history of EMBnet which was celebrating its 10 years of existence. Graham witnessed the complete evolution of the EMBL databank, from the initial few hundred entries to the present multimillion sequences databank. He reminded us that EMBnet was the idea of a few people like Ashman, Kennard and Cantley at the EC in the early eighties.

These ideas brought Berendsen to present the concept of a `bioinformatics workstation', part of a network of peers spread around Europe. These `workstations' were to be based in universities significantly involved in areas like DNA sequencing, protein modelling and generally speaking in computer handling of biologically relevant data. A sketch map of Europe displayed partner nodes in Switzerland, Germany, Holland, Sweden, the U.K, France, Italy and Belgium.

In 1988 Greg Hamm and Roy Omond developed the Berendsen idea and the EMBL council asked member countries to identify their national representatives. In those times it was even decided what hardware a partner of the system should have: one VAX/VMS computer with at least 8 Mbyte of memory, sufficient disc storage to hold the current version of the sequence databanks, at least one tape backup unit and some high speed communication link. A VAX/VMS licence for at least 16 users was also considered as a minimum.

The three first nodes to join their efforts into EMBnet were CITY2 in France, the CAOS/CAMM centre in Holland and SEQNET in the United Kingdom. EMBnet was born! Very soon, nodes designated by several other European countries joined. At the first workshop, held in Heidelberg in 1989, EMBnet already consisted of 14 members. Daily updates of the databanks were underway, and some limitations of the DECNET data exchange software started to turn people's attention to the TCP/IP protocol. In 1990 the first grant for EMBnet was obtained from the EC; it was managed by the Italian node under the leadership of Cecilia Saccone. Some years later, Jack Franklin became manager of the EC grant and, in view of the still raising number of members, the need was felt for the creation of an international organisation with a legal status that would co-ordinate the work between the nodes and the funding agencies. The EMBnet Stichting came to life during the end of 1993.

Members of the EMBnet produced, in close collaboration, several important pieces of software, among which were packages now consigned to the dusty recesses of the history of bioinformatics; like Hassle (R. Doelz, Switzerland), NDT (Peter Gad, Sweden) and EGCG (Peter Rice, Sanger Centre). These software packages played decisive roles in the progress of the technology inside the organisation. On the other hand, much of the hardware that was used in those early days is now part of local technology museums, but Graham mentioned that undisputedly the most enduring piece of software inside EMBnet is Alan Bleasby's leather jacket!
These days the EMBL databank is no longer managed within the EMBL laboratories in Heidelberg, but rather in the purpose-built EBI institute, suitably vested close to Europe's largest genome sequencing institute, the Sanger Centre. In 1997, the growth of the nucleotide databank reached the level of a new entry every minute! Whereas in the initial release of the databank a lot of organisms where equally represented (by a single entry), the composition of today's bank shows a clear advantage for human sequences, good for about 42% of the total. This is followed by the mouse (9.8%), the nematode (8.3%) and baker's yeast (3.6%). Close collaboration between the three major nucleotide sequence producers (EMBL, GENBANK and the DDBJ) provides virtually equivalent data sets, the growth rate of which shows no sign of slowing down.

Graham concluded by saying that these days a lot of technical problems have been solved. EMBnet spans Europe and many areas of the world with a dense network of bioinformatics centres, managed by a community of committed professionals, making this organisation unique in the world.

Martin Bishop, head of the Human Genome Mapping Project Resource Centre (HGMP-RC), explained how this institute fulfils the needs of the international scientific community of genome researchers. Martin explained how the classical taxonomy, based on morphological similarities, had to yield in face of the molecular taxonomy, based on similarities between molecular structures. Around 1965, this science made its first steps with the study of the haemoglobin of several species. Later, it became possible to start comparing chromosome maps of various organisms, revealing that a lot of the basic genome structure is in fact highly conserved among e.g. the mammals. Martin reminded us that Gary Williams of the HGMP-RC initiated the port of the GCG package to the UNIX world. Today, HGMP-RC is a reference place in Europe where most data about genomic organisation is studied and made available for public access.

He was followed by Richard Durbin, from the Sanger Centre, who described the current progress in the sequencing of the many complete genomes the institute has undertaken. Currently, the human genome is progressing at a steady rate and the prospects are that it should be finished by 2003. The Sanger Centre is also making fast progress in the sequencing of the Caenorhabditis elegans genome which should be completed by the end of this year. Many genomes of important pathogens are reaching completion virtually every month. Richard gave a brief overview of the numerous efforts accomplished at the Sanger Center to develop ad-hoc software for genome analysis. It spans the whole range of functionalities, from assembly of large sequencing projects to comparative functional analysis, fingerprint contig mapping, managing of genomic maps (Richard Durbin is one of the authors of the famous AceDB software) and the EMBOSS project for advanced sequence analysis, under theleadership of Peter Rice, member of the EMBnet board.

To the outside world, the Wellcome Trust Centre Genome Campus appears as the Mecca of European Genomics and Bioinformatics. It clearly deserves this envied reputation.

INTERviewNET

This month Alan Bleasby interviews David Kristofferson, about the past, present and future of BIOSCI/Bionet

AJB: Hi Dave. I'll start by asking why you perceived a need for BIOSCI in the first place?

DK: When I first started working at IntelliGenetics in August 1986 and saw the electronic bulletin board system that ran only on the BIONET National

Computer Resource's DEC 2065 computer, it was immediately obvious that this could be a very powerful communication tool if enough people were involved in its use. It would enable scientists to conduct continuously the kind of discussions that normally only occurred during professional society meetings. The system also held the potential to avoid the costly "reinventing of the wheel" that occurs in so much research.

The published literature only records the successes of the scientific enterprise. Newsgroups allow scientists collectively to discuss unpublished "failures" and debug experimental protocols. This constitutes the bulk of scientific work and is not being served by the journals. A newsgroup system can accelerate the pace of research tremendously. The opportunity to have such a broad impact on molecular biology initially, and other fields of biology subsequently, was immensely exciting to pursue.

AJB: How and why did BIONET come into existence?

DK: BIONET was a central molecular biology research computing center funded at IntelliGenetics (IG) from 1984-1989 by the Division of Research Resources at NIH. The original grant application was co-authored by Larry Kedes, Doug Brutlag, Peter Friedland and Ed Feigenbaum of Stanford who co-founded IG. It was created to provide scientists with an inexpensive facility for accessing a comprehensive set of sequence analysis tools that could analyse a unified set of sequence databases in a common format. Part of BIONET's mission was also to facilitate electronic communications. I became the manager of the BIONET Resource in December of 1986 and supervised all aspects of its mission. The electronic communications aspect of BIONET was my own "pet project" to work on in addition to my other managerial duties.

At that time (pre-Arpanet connection) BIONET had around 700 labs worldwide using modems and X.25 networks to connect to the system and charged $400 a year for unlimited access. I started trying to promote the system by giving free accounts to people who would serve as discussion leaders and help organise online communities. If I recall correctly there were less than 20 "BBOARDS" as they were called on the BIONET system in '86. By the time the project was transferred to its current home at Stanford in 1997, there were almost 105 forums!

When BIONET in the U.S. was connected to the Arpanet in early 1987 and began collaborating initially with Michael Ashburner and Martin Bishop at Cambridge UK, the system became accessible to a broader audience. Other sites joined later, including the BIOTECH group at the University of Maryland under Deba Patnaik, the University of Uppsala (Mats Sundvall), the supercomputer center in Helsinki (Rob Harper) and IRLEARN in Ireland (Niall O'Reilly). The SEQNET facility became the responsibility of the Daresbury laboratory in the UK and the newsgroup work there was initially run by Royd Whittington and subsequently by Alan Bleasby. The extension of the service beyond the BIONET system in the U.S. gave rise to the "BIOSCI" name for the overall collaboration. However, the USENET system retained the "bionet" name reflecting its origin at BIONET and the difficulty of renaming something which had already propagated around the network. The collaborative effort became a bit too complex (too many mailing list sites to co-ordinate) and, following a meeting between several of the principals at the EMBL, it was finally decided to centralise all of the work between the Daresbury site in the UK (national EMBnet node) under Alan Bleasby and the IntelliGenetics site in the U.S. under myself.

AJB: What does BIONET do?

DK: BIOSCI continues to this day its mission of facilitating discussions and information exchange among biologists. I believe that it remains the largest unified collection of scientific discussion groups on the Internet.

AJB: How do you see the future of internet discussion groups and email lists?

DK: After 10 years of administering the U.S. side of the BIOSCI system both as part of my career and as a labour of love in my "spare" time, the torturous turns of my career and increasing work demands of the service were making it too difficult for me to assure BIOSCI's safe continuance without exacting a very high personal price. I thought it was therefore in the best interest of both the service and myself if I found the project a more stable home. The Stanford University Library, in particular the group that runs HighWire Press, appeared to be such a site and I arranged for the transfer of the project from my employer at that time to Stanford. It was my expectation that new resources would be devoted to the project to enhance the service, possibly link it to HighWire's electronic journals and combat the increasing tide of "spam" that was affecting the Internet.

I am very pessimistic about the future of *unmoderated* Internet discussion groups because of the tidal wave of "yahoos" that have swept over the net since the days in which it was primarily a research network. However, before I left the project, about a third of the BIOSCI newsgroups were moderated and that number has continued to increase. The future of these moderated groups can be very bright. Basically BIOSCI's future remains whatever the organisers want to make of it! The original motivations for BIONET's foundation described above will always remain fundamental needs of science, and BIOSCI/bionet can continue to fulfill those needs if it continues to adapt to meet new challenges.

BITS

Getting information out of BLAST and FASTA runs

by Guy Bottu, Belgian EMBnet Node

One of the things that users of bioinformatics services do most is to search protein or nucleic acid sequence against a sequence databank for similarities using the popular BLAST or FASTA algorithms. In the early days, when there were not so many sequences in the databanks, a search frequently yielded nothing. Nowadays, because of the sheer size of today's databanks, the problem is often that a search yields too many hits. Not only is it tedious to read through a "kilometric" output file, there is even a serious risk of missing interesting information. Indeed, an potentially interesting sequence could appear in the list of hits preceded by hundreds or thousands of other sequences and not be reported. This is because the software limits the number of sequences in the output or, what leads to the same result, not be seen because people have a tendency to look only at the top of the list. Here we review a number of procedures and tools that have been designed to tackle this problem.

Before you search : remove uninteresting parts from the sequence Proteins are typically composed of several domains. If your query protein shares a domain with hundreds of proteins in the databank and also shares a (shorter and therefore lower scoring) domain with only a few other sequences, you are likely to miss the second domain.

Similarly, DNA sequences often contain repetitive elements such as Alu that can produce a long but uninteresting list of hits at the top of the output. Therefore, if you find in the alignments at the top of a BLAST or FASTA output same rregion of your query sequence over and over again, it is a good idea to cut that range out (or replace it by XXX) by using a sequence editor and then run the search again

At the NCBI some software tools have been developed that automatically remove uninteresting parts of a protein : XNU (Claverie and States, 1993) removes short-range repeats like the poly-GXX of collagen tails and SEG (Wootton and Federhen, 1996) removes ranges with a very biased amino acid composition such as histidine-rich domains. The SEG and XNU algorithms turned out to be inappropriate for nucleic acids. Therefore, the NCBI has more recently developed DUST (Tatusov and Lipman, in preparation). The BLAST programs have a command-line option for submitting the query sequence to a "filter" program before starting the search. BLAST WWW or E-mail servers usually by default allow you to "filter" your query sequence. FASTA has no provision for "calling" a "filter", but there is still a possibility to submit the query sequence to the filter program by a separate command. Note that "xnu" and "seg" are available as stand-alone programs in the GCG package from release 9.0 onwards.

Help for browsing through long outputs

WWW BLAST and FASTA servers generally return their output in the form of hypertext. So you can easily shift back and forth between the list of hits and the actual alignments, as well as retrieve the sequence from the server with a simple "click" of the mouse. Among others, WWW2GCG, the web interface for GCG developed at the BEN site, returns "blast", "fasta" and "tfasta" outputs as easily parsable hypertext. WWW2GCG produces the output in a file in the user's account on the remote computer (contrary to conventional WWW servers, including the SeqWeb interface from GCG, which has no memory of previous sessions) and returns at the same time a hypertext version of the output to the computer of the user. The hypertext is generated by BOV (Blast Output Viewer), a perl script developed by P. Alard from the BEN staff. BOV inserts sequence retrieval hyperlinks pointing to an SRS WWW server .. Note that a user can, whenever he likes, re-read his old BLAST/FASTA outputs, either as hypertext using the "Directory" page of WWW2GCG or as simple text using the standard UNIX tools after logging into an interactive session.

Users who have no Internet/WWW access but have only a text mode login to a remote computer are not entirely left in the shadows. P. Rioux at the U. of Montreal has developed "tbob", a perl script that allows you to shift back and forth through a BLAST output with simple keystrokes.

Automatic parsing of outputs

There is also software that automatically parses BLAST/FASTA outputs and selects hits based on objective criteria. Sander and Schneider (1991) from the EMBL proposed a criterion based on % identical amino acids and alignment length that allows sorting of gapped protein alignments.

This means that only proteins sharing at least 70% secondary structure identity are retained. This criterion is used by the popular PredictProtein server to select from the SwissProt databank those sequences related to the query sequence. It is also used for making the HSSP databank. It has been implemented in the egcg program "fastacheck" which parses fasta outputs. The pair of programs fasta-fastacheck answers the question "Which proteins in the databank are so similar in primary structure to my protein that they are likely to have a similar secondary/tertiary structure?" It is therefore most useful for searching a sequence against the sequences of the Brookhaven PDB to see if there is a protein with known structure that can serve as a guide for modelling the structure of the query protein. Unfortunately, as we all know, egcg is dying...

The MSPcrunch software has been developed by E. Sonnhammer from the Sanger Centre (Sonnhammer and Durbin, 1994). It parses BLAST outputs and selects HSP's (BLAST parlance for local ungapped alignments). It first limits (to 10) the number of HSP's covering the same range of the query sequence. It then corrects the score of the alignments for biased base/amino acid composition. Finally it selects HSP's by a criterion based on score and adjacency (= correct order and distance constraints) to other HSP's from the same databank sequence. The pair of programs BLAST-MSPcrunch answers the question "Which sequences in the databank are likely to belong to the same family of sequences as my sequence ?" Unfortunately, MSPcrunch works well with the old BLAST and with WUBLAST but not with BLAST2 from the NCBI for other options than blastp (search protein against protein databank). This is because, for unexplainable reasons, information about strand/reading frame has vanished from the output.

Note that if you want to find as many possible sequences as you can of the family, rather than just a few representatives, you should relax the coverage constraint of 10. It can also be advantageous to relax the limit that BLAST sets on the number of reported alignments, leaving it up to MSPcrunch to reduce the size of the list.

Software Availability

The following software can be obtained by anonymous ftp :

XNU at ftp://ncbi.nlm.nih.gov/pub/jmc/xnu

SEG at ftp://ncbi.nlm.nih.gov/pub/seg/seg

DUST as an integral part of the NCBI toolkit at
ftp://ncbi.nlm.nih.gov/toolbox/ncbi_tools

BOV as part of WWW2GCG at ftp://alize.ulb.ac.be/pub/www2gcg tbob at ftp://megasun.bch.umontreal.ca/pub/tbob

MSPcrunch at ftp://ncbi.nlm.nih.gov/pub/esr/MSPcrunch+Blixem

The *PredictProtein* server can be accessed by WWW at http://www.embl-heidelberg.de/predictprotein/predictprotein.html and by E-mail at PredictProtein@EMBL-Heidelberg.DE.

References

Claverie, J.-M. and States, D.J. (1993). Information enhancement methods for large scale sequence analysis. Computers Chem. 17(2): 191-201.

Wootton, J.C. and Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. Methods in Enzymology 266: 554-571.

Sander, C. and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9(1):56-68

Sonnhammer, E.L.L. and Durbin R. (1994). An expert system for processing sequence homology data. In : ISMB-94; Proceedings Second International Conference on Intelligent Systems for Molecular Biology,

Altman R., Brutlag D., Karp P., Lathrop R., Searls D., Eds., pp363-368, AAAI Press, Menlo Park

TIPS

Find, Grep & Xargs:
Different ways to crack an egg

by Alan Bleasby, SEQNET.

A previous TIPS column introduced you to simple invocations of the 'find' and 'grep' commands. For example, the command:

```
grep Fred *
```

will search all files (*) in the current directory for the pattern (or searchstring) "Fred". If you wanted the search to be case insensitive you would have typed:

```
grep -i Fred *
```

This is all well and good however, you may well wish to search through a directory tree structure for a pattern. There are many ways to do this in UNIX depending on what level of complexity you require. The most simple way is arguably by using a combination of grep and find by using a command of the form:

```
grep pattern `find . -name \*.txt -print`
```

which will look for your pattern in all your `.txt' files in the current and sub-directories. This example uses back-quoting; make sure you don't use the normal quote character. Another point to note is that the asterisk has to be `escaped' with a backslash to prevent expansion (or globbing). An alternative way of typing this would be:

```
grep pattern `find . -name "*.txt" -print`
```

We can crack this egg another way and the following command uses the exec (execution) facility of the find command to invoke grep:

```
find . -type f -name \*.txt -exec grep pattern
{} \;
```

This is the form of command most wizards use. Grep is invoked with all the things (i.e. {}) that find locates. The command has to be terminated using a semicolon which itself has to be escaped to avoid mistranslation. A little wrinkle has been added in using "-type f" which restricts the search to regular files i.e. it will not find symbolic links (if you wish to do this add the flag -follow).

Now, there are two potential problems here. The first one is that grep will only report the filename and the search string if it has to search through more than one file. If the find only returns one file to grep then only the search parameter will be reported. To get around this problem you can type:

```
find . -type f -name \*.txt -exec grep pattern
{} /dev/null \;
```

This means grep all the files found and also /dev/null which, after all, is a zero length file.

The second problem occurs if the find command generates a lot of hits for *.txt . Both forms of command expand either *.txt in the case of back-quoting or {} in the case of exec'ing to a list of filenames. This list would then overflow the input buffer. Thankfully most UNIXs have the `xargs' command. What this does is to split up such lists of filenames or text into manageable chunks and then invoke another command (e.g. grep) as many times as necessary. We've used back-quoting and exec'ing up to now; to use xargs the pipe command (|) is handy:

```
find . -name \*.txt -print | xargs grep pattern
```

Note that we revert to using '-print' as we're no longer using the exec facility. The find command gets all the matching filenames and passes (pipes) them to the xargs command which splits up the filenames, if necessary, and hands them to grep to look for the pattern.

If you're lucky enough to be using the GNU versions of find and grep then you can add a neat trick. Some people accidentally create files with atrocious names (generally by accident). An example would be files having a newline (\n) in their name. The following command stops any havok associated with finding such files by separating the tokens (in this case the filenames found) with null characters and not by whitespace:

```
find . -name \*.txt -print0 | xargs -0 grep
pattern --
```

A further boon is the "--" syntax which also allows filenames like "-a" or "aaa;bbb\ccc" to be handled.

A caveat to all the above commands is that you're usually searching text files and certainly don't want to search binary files. An easy way around this is through use of the `file' and `cut' commands e.g.:

```
find . -name \*.txt -type f -print|xargs
file|grep -i text|cut -fl -d:  | xargs grep
pattern
```

If the file command reports the word "text" (as found by grep) then the file is not an executable so it passes things on via the cut command (for tidying up the string) to xargs and grep.

So, you can make things as simple or as complex as you like. There is no one correct solution and usually something simple will be adequate. If you wish to be specific about what you wish to do then the flexibility is there in UNIX.

Xargs also has many other useful facilities such as echoing or invoking commands once per line etc. A look at the man page for this command is always worth it if you deal with long command lines on a regular basis.

Collaborative Molecular Visualization in ToolSpace

T. Goddard, V. S. Sunderam
Emory University
{goddard,vss}@mathcs.emory.edu

Abstract

ToolSpace allows interactive molecular visualization "conference calls" for people and molecules located anywhere on the internet. Simply by pasting in the URLs, any number of molecules can be imported together for measurement and visual docking. A group of participants can then discuss the structures using conventional and three dimensional conferencing tools.

Introduction

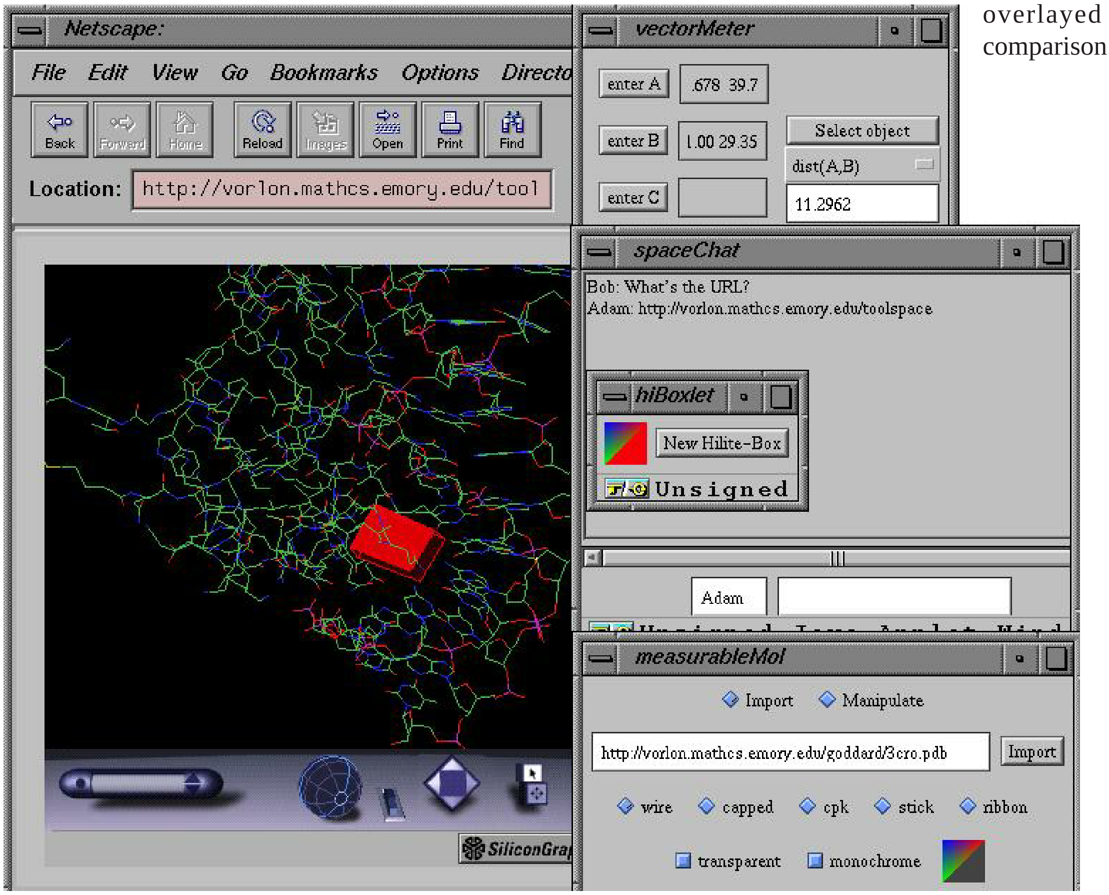
Three researchers on separate continents (having overcome time zone differences) use their VRML-enabled web browsers to meet in ToolSpace by visiting the ToolSpace web site that one of them has set up. They've placed the molecules they want to discuss as .pdb files on their own web sites and are talking over a telephone conference call or internet phone (such as the CCF audio tool). For tools, they've selected the molecule importer, measurer, highlight box, and text chat. (Picture 1)

After importing some structures (by pasting in the URLs to the .pdb files), they can collaborate by taking measurements and pointing out the features they wish to talk about. Each of them has their own point of view in the space as seen in their web browser, but the positions and orientations of the objects in the space are shared (this is analogous to a group of people discussing a collection of objects together in a room). To share a point of view, they place "viewpoints" in the space that anyone can warp to simply by selecting from a list. When they're done, they archive the session by taking a "snapshot". The snapshot can be viewed later in three dimensions just like the original space (but the objects within are no longer interactive).

Tools

Molecular Modeling Tools

ToolSpace presently has three molecular modeling tools: a molecule importer, a molecule measurer, and a Van der Waals highlighter (but the modular design of ToolSpace makes it easy to expand this list). The molecule importer accepts a hyperlink (URL) to a .pdb file and can import in a variety of styles; for example overlayed below for comparison. (Picture 2)



Location: http://vorlon.mathcs.emory.edu/tool

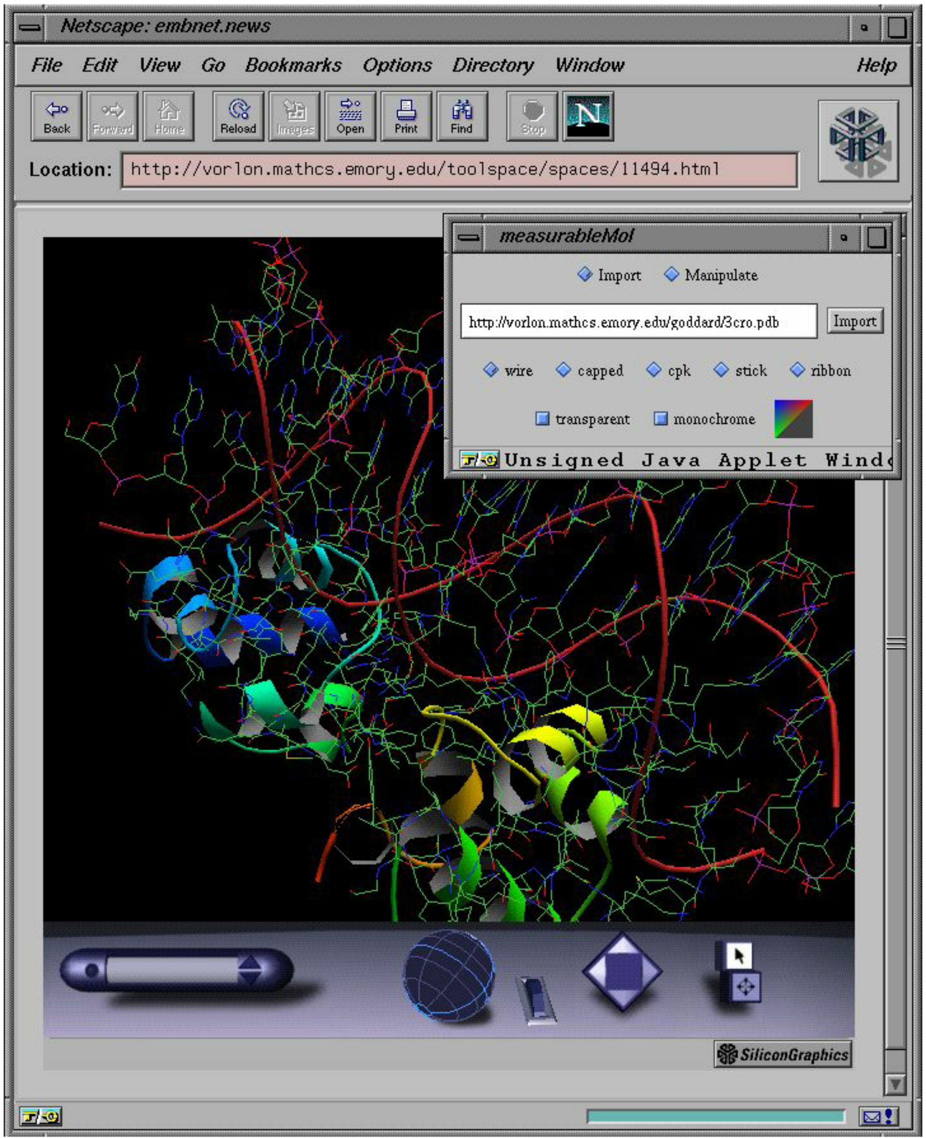
enter A: .678 39.7
enter B: 1.00 29.35
dist(A,B): 11.2962

Bob: What's the URL?
Adam: http://vorlon.mathcs.emory.edu/toolspace

Adam

◇ Import ◇ Manipulate
http://vorlon.mathcs.emory.edu/goddard/3cro.pdb
◇ wire ◇ capped ◇ cpk ◇ stick ◇ ribbon
□ transparent □ monochrome

Picture 1


Netscape: embnet.news
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Images Open Print Find Stop
Location:
measurableMol
Import Manipulate
Import
wire capped cpk stick ribbon
transparent monochrome
Unsigned Java Applet Wind
SiliconGraphics

Location: http://vorlon.mathcs.emory.edu/toolspace/spaces/11494.html

http://vorlon.mathcs.emory.edu/goddard/3cro.pdb

Picture 2

This shows wireframe (from pdb2vrml) and ribbon (from MolScript). Since ToolSpace extracts its own data directly from the .pdb files and displays the output from these other programs, enhancements made to pdb2vrml and MolScript will automatically be accommodated.

The "vectorMeasurer" allows such imported molecules to be measured. This is accomplished by clicking on an atom and then clicking a "label" button (A,B,C, or D). Computations such as: the distance from A to B, the angle ABC, the point to plane distance D to ABC, and the dihedral angle defined by AB and CD are then available.

Finally, we have the Van der Waal highlighter, which is a collaboration tool especially useful with wire frame models. It simply places a transparent appropriately colored and sized sphere around the currently selected atom.

Collaboration Tools

ToolSpace provides a variety of collaboration tools suitable for chemists. These include:

- a text annotator,
- a draggable arrow,
- a highlight box,
- an orientable plane, and
- a general object importer.

The text annotator allows you to place colored text anywhere in space, or, to attach that text to a particular object. Thus, if the molecule is later moved or rotated, the text annotation will remain in its correct position relative to the molecule.

The arrow is useful for pointing at objects in the virtual space, but in many cases the highlight box may be preferable since it can be used to restrict as well as focus attention. This is accomplished by having a box that is always transparent on the sides facing the viewer. Thus, what is behind the object of interest is hidden, but the object itself is still visible. Different users can choose highlight boxes of different colors.

The orientable plane is useful for illustrating certain types of symmetry. It can be positioned, oriented, and created in a variety of sizes and colors.

Finally, it may be desirable to import an arbitrary object. Currently, ToolSpace supports importing any JPEG or GIF image, or VRML 2.0 object from the web. Examples of such objects could be: models of physical systems or laboratory apparatus, diagrams, or molecular structures created with other software.

Collaboration

Scenarios

The original concept for ToolSpace was to create a collaboration environment for researchers distributed across the internet, but many other uses are being found:

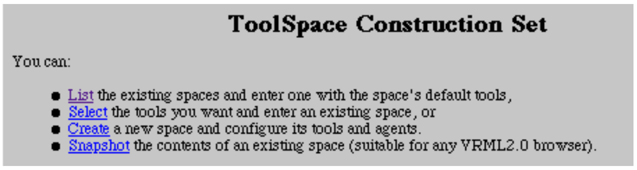
Collaboration over time. By using a "persistent" space researchers can work over time as well as distance (they need not both be viewing the space at the same time). The text annotation tool can be very useful in this mode of collaboration. An analogy for this might be a long-running chess game. For instance, two researchers in the same lab but on different schedules might find this useful.

On-line tutorials. Tutorials and other lectures can be held over the internet or over a campus network. For this, an audio or video broadcast solution is also desirable.

Interactive learning. Computer networking is not merely useful over large distances. Using ToolSpace, a classroom equipped with computers for each student and one for the lecturer allows questions and exploration not normally possible. Since viewing a three-dimensional object is an interactive process, students can become more involved in the material than if they were just watching it on video.

Setup

For collaboration to begin, a meeting place must be present. In ToolSpace, the place to meet is a "space". Management of spaces is done from the following menu (as it appears in your web browser)

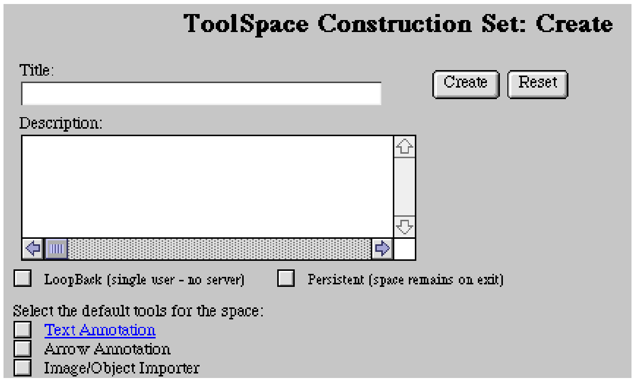


ToolSpace Construction Set

You can:

- List the existing spaces and enter one with the space's default tools,
- Select the tools you want and enter an existing space, or
- Create a new space and configure its tools and agents.
- Snapshot the contents of an existing space (suitable for any VRML2.0 browser).

Existing spaces can be listed and entered, or entered with a different set of tools than the original configuration. An important function is the creation of a space:



ToolSpace Construction Set: Create

Title:

Create Reset

Description:

□ LoopBack (single user - no server) □ Persistent (space remains on exit)

Select the default tools for the space:

□ Text Annotation
□ Arrow Annotation
□ Image/Object Importer

(People cannot meet until someone creates a space.) Spaces are created with a title, a short description, and a selection of default tools. If the space is made persistent then the space will not be disposed of even if no participants are currently active (normally a space is deleted when the last participant leaves as this allows the server to run unattended without consuming unbounded resources).

Future Work

ToolSpace can be enhanced both in terms of chemistry-specific tools and general collaboration tools.

For general collaboration, one of the most important aims is to enhance navigation beyond what a "stock" VRML viewer provides. A three dimensional "map" would be very useful since structures with their own coordinate systems (that often must be preserved to maintain their relationships with other structures) may initially be beyond the field of view when they are imported. In addition to that, movement through the virtual space needs to be improved with more absolute positioning (for example, dragging a page with a "hand" tool is often preferable to scrolling with scroll arrows). Much of this will be added as VRML browsers improve, but it is also possible to add some of this functionality through ToolSpace tools.

For chemistry in particular, we plan to add a number of tools focusing on manipulating structures, such as rotation around bonds, and best-fit alignment between structures. Further in the future, we plan to add more actual applications and interfaces to existing packages such as Amber.

References

1.ToolSpace, http://vorlon.mathcs.emory.edu/toolspace
2.http://ccf.mathcs.emory.edu/toolspace
3.PDB file format, http://www.pdb.bnl.gov/
4.pdb2vrml, http://ws05.pc.chemie.th-darmstadt.de/vrml/pdb2vrml.html
5.MolScript, http://www.avatar.se/molscript/
6.Amber, http://www.amber.ucsf.edu/amber/amber.html

Thanks

Research supported by Army Research Office grant DAAH04-96-1-0083 U. S. Department of Energy Grant No. DE-FG05-91ER25105, and the National Science Foundation Award No. ASC-9527186.

Book Review

Molecular Systematics (2nd Edn - Jan 1996)

David Hillis, Craig Moritz, Barbara Mable (Eds) Publ. Sinauer ISBN 0-87893-282-8- 137.95GBP/54.95USD Pback. 650 pages

reviewed by Andrew Lloyd, Irish EMBnet Node.

This is an edited book with 12 Chapters and, all told, 26 contributing authors. It covers in a Maniatis style some dozens of different techniques for abstracting information from biological specimens with a view to determining the relationships amongst individuals, populations and species. The techniques elaborated include Isozyme Electrophoresis, Molecular Cytogenetics, DNA-DNA hybridisation, PCR, RFLPs, Cloning and Sequencing. Much the longest chapter, Chapter 11 on phylogenetic inference, goes on to indicate what are the best ways of analysing the data obtainable from all these techniques in order to determine the evolutionary relationships among the samples and the taxa they represent.

Accordingly the range and quality of the information in the book is enormous. So at one end, it has a recipe for Ringer's Solution and step by step protocols for obtaining mitotic chromosomes from peripheral blood, fibroblast cultures, corneal epithelium and embryos. At the other end there is densely algebraic material on substitution matrices, and a comprehensive review of the philosophical and methodological underpinnings of almost every method of inferring evolutionary relationships that have been published in the last 20 years. No single person is going to be satisfied with the whole book. If you need to be told how to make up Ringer's solution then how much more imperatively do you need be to told how to implement Camin-Sokal Parsimony or Hadamard Conjugation?

The authors acknowledge that "debates between proponents of rival methodologies are often intense and sometimes unnecessarily acrimonious" and try not to approach the different analytical methods with too much 'baggage'. This is fair enough. Every dataset is unique, and specific data requires specific analysis. So no method, let alone any particular combination of options and parameters, is going to suit all problems. However such a non-judgmental approach rather leaves potential users in the dark about when, if ever, they can legitimately apply the techniques described. I am sure that many molecular biologists, acknowledging their ignorance of the theory and practice of phylogenetics, would welcome the assertive certainty of a committed Parsimonist to counteract the case made by Nick Goldman in favour of maximum likelihood (see review embnet.news vol4_2). Chapter 11 is not, and does not intend to be, the kind of document that will give you an unequivocal answer after one reading.

There is five page appendix listing programs and packages and a deferral to Joe "Phylip" Felsenstein to maintain and update such a program listing. So I suspect that a typical protocol will be

1) read Chapter 11
2) if necessary and if they are not too intimidating read the original papers cited
3) do a key word search of the Appendix to Chapter 11
4) download the appropriate software
5) hope that you can understand the documentation (if any)
6) format input data (almost certainly a non-trivial issue)
7) run program
8) scratch head over the output.

And I fear that, along the way of understanding the philosophy, the algorithm, the implementation, the output and the interpretation most of us will stumble at least once.

As a textbook for graduate students in molecular biological disciplines who want to chart a course in the controversial universe of phylogenetic inference, however, this book has much to recommend it.

Vector Sequence contamination in the public sequence databases

Rodrigo Lopez, Gunter Stroesser and Robert Harper

Services Programme - EMBL Outstation EBI - European Bioinformatics Institute
Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

More than 219 reports can be found in the public literature on the subject of contamination by vector sequences of the major sequence databases. All of these reports try to alert the scientific community of the potential pitfalls associated.

Vector sequences can be found in the sequence databanks for various reasons. The most common cause is that the submitters forget to remove them. Another is that they are submitted to the databanks because they are, after all, vectors which others might find useful.

The publication of a newly discovered gene or gene fragment requires that people submit the sequence to the public databanks in order to obtain an accession number. This number is unique for each submission and helps scientists identify their sequence amongst the myriad of sequences being produced each day by the many world-wide sequencing efforts. Furthermore, without an accession number, there can be no publication.

So scientists are often in a hurry to submit their sequences and innocently forget to remove a crucial part of the cloning vector they used to obtain it. This is frequently the poly-linker and inserts related to it. Other parts of cloning vectors are also to be found in eukaryotic sequence submissions, which by accident, have made their way into an otherwise genuine gene. In general this implies that some sort of rearrangement has taken place during the cloning experiments. Accidents do happen and the more we sequence the more submission with vector contamination will occur.

Nevertheless, the fact is that these anomalous sequences are already in the databases and there is very little that database curators can do to remove them. The task of the databanks is not to police the submitters but to help them in making their findings public.

Some reports suggest that when scientists are searching the databanks for similarities or homologies they should be made aware that the databanks are contaminated with vector sequences. Another option is that when scientists submit sequences to the databanks they should be reminded that it is their duty to submit only relevant scientific data. The estimates for contamination in various reports range from 0.23% to 4.00%, and with recent advances in the speed of sequencing and with fast submission throughput via web interfaces the problem of vector contamination is likely to be getting worse.

It is not all that gloomy though, for there are important lessons to be learned from reading these reports. At least two of the reports dealing with vector sequence contamination contain estimates of the degree of contamination in the databases. Furthermore, they point out that since the vector sequence is there (i.e. the poly-linker) then one has a well known template to test sequencing fidelity. They claim that quite often the similarity matches are quite good although degenerate. They conclude that quite probably one or more of the sequencing procedures failed. Genetic recombination is responsible for some of the observations but these reports claim that failure to check for traces of vector sequences is quite likely the most common source of contamination.

In an effort to assist the submitters the EBI is now providing a Vector Screening Service using the latest implementation of the BLAST algorithm and a special sequence databank know as EMVEC.

EMVEC is an extraction of sequences from the SYNthetic division of EMBL containing 2056 sequences commonly used in cloning and sequencing experiments. EMVEC is by no means a complete vector databank but EBI believes it is representative of the kind of material used in modern sequencing and should be useful to submitters. The databank will be updated with each release of EMBL and made publicly available on the EBI's ftp server (ftp.ebi.ac.uk) for those who wish to have it.

The service, which is currently provided via a WWW interactive form, can be found at: http://www2.ebi.ac.uk/blastall/vectors.html and in pages relevant to WWW interactive sequence submission at the EBI: http://www.ebi.ac.uk/submission/webin.html and http://www.ebi.ac.uk/ebi_docs/embl_db/ebi/databasehome.html

For some of the more interesting reports on this subject please refer to:

Reynolds TL. Technical report. Vector DNA artifacts in the nucleotide sequence database. Biotechniques. 1994 Jun; 16(6): 1124-1125.

Savakis C, et al. Contamination of cDNA sequences in databases. Science. 1993 Mar 19; 259(5102): 1677-1678.

Kristensen T, et al. An estimate of the sequencing error frequency in the DNA sequence databases. DNA Seq. 1992; 2(6): 343-346.

Yoshikawa T, et al. Contamination of sequence databases with adaptor sequences. Am J Hum Genet. 1997 Feb; 60(2): 463-466.

Little P. DNA sequencing. Complementary questions. Nature. 1991 Jul 4; 352(6330): 20-21.

Sibbald PR, et al. Consistent policy? Nature. 1992 Jan 9; 355(6356): 103.

Hodgson CP. Cloning vector artifacts in the DNA database. Biotechniques. 1990 Jul; 9(1): 54-55.

Lamperti ED, et al. Corruption of genomic databases with anomalous sequence. Nucleic Acids Res. 1992 Jun 11; 20(11): 2741-2747.

Rice CM, et al. Submission of nucleotide sequence data to EMBL/GenBank/DDBJ. Methods Mol Biol. 1994; 25: 413-424.

Binns M. Contamination of DNA database sequence entries with Escherichia coli insertion sequences. Nucleic Acids Res. 1993 Feb 11; 21(3): 779.

Lopez R, et al. Database contamination. Nature. 1992 Jan16; 355(6357): 211.

Asia Pacific Bioinformatics Network

Jingchu LUO(1), Christoph Sensen(2), Tin Wee TAN(3) 1 Centre for Bioinformatics, Peking University, China 2 Canadian Bioinformatics Resource, National Research Council, Canada 3 Bioinformatics Centre, National University of Singapore, Singapore

The Asia Pacific Bioinformatics Network (APBioNet) is dedicated to the advancement of the field of bioinformatics among Asia Pacific economies. Specifically, APBioNet focuses on the development of the bioinformatics infrastructure, the exchange of data and information, the development of training programs, workshops and symposia and the stimulation of collaborations in the field of bioinformatics. Nearly twenty representative organisations and institutions from more than ten economies including Australia, Canada, China, Hong Kong, India, Japan, Korea, Malaysia, Russia, Singapore and USA are currently part of this growing network.

Researchers in these institutions, and beyond, are able to intercommunicate via the Internet, especially through the high-performance advanced second generation network. This ATM-based Asia Pacific Advanced Network (APAN) was set up between some of the APBioNet member countries and will be extended to the other members. The use of the APAN system is due to the successful joint collaboration between APAN and APBioNet through the APAN

Bioinformatics Working Group. Soon, based on the recommendation of the European organisation responsible for advanced research networking (DANTE), the European research network connecting all national research networks in Europe may link to the APAN via a potential landing point in South Korea. This also links with Japan and from thence to other Asia Pacific networks, and to the Science Technology and Research Transit Access Point (STARTAP).

It is hoped that most of the Asia Pacific Economies will be able to establish national bioinformatics networks. Whereas some of the APBioNet countries (USA, Japan and Australia) have many years of experience, widespread dissemination of the technology and a high level of awareness, bioinformatics is still relatively new in a large part of Asia. Consequently, there is a need for a vehicle to drive manpower awareness and skill development through co-operation and assistance between developed and developing economies. One of the first priorities is to conduct courses, workshops, seminars, conferences etc. at institutional, national and international levels. In the span of the first nine months since its inauguration, APBioNet has co-organised and endorsed more than ten events.

APBioNet's mission also includes the production and provision of bioinformatics resources such as databases, software, guidelines and network communication. In this regard its partnership with APAN has provided the basic framework for further network communication. Another aspect of APBioNet's mission is to provide specialist expertise and a forum for regional communication. The ultimate aim is that it can foster productive and outstanding research collaboration between Asia Pacific researchers and the rest of the world.

It is important for APBioNet to initiate discussions on how to partner EMBnet in reaching out to those countries in Asia Pacific and help them to set up national nodes and national bioinformatics networks. Part of APBioNet's mission is to act as a consolidated regional body that can act as a single entity in dealing with external groups, thereby presenting a unified voice in representing the region. In this regard a strategic partnership with EMBnet at the regional level would be most desirable.

APBioNet is a young nascent organisation that is rapidly growing to meet the challenges of a global information infrastructure that knows no boundaries. As it serves a very technologically heterogeneous and culturally diverse region, APBioNet faces numerous challenges. As such, strategic partnership with and assistance and advice from more experienced organisations would be most helpful.

For more information: http://www.apbionet.org

New Web Resources

AMAS

http://barton.ebi.ac.uk/servers/amas_server.html

This server allows you to run AMAS on your own multiple sequence alignment.

MutRes, Mutation resources

http://srs.ebi.ac.uk/srs5bin/cgi-bin/wgetz?-fun+PageQueryForm+-l+MUTRES

This is collection of resourses useful for sequence level mutation research. It's primary goal is to collect information about various single locus databases

Jpred

http://barton.ebi.ac.uk/servers/jpred.html

A consensus method for protein secondary structure prediction

EBI - Genetic Code Viewer

http://www2.ebi.ac.uk/cgi-bin/mutations/trtables.cgi

This is a tool for browsing different versions of genetic code used by various taxonomic groups.

EBI - DNA Mutation Checker

http://www2.ebi.ac.uk/cgi-bin/mutations/check.cgi

This is a tool to help researchers and database curators to verify the transription and translation effects of DNA level sequence variation.

EBI - Reverse Translator

http://www2.ebi.ac.uk/cgi-bin/mutations/revtransl.cgi

This is a simple tool for calculating of probable DNA level point mutations underlying known amino acid substitutions.

MolScript Web Site

http://www.avatar.se/molscript/

MolScript is a program for displaying molecular 3D structures, such as proteins, in both schematic and detailed representations.

Biochemistry Easy Search Tool

http://www.worthington-biochem.com:8080/

Indexes educational sites with biological content useful to biochemists and other life sciences researchers

Biochemistry Journals

http://www.sciencekomm.at/journals/biochem.html

Directory of internet addresses for biochemical, biophysics and molecular biology journals

PCR Forum

http://sunsite.Berkeley.EDU/pcr/

This is an interactive site developed at UC Berkeley and devoted to discussion and debate about the past, present, and future of PCR (polymerase chain reaction) technology

EMBnet Nodes news

Node News from Norway

EMBnet Norway has recently undergone a change in its economic funding. Previously we recieved most of our funding from the Norwegian Research Council. Now we are supported by the University of Oslo in collaboration with other Norwegian universities, colleges, hospitals and other institutions that are using our services.

Another notable change is that the node has a substitute manager, Karin Lagesen. Linda Akselberg Langholm has taken a year's leave from the node. Karin' was educated at the University of Bergen. Her background is in both informatics and in molecular biology.

Node News from Ireland

The funding of the Irish National Centre for BioInformatics (INCBI) has had a tendency to be uncertain and intermittent, with comparatively short-term contracts and little money available for hardware upgrades or software development. The host institution, Trinity College Dublin, has recently allocated a two year grant to support bioinformatics in full from its Strategic Research Fund. This will involve us in both undergraduate and post-graduate teaching and course development.

Indian Node News

The Centre for DNA Fingerprinting and Diagnostics (CDFD) has obtained a 64 Kbps leased line link from the Department of Telecommunications (DOT), Hyderabad for its EMBnet activities. In addition, we have a 28.8 KBps Dial-up line. We are equipped with state-of-the-art hardware and software in bioinformatics. We will become fully operational as the EMBnet node once our Origin 2000 arrives. We have organised a workshop on macromolecular sequence analysis using GCG for our scientific and technical staff in September 1998. Please visit our homepage http://www.cdfd.org

10 years of Biocomputing at ICGEB

ICGEB, the International Centre for Genetic Engineering and Biotechnology in Trieste, Italy, has its 11th anniversary this year. Biocomputing started in 1988 with a user course that became an annual event. It was a modest start with borrowed computer facilities and it was only in 1990 that we started our own computer resource called ICGEBnet. Since then ICGEB has grown into an organisation of 60 member countries. A second laboratory in New Delhi has been built. There is continued interest in providing biocomputing services to developing countries, for many of which ICGEB provides a unique link to updated molecular biology databases and to up to date services. ICGEB spends an estimated 300,000 USD on bioinformatics each year, and this includes fellowships, grants, a series of courses as well as the maintenance of the computer resource which is done by 3 people. We are especially proud of the courses which are often co-sponsored by EMBnet. We have had over 700 students who received hands-on computer training in Trieste. We also contribute to a biophysics PhD course and this year we have organised a NATO workshop on structural biology and genomics. The rest is science...

Sandor Pongor, Maria Verina, Kristian Vlahovicek

ICGEB homepage http://www.icgeb.trieste.it/

ICGEBnet Biocomputing Resource http://www.icgeb.trieste.it/net/

Bioinformatics courses at ICGEB http://www.icgeb.trieste.it/net/netcourse.html

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:jrvalverde@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_3/contents.html***
- ***http://www.be.embnet.org/embnet.news/vol5_3/contents.html***
- ***http://www2.ebi.ac.uk/embnet.news/vol5_3/contents.html***
- ***http://www.ie.embnet.org/embnet.news/vol5_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.ebi.ac.uk in the directory pub/embnet.news/***
- ***ftp.ie.embnet.org in the directory pub/embnet.news/***

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

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

Back issues are available at most of these sites.

Publisher:

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

Editorial Board:

embnet.news
Vol.5, No.3, 1998
30 September 1998
ISSN 1023-4144