embnet.news

Volume 3 Nr2 (ISSN1023-4144)

July 21, 1996

SWISS-PROT should have been 10 years old in July 1996 but it may disappear on June 30, 1996". Imagine the impact that such headline news would have on the international bioinformatics community. Well, that was the news that was broadcast from Geneva at the beginning of May 1996. The SwissProt team, led by Amos Bairoch of the Medical Biochemistry Department of the University of Geneva and Rolf Apweiler of the EBI produce a number of key, high added-value protein databases, including SWISS-PROT itself, the PROSITE motifs database, the ENZYME nomenclature database and the SWISS-2DPAGE database.

About two years ago, the Swiss decided that, as SWISS-PROT was used by the international community, it should not be funded solely by money reserved for national projects. Accordingly, an EU grant proposal was submitted asking for some 12 positions in Switzerland, the EBI, Ireland, Israel and France to maintain, develop and extend the existing service. This grant was favourably reviewed but ultimately rejected by Brussels. Furthermore, grants from Swiss funds which were contingent upon European funding were put into jeopardy. Hence the "End of SWISSPROT" announcement which was coupled with a request for demonstrations of support.

The letters of support were immediately forthcoming and within a month some 1,500 scientists from 39 different countries responded. An interim solution has now been worked out which will keep SWISS-PROT afloat and in the public domain for another six months until a new funding proposal can be submitted. Good software and valuable databases really do cost money to maintain.

You can get a status report on the SWISS-PROT situation via the Announcements section of this issue of embnet.news. In this issue you can also find an interview with Doug Brutlag of Stanford University, a new method for creating non-redundant databases and a tip for running menu-driven software more efficiently. We hope you have good weather for your summer vacations.

The embnet.news editorial team:

Alan Bleasby
Reinhard Doelz
Robert Herzog
Andrew Lloyd
Rodrigo Lopez

Database and Software Developments

The BioImage Data Base

Jose Maria Carazo, Madrid

We are currently experiencing an explosion in the amount of volume image data obtained from the various kinds of techniques used in basic research of biological structures. By volume image data we refer to three-dimensional information that is represented in the form of multidimensional raster images. Provided that volume data are of sufficient resolution, atomic coordinate data can be derived from these to model the data in terms of individual amino acids or nucleotides. Similarly, atomic coordinate data can be converted to volume data for quantitative comparisons with data of a lower resolution. The information is in the form of multidimensional images of structures whose sizes span several orders of magnitude, from macromolecules at atomic scale resolution to entire cells and organisms.

Volume image data are acquired by X-ray crystallography, neutron diffraction, NMR, various forms of electron microscopy, scanning probe microscopy and light microscopy, and other less common techniques such as acoustic and X-ray microscopy. As a result of advances in instrumentation and computer technology, these techniques have all made substantial progress recently, and we have now truly entered an era of electronic image representation. This development is about to profoundly change the ways

Contents

Editorial	1
Database and Software Development :	
The BioImage Data Base	1
Software Development : CLEANUP	3
BITS : How safe is Email ?	6
INTERviewNET - Doug Brutlag - Stanford University	7
TIPS from the Computer Room	8
Announcements	9
The EMBnet Nodes	11
embnet.news information	12

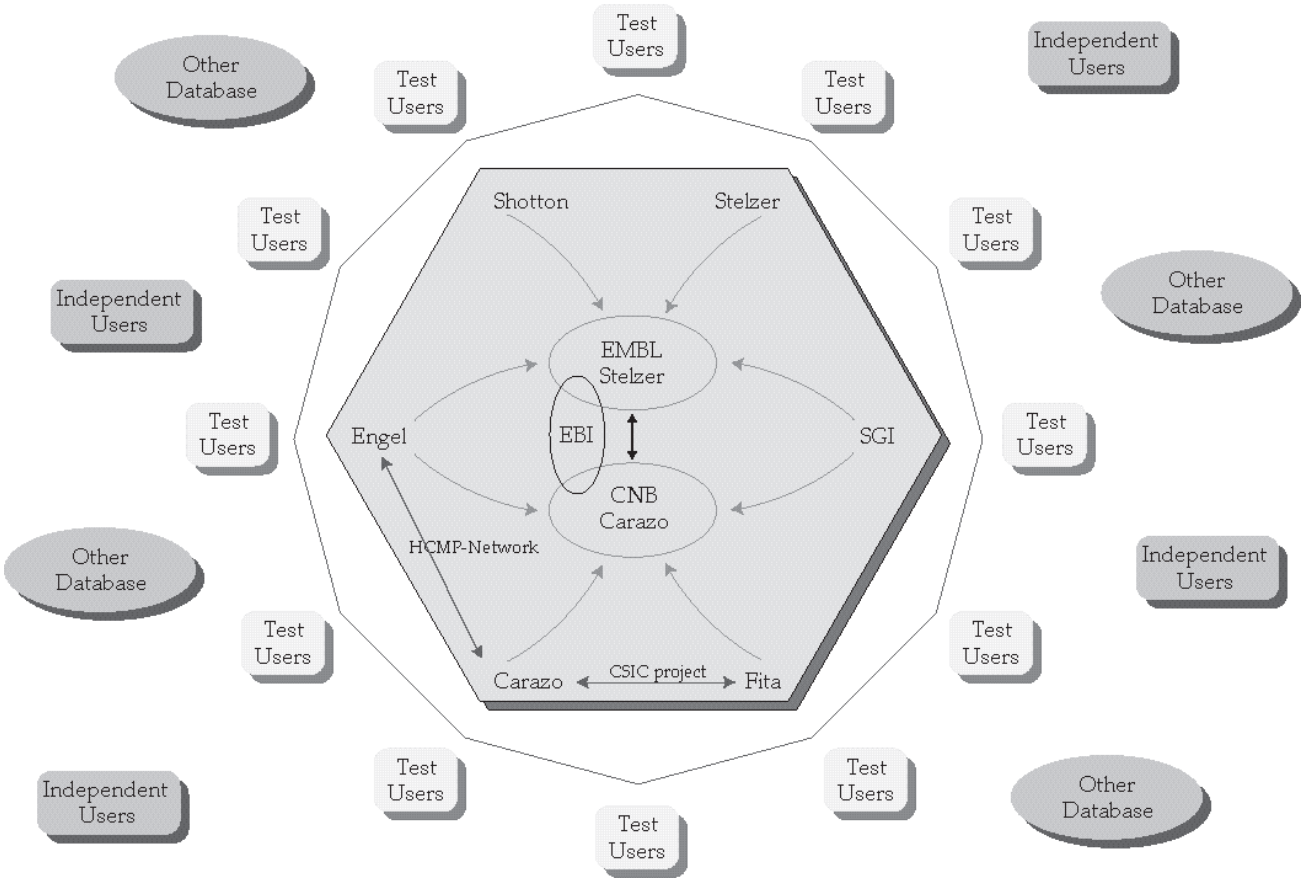
in which biomolecules, macromolecular assemblies, organelles and cells and their mutual interactions may be studied.

Whereas databases exist for atomic coordinates and sequences of proteins, none are available for volume image data. It is in this context that our group, together with the American group of Dr. J. Frank, started thinking, back in 1993, about organizing this type of information into a new data base. Indeed, EMBnet has already been very much aware of the importance of organizing key biological data into well structured data bases, and the initial support of EMBnet in the form of "seed funds" as well as general advice has been instrumental in correctly placing this topic into the more general area of biological databases.

We are all aware that scientific data is being produced all over the world, and that it is precisely the potential to combine data from all these distributed sources that constitutes one of the most valuable assets of a good database. However, we are also all aware that differences in funding schemas between different institutions (mainly between the EU and the US) may create some difficulties in organizing a concerted global action generally accepted by the scientific community at large, especially if extensive overlaps between different initiatives needs to be avoided. This situation certainly arises since different, independent projects have to be launched in the US and the EU in order to obtain appropriate financial support. We certainly would like to learn from the history of other databases how they have handled these problems, and EMBnet could well play an important advisory role in this matter in the coming years.

It has been in this context of several years of preliminary studies that the European Union has recently agreed to start contract negotiations with a team of European laboratories to undertake an initiative that, under the name of "BioImage", intends to develop such a data base of volume data in close coordination with other emerging initiatives in the US (most notably, the NSF-funded initiative in the macromolecular field, led by Dr. J.Frank). The BioImage European initiative involves seven laboratories, and is coordinated from the EMBnet National Node in Madrid. Additionally, a group of "test users" laboratories has beeen identified.

The Figure below indicates this situation:



Management Structure of BioImage

The general objectives of this project are: (1) To organize the diverse forms of volume image data, that are produced by a number of different techniques into a single archive that is fully compatible with other databases such as PDB. (2) To develop the infrastructure required to access such volume image data world-wide. (3) To devise ways to search the database efficiently. (4) To develop specific user-friendly volume handling tools that enable us to make new data correlations. (5) To link the volume image data sets to other sources of scientific data (protein and nucleic acid sequence data, atomic coordinates, literature citations, etc.) in an efficient and transparent way. And (6) to facilitate collaborations with software companies, by generating precise image definitions based on well-accepted standards.

This project has the potential to influence future biological research, as it will provide possibilities to analyze and combine structural data at the molecular and cellular level in ways that have hitherto been impossible. In addition, we anticipate that it will stimulate developments in the commercial sector in several areas.

We wish to acknowledge with great appreciation the support that the EMBnet community has always provided to this initiative and which will hopefully continue into the future.

Jose Maria Carazo, PhD
BioComputing Unit, Head
Centro Nacional de Biotecnologia-CSIC
Universidad Autonoma
28049 Madrid

Tf.: + 341 585 4543 or + 341 585 4510
Fax: + 341 585 4506

(From June to August 96 on a sabatical stay at the following address:

Dr.J.M.Carazo (c/o Dr.J.Frank), Wadsworth Center,
Empire State Plaza, P.O. Box 509,
Albany, NY 12201-0509.
Tf.: + 1 518 486 7318
Fx.: + 1 518 486 2191
e-mail: carazo@wadsworth.org)

Software Development

CLEANUP: a fast computer program for cleaning nucleotide sequence databases from redundancies

Graziano Pesole[1,2], Giorgio Grillo[3], Marcella Attimonelli[4] and Sabino Liuni[1,3]

1 Area di Ricerca del CNR, Bari, Italy
2 Dipartimento di Biologia, Difesa e Biotecnologie Agro-Forestali, Universita della Basilicata, Potenza, Italy
3 Centro di Studio sui Mitocondri e Metabolismo Energetico, CNR, Bari, Italy
4 Dipartimento di Biochimica e Biologia Molecolare, Universit di Bari, Italy

Correspondence to:
Dr. Graziano Pesole
Area di Ricerca del CNR
via Amendola 166/5, 70126 Bari, Italy
Tel. +39-80-5443305 - Fax +39-80-5443317 - E-mail: graziano@area.ba.cnr.it

INTRODUCTION

A key concept in comparing sequence collections is the issue of redundancy. The production of sequence collections cleaned from redundancy is useful both in performing statistical analyses and accelerating extensive database searches on nucleotide sequences. Publicly available databases contain multiple entries of identical or almost identical sequences. Performing statistical analysis on such biased data runs the risk of assignation of high significance to non significant patterns.

Given that an unambiguous definition of redundancy is impracticable for biological sequence data, the program presented uses a quantitative description of redundancy based on a measure of sequence similarity. A sequence is considered redundant if it shows a degree of similarity and overlap with a longer sequence in the database greater than a threshold fixed by the user. Here we present a new algorithm, based on an “approximate string matching” procedure, which is able to determine the overall degree of similarity between each pair of sequences contained in a nucleotide sequence database without performing the time consuming task of pairwise global best alignments. The algorithm generates a new purified dataset from the redundant nucleotide sequence collection using cutoff parameters set by the user. The method is fast, sensitive and automatic.

Sequence 1	Sequence 2	L1	L2	P1	P2	OS1	OS2	%Ident	%Overl.
MIHSCG	HSCOXII	16569	708	7584	1	709	708	99.86	100.00
.	HSMTALT1	.	360	16024	1	360	360	100.00	100.00
.	HSMTNA02	.	360	16024	1	360	360	98.33	100.00
.	HSMTNA03	.	360	16024	1	360	360	98.06	100.00
.	HSMTNA04	.	360	16024	1	360	360	98.89	100.00
.	HSMTNA05	.	360	16024	1	360	360	98.33	100.00
.	HSMTNA06	.	360	16024	1	360	360	98.06	100.00
.	HSMTNA07	.	360	16024	1	360	360	97.78	100.00
.	HSMTNA08	.	360	16024	1	360	360	98.61	100.00
.	HSMTNA09	.	360	16024	1	360	360	98.89	100.00
.	HUMMTALT0	.	360	16024	1	360	360	99.72	100.00
.	MIHS01	.	608	324	1	610	608	99.51	100.00
(c)									
.	MIHSAIA	.	360	16024	1	360	360	98.89	100.00
.	MIHSAIAA	.	360	16024	1	360	360	99.44	100.00
.	MIHSAIAB	.	360	16024	1	360	360	99.17	100.00
.	MIHSAIB	.	360	16024	1	360	360	99.17	100.00
.	MIHSAIC	.	360	16024	1	360	360	99.44	100.00
.	MIHSAID	.	360	16024	1	360	360	99.17	100.00
.	MIHSAIE	.	360	16024	1	360	360	98.06	100.00
.	MIHSAIF	.	360	16024	1	360	360	98.61	100.00
.	MIHSAIG	.	360	16024	1	360	360	98.06	100.00
.	MIHSAIH	.	360	16024	1	360	360	98.33	100.00
.	MIHSAII	.	360	16024	1	360	360	98.33	100.00
.	MIHSAIJ	.	360	16024	1	360	360	98.33	100.00
.	MIHSAIK	.	360	16024	1	360	360	98.61	100.00
.	MIHSAIL	.	360	16024	1	360	360	98.89	100.00
.	MIHSAIM	.	360	16024	1	360	360	98.61	100.00
.	MIHSAIN	.	360	16024	1	360	360	98.33	100.00
.	MIHSAIO	.	360	16024	1	360	360	98.06	100.00
.	MIHSAIP	.	360	16024	1	360	360	98.89	100.00
.	MIHSAIQ	.	360	16024	1	360	360	98.89	100.00
.	MIHSAIR	.	360	16024	1	360	360	98.89	100.00
.	MIHSAIS	.	360	16024	1	360	360	98.61	100.00
.	MIHSAIT	.	360	16024	1	360	360	98.61	100.00
.	MIHSAIU	.	360	16024	1	360	360	99.17	100.00
.	MIHSAIV	.	360	16024	1	360	360	99.17	100.00
.	MIHSAIW	.	360	16024	1	360	360	98.89	100.00
.	MIHSAIX	.	360	16024	1	360	360	98.89	100.00
.	MIHSAIY	.	360	16024	1	318	318	98.43	88.33
.	MIHSALT01	.	360	16024	1	360	360	98.61	100.00
.	MIHSALT02	.	360	16024	1	360	360	98.89	100.00
.	MIHSALT03	.	360	16024	1	360	360	99.17	100.00
.	MIHSALT04	.	360	16024	1	360	360	98.61	100.00
.	MIHSALT05	.	360	16024	1	360	360	98.61	100.00
.	MIHSALT06	.	360	16024	1	360	360	99.17	100.00
.	MIHSALT07	.	360	16024	1	360	360	98.89	100.00
.	MIHSALT08	.	360	16024	1	262	262	98.47	72.78
.			.	16287	264	97	97	98.97	26.94
.	MIHSALT10	.	360	16024	1	360	360	98.61	100.00

(..... Continued .....)

Table I. Output produced by CLEANUP on a sample of 362 sequences extracted from release 41 of the EMBL database (L1, L2 : length of primary (L1) and secondary (L2) sequence; P1, P2: starting position of the similarity region in the primary (P1) and secondary (P2) sequence; OS1, OS2: length of the overlapping segment in the primary (OS1) and secondary (OS2) sequence, (c) : similarity match found on the reverse / complementary strand of the secondary sequence).

A detailed description of the algorithm is reported in Grillo et al. (1996).

RESULTS AND DISCUSSION

The CLEANUP algorithm has been tested on a dataset of 362 sequences accounting for all the human mitochondrial DNA sequences available in release 41 of the EMBL database. This dataset provides a very simple check of the cleaning accurateness of the algorithm as we can expect that after cleaning only a single sequence corresponding to the complete human genome is retained and all the others are erased. Table I reports part of the output produced by the CLEANUP application when run against the human mtDNA collection.

In this test application, 20 sequences have been retained out of the 362 instead of the expected single one corresponding to the complete human mtDNA genome.

Figure 1 shows a CLEANUP session on a dataset of 2400 sequences (5,523,925 nt) comprising all Drosophila entries in the Invertebrate division of Genbank collection (release 90). 400 out of the 2,400 sequences (420,707 nucleotides) were removed by CLEANUP as both their degree of sequence similarity and overlap exceeded the fixed threshold of 95%.

CLEANUP is an invaluable and efficient tool for all database users/developers to remove or classify redundant information.

Acknowledgements

We thank C. Saccone for helpful comments. This work was partially financed by MURST, Italy, Progetto Finalizzato Ingegneria Genetica, CNR, Italy and by the EMBnet Research and Development Project Committee.

Bibliography

Etzold, T., and P. Argos. 1993. SRS, an indexing and retrieval tool for flat file data libraries. Comput. Appl. Biosci. 9: 49-57.

Gouy, M., C. Gautier, M. Attimonelli, C. Lanave, and G. Di Paola. 1985. ACNUC - a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. Comput. Applic. Biosci. 1: 167-172.

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

% **Cleanup -sim=95 -overlap=95**

Cleanup generates a non redundant sequence data library from any set of sequences.

Cleanup of what sequences ? **in:dro***

What should I call the output searching file (* in_dro.s *) ?
What should I call the file listing erased sequences (* in_dro.e *) ?
What should I call the file listing retained sequences (* in_dro.nr *) ?

Sequences	= 2400	Nucleotides	= 5523925
% Overlapping for cleaning	= 95.00	% Similarity for cleaning	= 95.00
Sequence matches	= 424	Searching C.P.U. time	= 158.84
Erased sequences	= 400	Erased nucleotides	= 420707

Figure 1. Sample run of a CLEANUP session on a dataset of 2400 Drosophila entries. User input are boldface typed. CLEANUP generates a file listing all non-redundant sequence entry names which can be used as input to programs of the GCG package as well as with retrieval programs such as SRS (Etzold and Argos 1993) or ACNUC (Gouy et al. 1985). It optionally also generates a sequence data file in Pearson/FASTA format.

BITS

How safe are electronically communicated messages?

Reinhard Doelz, formerly of BioZentrum Basel.

Principle

If two persons wish to communicate, they may decide to travel (and hence meet personally), use the telephone, or exchange written messages. Writing down statements, messages or opinions using pen and paper changes the paradigm of interactive information exchange (as encountered in meetings) to “asynchronous” procedures, which are required by the communication model of exchanging “letters”. The latter have two important characteristics:

1. Messages are typically a summary of statements and facts rather than singular arguments.
2. Messages take time to travel, and therefore the time scale of a dialogue expands to hours, days or even weeks.

Modern electronic documentation reduces the travel time of the messages and relies on message exchange programs. The most common way to transport a message from one user to another is referred to as "sending electronic mail".

>From a technical standpoint, electronic messages represent data files, which are copied from one user to another. As millions of users might use millions of different computers in a shared fashion, two basic prerequisites must be met:

1. The user must be known outside the boundary of his or her computer. This implies an address convention, and a technology which uniquely maps these addresses onto the international networks, hence, procedures must be provided to transport the message transparently.

2. The user must rely on the confidentiality of the system ‘as is’, and be aware of the fact that system and security auditing might reveal the contents of his or her message to third persons.

>From the perspective of security, two areas deserve attention: The transport machinery (a system manager’s issue), and the message alteration caused by the option to format, encode or encrypt confidential or binary messages (an action to be taken by the sender). Specifically, the processing of the message by the receipient is potentially insecure as the data received is typically “trusted” to be message text and not malicious material.

Impact of the transport system

Users send and receive messages as private persons. As the receiving system will need to deliver the mail individually, it is required that the program performing this task may act with ‘super-user’ privileges and write into the individual’s mailbox.

This behaviour has been recognized as sensitive in security terms and considerable effort is taken to avoid unauthorised access to this mechanism. Regardless how secure the machine is set up, it is usually not possible for the mail program at the receiving end to authenticate the sending person. Unless special caution is in effect, electronic mail can be faked easily. Authentication and security are possible but add considerable overhead, specifically on personal computers. For the sake of simplicity and ease of use, therefore, security is not applied at many installations.

Therefore, be careful if you receive exciting messages from apparently well-known sources as the originator might be faked on purpose. Recently it has been shown that the mail function in web browsers might get abused to send unsolicited messages without approval. This abuse of a submission page might even trespass firewalls and secure environments.

Impact of the mail contents

Messages transferred by electronic mail are not necessarily restricted to plain “text” containing the characters 0-9, a-z and A-Z. Formatting, national characters and any image information enrich today’s scientific documents and would be lost if only text could be transferred. Therefore the Multimedia extensions to mail (MIME) have been invented. This standard permits the smooth transfer of any kind of binary information in a simple but effective mechanism: all data are “encapsulated”, i.e. converted into a format which allows the reconstruction of the data at the receiving end. The advantage is obvious. Assuming that the electronic mail partner can recognize the program used by the sender (such as a word processor, charting or drawing software) any document can be transmitted without loss of information. The drawback of the method arises from the encapsulation procedure. Without a conversion program (a so-called MIME aware system) and the corresponding software no utilization of the transmitted message is possible. However, technology has improved and two mechanisms avoid a deadlock: Either importing filters as delivered by the software vendor or

general-purpose conventions (e.g., the Rich Text Format (RTF) for word processing) allow message exchange even if the application software packages do not match exactly.

The security aspect of MIME is twofold. First, recent developments have shown that data may carry malicious information which will damage the receiving user's environment (e.g., the WORD "macro virus"). Secondly, MIME-aware mailers have been developed which will view the document directly by calling the corresponding program for inspection of the message. This is very convenient but implies a computation at the receiving end with potentially malicious effects as the converted data will cause a program launch regardless of the user's precautions against viruses.

It might be self-evident but should be kept in mind that any binary or other download from public sources, be it via electronic mail or directly from FTP servers, might be designed to purposely compromise your system's security or operation in general. It should be noticed that even time-expiring shareware, including the released 'beta test' software systems, might be a threat for your service because a well-known tool stops functioning.

Summarizing, the user benefits from electronic communication but must keep in mind that security of today's message systems is lower than the expectation.

InterVIEWnet

Andrew Lloyd talks to Doug Brutlag from Stanford University

ATL. What brings you to Europe ?

DB. I was invited to give an overview to the SmithKline Beecham Scientific Advisory Boards on the value of bioinformatics in target discovery and drug design. Also David McConnell invited me over to visit INCBI in Dublin.

ATL. How long have been labelling yourself "bioinformatician" ?

DB. Let me first try my definition of bioinformatics?

ATL. Sure.

DB. Bioinformaticians study the flow of information from the genotype to the phenotype. This is also the primary goal of molecular biology. So bioinformatics is a subset of computational biology which overlaps genomics, databases and population biology. Genomics and bioinformatics are closely related: genomics is the science of acquiring genome information and bioinformatics deals with understanding that information. But, of course, the two interact with each other: you cannot organise the information unless you, to some extent, understand it.

ATL. So you are a "bioinformatician" ?

DB. Bioinfomagician maybe ! And to answer your original question, I guess I've been in the field for at least 30 years, since High School. Although I have been in turn a Biochemist, an enzymologist, a molecular biologist and a bioinformatician, the common thread has always been my interest in the biological information.

ATL. Under what circumstances did Intelligenetics come into being ?

DB. IG was formed to widely distribute software that had been developed on DEC computers at Stanford. So it was involved in porting the software to other platforms, providing software support and acquiring other software to incorporate into the suite.

I was particularly happy with the progress we made with the support of the US government between 1982 and 1987 with Bionet. Bionet was a resource, providing databases and software, first over commercial networks and then over ARPAnet, that was eventually accessed by over 2,000 labs. Each lab paid a $400 annual subscription, which about covered our telecommunication costs. It was not only US labs, but worldwide and was instrumental in facilitating a number of important European/US collaborations, in AIDS research for example.

IG spent five years from 1987-1992 running GenBank. In that time we shifted the paradigm, notably with respect to who was responsible for the annotation. Over the five years we reduced a two year lag beftween publication and entry into the database down to 2 days. This was achieved partly by providing the submitters tools to do the annotation and also by increasing the proportion of data contributed electronically from 5% up to 98%. Between 1987 and 1992, GenBank increased in size by a factor of x10, but we managed the project on a fixed cost. So productivity increased tenfold.

We subcontracted Los Alamos to deal with the sequences not submitted electronically and also annotated the backlog so that we were able to hand an up-to-date GenBank over to NLM in 1992. Some of the very early sequences are still not very well annotated and some of the new sequences are also poorly dealt with because the submitters are not highly motivated to provide all relevant information. In 1987 only

3 journals required sequence submission before publication, in 1992 that was up to 17 and now very few journals do not require this.

ATL. Do you think that the pharmaceutical industry is correct to be pouring money into bioinformatics ?

DB. Absolutely right! The potential for designing cures for many infectious and hereditary diseases is enormous. The pharmaceutical industry is pouring large amounts of money into the field, not only internally by hiring more bioinformaticians, but also externally towards training a new generation.

As the level of understanding of diseases and therapies can be increased by orders of magnitude through bioinformatics, so the pharmaceutical industry must increase its funding by orders of magnitude in the area of training. After all, if the industry is scooping up most of the available molecular biologists and bioinformaticians, it has a duty to ensure that there is a next generation of trained scientists. They cannot rely on the government to provide this. And they also have a duty to return sequences and other data to the public. Not only in the form of patented sequences but also formally published and submitted to GenBank/EMBL.

ATL. How do you cope with the data deluge at Stanford ?

DB. By ensuring that we have really high bandwidth networks. If there is essential data that needs to be available on a millisecond time scale then we cache that data at Stanford. Otherwise, not all the databases are available locally. We have distilled versions of various databases of local interest in ACeDB format - Drosophila, C.elegans and one human chromosome. And of course the cost of disk space is going way down. We treat disk-space as a fixed cost each year and find that the cost is falling about as fast as the space requirement is increasing. Our sequence server is managing fine with 20Gb of storage at the moment. Naturally the Internet outside of Stanford is somewhat beyond our control but, in the US, the National Infrastructure Initiative (NII), set up by Al Gore is doing a good job. Networks within the US seem to be getting faster and faster, but a few sites in the UK which I need are not very accessible.

ATL. What degree of bioinformatic infrastructural support do you supply at Stanford ?

DB. We provide a Bioinformatics Resource that makes sequence databases and tools and molecular modelling software available to about 1,000 users in 100 research labs in 35 different departments. Some departments are quite well served in themselves - the Genetics Department, for example, has a yeast genome group and a human genome group with their own computational resources.

We also have a system whereby the Stanford University librarians will search the databases for you. I've written them a protocol for how to do a homology search, so that if you are an infrequent user you can get this much information over the counter at the library.

Our core hardware is an internet server (Sun 690MP), a sequence database server (a 2 processor SPARC1000) and a 2 processor SGI Challenge L for molecular modelling.

ATL. And how do you find Ireland?

DB. Ireland quickly charmed me and I look forward to coming back when I have time to visit more of the country as a tourist. The most memorable feature was the genuine warmth and outgoing nature of the people I met, even strangers in the street and clerks in the shops.

Tips from the computer room

Driving menu driven programs

Andrew Lloyd, EMBnet Node Ireland

Programs that prompt you for information are very useful when you are just starting to use them; that way you are not allowed to omit any essential information. When you have run the same program many times with the same parameters but different input data, you may suddenly realise that repetitive tasks (like pounding the keyboard) are better handled by the computer than with human intervention. In GCG you can speed things up by appending all the parameters on the command line and editing the appended -options between each run of the program.

In this article I suggest ways in which you can automate menu-driven programs in a Unix environment. As an example, imagine that I need to perform a number of multiple sequence alignments with clustalW using different input files. If I create a script, or a file which contains the instructions that clustalW needs to create the alignment, then I can execute the script rather than clustalW itself; in this way I can save keystrokes and speed up the process.

The central idea in what follows is that you can use a script to feed a series of input commands, one at a time, to a program.

Note: The second line of clustalW instructions (where an Input File name is expected) is a $1. You could enter a real file name here, but the $1 allows you to enter the input file from the same line when you run your script. Assuming the

```
#!bin/csh
#The previous line specifies the 'shell' for this script
#Everything written to the right of a '#' is a comment
#and ignored.
#The next line means: run clustalw taking as input all the
#commands on each line until the next "!" appears
clustalw <<!
1                    #Sequence Input from Disc
$1                   #Input file from command line
2                    #Multiple alignments
1                    #Do multiple alignment now
                     #[RETURN] Take default output filename
                     #[RETURN] Take default name for guide tree
x                    #Stop displaying alignment
                     #[RETURN] to go to main menu
x                    #EXIT (leave program)
!
```

Script: Multiple sequence alignment from a file of catenated fasta (Pearson) format sequences.

above commands are saved into a file called 'cluscript', then:

%cluscript infile.seq

will run clustalw on the sequences contained in infile.seq.

You can run this job several times just changing infile.seq for different input file names:

%cluscript infile2.seq

Finally, what do you do if you want to run a computationally intensive job (as clustalW can be) in the middle of the night to spare the system and your fellow users. You can do this with the unix command "at":

% at 0300 script

which means: at 3 am execute the file called script. Unfortunately you cannot run this command with two input parameters so

%at 0300 cluscript infile.seq

will give you an error. Therefore, you have to do a little unix fudge to achieve what you want:

% echo cluscript infile.seq | at 0300

will 'pipe' (the vertical bar symbol) the command line you want through the 'at' command and allow you to run the program with the specified input file at the specified time.

You should be able to generalise the principals outlined above to run any program or sequence of programs: fewer keystrokes, more control, greater efficiency.

However, note that it is considered very antisocial to run jobs using the 'at' command on systems which run batch software. Use the batch queues instead.

Announcements

EMBnet Scientific Meeting

In connection with EMBnet Annual General Meeting on November 21 -24 1996, EMBnet Finland is organising a short scientific meeting under the theme:

Sequence information over-flow
How to cope with 100-fold increase in next five years?

The meeting will take place at Datatorium, CSC, Espoo on Friday 22nd November 9.00-12.00. The speakers include:

- William R. Pearson, University of Virginia
- Stephen Altschul, NCBI

The third speaker has not yet confirmed his participation. The meeting is open to all interested parties. No registration needed. Up-to-date information about the AGM and the meeting is available at this site
http://www.csc.fi/molbio/embnet/agm96.html

SwissProt Funding Crisis

You can get information about the initial funding crisis at SWISS-PROT, which is reported in the Editorial of this issue, from:

http://expasy.hcuge.ch/sprot/help-sprot.html

Recent developments on the subject are reported at:

http://expasy.hcuge.ch/sprot/rec_devlp.html

It is probably not too late to send letters of support or offers of help to:
e-mail: help.sprot@hon.ch
Fax: + 41-22-346 87 58
Mail: Amos Bairoch, Dept. Medical Biochemistry,
1 rue Michel Servet, 1211 Geneva 4, Switzerland

Careers in BioInformatics

Science, organ of the AAAS, is featuring Careers in BioInformatics as its current "NextWave" project. It features several role models who now call themselves bioinformaticians as well as a number of resources for obtaining appropriate training and education. See

http://www.aaas.org/nextwave/niches-bio

or if you have good connections to France try the European Science Foundation pages at;

http://www.esf.org/nextwave

These pages invite your contributions and questions.

Protein Structure Course

The Crystallography Department of Birkbeck College is running a part-time, London University course on the PRINCIPLES OF PROTEIN STRUCTURE on the Internet.

1. Introduction to Internet Resources
2. Protein Structure
3. Dissertation: structure, function and dynamics

The course covers one academic year of three terms, from September 30, 1996 to 4 July 1997. Costs are: 250 pounds sterling for EU students and 550 pounds sterling for other students.

For details of course contents, administration and registration URL http://www.cryst.bbk.ac.uk/PPS2/index.html

Contact: Jody McGill, Crystallography Department, Birkbeck College, London, WC1E 7HX. UK
Tel. +44 (0)171 631 6800 Fax. +44 (0)171 6316803
e.mail: j.mcgill@mail.cryst.bbk.ac.uk

The EMBnet Nodes

[AT] EMBNet (martin.grabner@cc.univie.ac.at)
VIENNA BIOCENTRE
University of Vienna, Vienna, Austria

[BE] BEN (rherzog@ulb.ac.be)
Brussels Free Universities,
Rhode-St-Genese, Belgium

[CH] Biocomputing Basel (info@ch.embnet.org)
Biozentrum der Universitaet,
Basel, Switzerland

[CH] ROCHE (doran@embl-heidelberg.de)
Hoffmann-La Roche,
Basel, Switzerland

[CH] SWISSPROT (bairoch@cmu.unige.ch)
Med. Biochem. Dept. CMU, University of Geneva
Geneva, Switzerland

[DE] EMBL (datalib@EMBL-Heidelberg.de)
European MolecularBiologyLaboratory,
Heidelberg, Germany

[DE] GENIUS
(dok419@genius.embnet.dkfz-heidelberg.de)
DKFZ, Heidelberg, Germany

[DE] MIPS (mewes@mips.embnet.org)
Max-Plank-Institut fur Biochemie,
Martinsried, Germany

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

[ES] CNB (carazo@samba.cnb.uam.es)
Centro national de Biotechnologia
CSIC, Madrid, Spain

[ES] TDI (dopazo@tdi.es)
Technologica para Diagnostico e Investigation
Madrid, Spain

[FI] CSC (Heikki.Lehvaslaiho@csc.fi)
Centre for Scientific Computing,
Espoo, Finland

[FR] CEPH (claude@genethon.fr)
GENETHON,
Evry, France

[FR] INFOBIOGEN (dessen@infobiogen.fr)
INSERM,
Villejuif, France

[GR] EMBnet Node (savakis@myia.imbb.forth.gr)
Institute of Molecular Biology and Biotechnology,
Heraklion, Greece

[HU] EMBnet (remenyi@abc.hu)
Agricultural Biotechnology Centre,
Godollo,Hungary

[IL] INN (isestern@weizmann.weizmann.ac.il)
Weizmann Institute of Science,
Rehovoth, Israel

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

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

[IE] INCBI (atlloyd@acer.gen.tcd.ie)
Irish National Centre for BioInformatics,
Dublin, Ireland

[NL] CAOS (jackl@caos.caos.kun.nl)
Katholieke Universiteit,
Nijmegen, Netherlands

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

[PL] IBB (piotr@ibbrain.ibb.waw.pl)
Institute of Biochemistry and Biophysics,
Polish Academy of Sciences, Warsaw, Poland

[PR] EMBnet (pfern@gulbenkian.pt)
Instituto Gulbenkian de Ciencia,
Oeiras, Portugal

[SE] EMBnet.se (gad@perrier.embnet.se)
Computing Department, Biomedical Centre,
Uppsala, Sweden

[UK] HGMP (mbishop@hgmp.mrc.ac.uk)
Human Genome Mapping Project Resource Centre,
Hinxton, Cambridge, United Kingdom

[UK] SEQNET (bleasby@daresbury.ac.uk)
Daresbury Laboratory,
Daresbury, United Kingdom

[UK] Sanger Centre (pmr@sanger.ac.uk)
Hinxton Hall
Cambridge, United Kingdom

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

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/vol3_2/contents.html***
- ***http://www.be.embnet.org/embnet.news/vol3_2/contents.html***
- ***http://www.no.embnet.org/embnet.news/vol3_2/contents.html***
- ***http://www.ie.embnet.org/embnet.news/vol3_2/contents.html***

A Postscript version (ISSN 1023-4144) 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.

Online:
- ***http://www.uk.embnet.org/embnet.news/info.html***

Postscript by ftp:
- ***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/***

Publisher:

EMBnet Administration Office.
c/o J.Franklin,
ASFRA BV,
Voorhaven 33,
1135 BLEDAM.
The Netherlands

Editorial Board:

Alan Bleasby, Daresbury Laboratory, UK
(bleasby@daresbury.ac.uk)
FAX +44-1-925-603100
Tel +44-1-925-603351

Reinhard Doelz, Sandoz, CH
(REINHARD.DOELZ@sandoz.com)
Tel +41-61-423 8214

Robert Herzog, BEN, Free University Bruxelles, BE
(rherzog@ulb.ac.be)
FAX +32-2-6509767
Tel +32-2-6509762

Andrew Lloyd, INCBI, Trinity College Dublin, IE
(atlloyd@acer.gen.tcd.ie)
FAX +353-1-679-8558
Tel +353-1-608-1969

Rodrigo Lopez, EBI, Hinxton Hall, UK
(Rodrigo.Lopez@ebi.ac.uk)
FAX +44 1223 494423
Tel ++44 (0)1223 494468

embnet.news
Vol.3, No.2, 1996
21 July 1996
ISSN 1023-4144