embnet.news

Volume 3 Nr.1 (ISSN1023-4144)

May 10, 1996

We live in interesting times. Growth, fusion, and 'right sizing' are today's buzz-words and they affect not only the environment to which we are accustomed but also the methods we will use in the near future.

The exponential growth of the databases is a pervasive problem for us all. In this issue, Andrew Lloyd summarises a debate about possible solutions and plausible futures that was held on the embnet.general newsgroup in November. One of the problems for service providers is trying to maintain a service, with which naturally conservative end-users are familiar, in the face of this deluge of data. This is not to say that users are unable to embrace new technology. SRS is a case in point, having experienced an enormous increase in popularity over the last couple of years.

The database growth also affects present structure of EMBnet in rather direct way. The cost of projected disk and memory requirements is too frightening for all but the most well resourced centres. Just as automation of data generation must inevitably drive out small lab based projects, so it seems impossible to imagine that small sites providing database access to a few dozen researchers will survive the turn of he century. This is not the end of EMBnet however, for the organisation can make significant contributions in such areas as user guidance and training. No amount of automation can substitute for creative insight in the design of new software and new approaches to information storage, retrieval and analysis. There is no room for complacency and all users of bioinformatics need to be vigilant when the concepts of infrastructure and central support of widely used resources becomes unfashionable or "unrealistically expensive".

This newsletter is designed to help keep you informed about new developments in bioinformatics. We hope that it does not encourage a passive approach to the problems that so deeply affect us all. Certainly we would like to hear from you if you have achieved an improvement in data collection or evaluation. Sharing information will make us strong enough to meet the challenge successfully.

The embnet.news editorial team:
Alan Bleasby
Reinhard Doelz
Robert Herzog
Andrew Lloyd
Rodrigo Lopez

GCG User Focus Meeting

EMBL Heidelberg, Germany, March 28, 1996

A report by Heikki Lehvaslaiho from EMBnet Finland.

The first ever GCG User Focus Meeting took place last month at Heidelberg. Although plain users were present, most of the participants were involved in maintaining the software, so it could have been called a GCG Manager Focus Meeting. GCG was represented by Maggie Smith-Edelman, Steve Smith and Michael Hogan.

Speakers gave presentations on GCG-related software or on ways to keep sequence databases up-to-date. These talks are listed at the end of this article. The main focus was, of course, on the direction GCG software will take in the next releases.

Proposed developments

1. The most significant development is called SeqLab, a new WPI/GDE-based Graphical User Interface, on which Steve Smith, creator of GDE, has been working for over a year. Most of the functionality of GDE seems to be there. If you have GDE, you can use its helper applications from SeqLab, too. A new thing is that SeqLab is able to display and use features table annotation from sequence files. There will be a printed tutorial on SeqLab in v.9.

2. There will be yet another sequence format to allow the use of features: RSF, rich sequence format. This is a tagged, non-interleaved multiple sequence format, which has only two required tags, name and sequence. The sequence type is DNA by default. Sequences are limited by braces. A GDE-like feature is that the tilde (~) is used for distinguishing trailing spaces from gaps.

Contents

3. The latest 2.xx version of FASTA will be in v.9. The output will be readable by sequence manipulation programs. The fasta format can be used for sequence databases which reduces disk space requirements in sites wanting to have local BLAST databases. Database maintenance in general will be easier (finally!). There will be three main files:

- list of databases (a database can be defined to be on many directories in separate disks)
- dbmap
- one farm file

Farm files used internally by GCG programs will be automatically generated from this one. Also, there will be tools for checking the consistency of database set-up.

4. Improved Fragment Assembly System and David Swofford's PAUP will be in some version after 9.

Licensing

GCG is concerned at the trend for sites continuing to drop the subscription of GCG as users consolidate to use it on fewer and fewer remaining sites. GCG has to find ways to compensate for this loss of income. Several proposals were made and discussed without any definite conclusion. These included an additional 10 % fee for any additional machine, and extra fees for external users such as USD 50/user/year or USD 1500/institution/year.

Whether the number of processors will be an issue was not bought up. GCG propose to have a file within the package where managers will add the computers which use GCG, the implementation will be on an honour system. The situation is still under negotiation and "GCG will work with institutions affected for a workable solution". This will be hard thing to negotiate, especially at sites like mine where all the users are technically "external". We will no doubt hear quite a lot about this subject!

Wish list

The last session, before beer and snacks, was a brain-storming about short term enhancements and future directions of the GCG package. There were a great many requests written on the black board! What did not quite dawn on me until a private discussion with Steve Smith is that they really are seriously thinking of a client/server model. One day, here might be real Mac and PC applications capable of doing simple things on their own and pass hard ones to the server.

Other talks

1. *EMBL Introduction* Hans Doebbeling (EMBL)

2. *Wisconsin Package* as part of EMBL biocomputing environment, Luca Toldo. He discussed also the usage of individual GCG programs. The scripts used (decode_gcg_vms.pl, decode_gcg_unix.pl) are available from Luca Toldo's home page (http://www.embl-heidelberg.de/~toldo/) .

3. *Demonstration of Compugen Biocellerator*
Eli Mintz, Amir Natan, Jonathan Kagan, Compugen, (http://compugen.co.il) Specialised hardware for rapid sequence database searching.

4. *Overview of EMBnet* (http://www.embnet.org/brochure)
Jack Leunissen (Dutch EMBnet node) gave the talk in the absence of Rodrigo Lopez (Norwegian EMBnet Node),

5. *EGCG* (http://biomaster.uio.no/egcg.html)
Peter Rice (http://www.sanger.ac.uk/~pmr/) Sanger Centre, (http://www.sanger.ac.uk/) Extended GCG, free set of programs that improve or add new functions to GCG package.

6. *WWW2GCG*
Marc Colet (http://dec5.ulb.ac.be/marc.html) Belgium EMBnet Node (http://www.be.embnet.org/) Web-interface to GCG with user authentication and graphics

7. *SRS*
Thure Etzold (http://www.embl-heidelberg.de/srsdoc/thure.html) SRS software at: http://www.embl-heidelberg.de:80/srs/srsc. Forthcoming SRS v. 5 has a new, more powerful description and manipulation language, ICARUS, which will allow easier adding of new databases and more flexible queries.

8. *Maintaining up-to-date databases*
Reinhard Schneider (http://www.embl-heidelberg.de/~schneide/) Described the DBupdate which is part of the massive GeneQuiz (http://swift.embl-heidelberg.de/genequiz/) program for automated analysis of protein sequences.
Nicole Redaschi EMBnet Switzerland (http://www.ch.embnet.org/) Described her software package, SynCron, (ftp://ftp.ebi.ac.uk/pub/software/unix/listtools) developed with EBI, for keeping the local copy of the EMBL database updated automatically.

EGCG 8.1 Released

Peter Rice (Sanger Centre, Hinxton, UK)
Rodrigo Lopez (Norwegian EMBnet node)
Reinhard Doelz (Swiss EMBnet node)
Jack Leunissen (Netherlands EMBnet node)

EGCG 8.1 continues the work of the EGCG team in the previous release [1].

The code has been further standardised, and critical parts of the internals are now ported to C so that routines can also be called from C main programs. The documentation has been reviewed, checkedfor omissions and further standardised to the point where we are able to produce an HTML Web version. The URL for the Web version, still under development and changing rapidly, is http://www.sanger.ac.uk/egcg/
New programs in EGCG 8.1 are provided by members of the EGCG team, and also by the French, German and Italian EMBnet nodes and David Mathog at Caltech.

New programs in EGCG 8.1:

POLYDOT - produces all-against-all dotplots (compare - word style) on many sequences. POLYDOT is intended to compare all contigs in a fragment assembly project, but can also compare groups of database entries to find overlaps, or compare protein families. The output is graphical, but POLYDOT also writes a report of overlaps and an input file for GCG's SEGMENTS program which can make the alignments.
PATTERNPLOT - produces a graphical representation of the results of GCG's FindPatterns program.

PROFILEPLOT - produces a graphical report of the frequency of patterns in a protein or nucleotide sequence.

SORTCONSENSUS - identifies the strong consensus regions of an alignment in an MSF file and reports them in sorted order.

STSSEARCH - looks for primer pairs in a set of sequences.

GENETRANS - extracts and/or translates coding regions as defined in the feature table of sequences stored in the EMBL or Genbank databases.

MULTALIGN - does a simultaneous alignment for two or more DNA or protein sequences. The program is based on a generalisation of the algorithm of Waterman, Smith and Beyer by Krueger and Osterburg.

ECLUSTALW, ***CLUSTREE*** and ***PROFALIGN*** - are parts of the original ClustalW distribution from Des Higgins [2], modified for inclusion in EGCG.

WORDUP - Reports unusual (statistically significant) nucleotide patterns of size 6 to 9 bases. The method used is that of Pesole et al [3].

CHAOS - makes a CHAOS game representation of a nucleic acid sequence using the method of Jeffrey [4]. We have used this program to demonstrate patterns down to 5 base resolution in E.coli sequences.
PRIMA - GCG's PRIME program is now extended in EGCG. The only change to date is to allow ranges to be specified relative to the end of the sequence, but many others are planned in the near future.

TANDEM - Looks for tandem repeats of a given size range in nucleotide sequences

QUICKTANDEM - Rapidly scans a nucleotide sequence for potential tandem repeats.

INVERTED - Looks for imperfect inverted repeats in nucleotide sequence.

CPGREPORT - Reports potential CpG islands in nucleotide sequences.

ECOMPOSITION - an extended version of GCG's COMPOSITION which calculates molecular weights for single and double stranded DNA and RNA.

EOVERLAP and ***FILTEROVERLAP*** - an extended version of GCG's Overlap and a quality filter program for use in database non redundancy checks and fragment assembly project validation.

CODFISH - calculates a set of codon usage statistics for a sequence using a specified codon usage table. The name comes from the original requirement for codon usage analysis of fission yeast.

WORDCOUNT - counts the commonest words in a sequence and reports them in order of frequency and sequence.

GAPFRAME - moves all gaps in a DNA sequence reading frame to be at codon boundaries.

PEPCORRUPT - randomly introduces small numbers of substitutions, insertions, and deletions into protein sequence(s).

RFINDPATTERNS - is a version of GCG's FINDPATTERNS that writes each hit to a separate sequence file.

CREFORMAT - a version of GCG's REFORMAT that allows base ranges to be selected or excluded, and some sequence characters to be replaced.

ECODONFREQUENCY .... ETRANSLATE - The remainder of Jaakko Hattula's conversions of additional GCG programs to use the command line (from EGCG 7.x) have been revived as his methods are often different to the GCG used in version 8.0, and in some cases we feel they still can be very useful. These programs also now support the new EGCG interface options (see below).

EFROMFASTA - a version of GCG's FROMFASTA that preserves the case of the output file name.

EPEPTIDESORT - a version of GCG's PEPTIDESORT with additional output options.

ELINEUP - a version of GCG's LINEUP allowing up to 500 sequences with improved row numbering and allowing extended screen sizes.

EPLOTSIMILARITY - a version of GCG's PLOTSIMILARITY with gaps where the sequences are gapped.

IG2NBRF - a utility program that converts an IG formatted file into an NBRF formatted database which GCG's PIRTOGCG can index.

PHYLIP2TREE - displays trees computed with one of the PHYLIP-programs in GCG style.

Enhancements:

An early release of EGCG 8.1 compiled on AIX, although we expect some further problems.

QUICKSEARCH and ***QUICKMATCH*** now support sequences longer than 32000, for example cosmid sequences being compared to the complete database. These programs also have new qualifiers to aid in database self-comparison.

PEPWHEEL and ***PEPNET*** can mark residues in their own style, or in the same style as GCG's HELICALWHEEL program.

New interface details:

All EGCG programs have a new qualifier "-help" which brings up the egenhelp text on that program.

When asking for a sequence range, most EGCG programs can accept "-100" to mean "100 bases from the end". If this is allowed, the prompt is "Start" rather than "Begin".

Most EGCG programs which read sequences are now able to handle sequences in GCG, FASTA, STADEN and TEXT formats by a slight change in syntax :

- **fa:abc.fasta** reads a single FASTA format sequence
- **fdb:xyz.fasta** reads a file with many sequences in FASTA format (including SRS getz sequence output files).
- **fdb:xyz.fasta:LACI** reads the sequence "LACI" from a multiple sequence FASTA file.
- **staden:abd.sdn** reads a sequence in STADEN format
- **text:abc.txt** reads a sequence in plain text format

We expect to extend this syntax rapidly in the coming months. Any suggestions are welcome for new sequence formats.

Not all EGCG programs support this style of sequence specification. Those that do will provide an additional message when prompting for sequences, for example:

TWORDSEARCH uses protein sequence data

TWORDSEARCH of what sequence ?

These programs will additionally check that all sequences are valid, so the program does not need to perform any additional checks (for gaps, ambiguity codes, DNA to RNA conversion, and so on).

Distribution

Current major version: 8.1 beta (March 1996)
URL: ftp://ftp.sanger.ac.uk/pub/pmr/egcg8
E-mail contact : egcg@embnet.org, pmr@sanger.ac.uk

References

[1] Rice P. et al. EGCG 8.0. embnet.news 2(2): 5-7 (1995)
[2] Thompson J.D., Higgins D.G. and Gibson T.J. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22 4673-4680 (1994)
[3] Pesole, G. et al. WORDUP: an efficient algorithm for discovering statistically significant patterns in DNA sequences.Nucleic Acids Res.20, 2871-2875 (1992).
[4] Jeffrey, H.J. Chaos game representation of gene structure. Nucleic Acids Research 18, 2163-2170 (1990).

Coping with the exponential

Andrew T. Lloyd, Irish EMBnet node

A central criteria for membership of EMBnet as a nationally mandated node is a computer large enough to store and make available local copies of the standard databases. Calculations on the back of an envelope suggest that the doubling time for the DNA databases is about 9 months. This means that in 5 years time, the disk space requirements may be 100 times what they are now. The RAM requirements for local blast searches will increase in a similar fashion. Here in EMBnet Ireland, we have a modest system with 96MB of RAM and 20GB of storage which is at the limits of affordability for Irish funders and the limits of usefulness for Irish bioinformatics. Of course, the problem of the exponential increase in database size is not felt only at the Irish EMBnet node, but by all database providers.

The problem is made worse by a number of factors

1. The fact that the rate of increase is increasing, so the growth of the databases is a second order function. This is largely because of automated generation of ESTs which now comprise more than 60% of the database entries.

2. The fact that there are three different DNA databases with near complete redundancy but different formats.

In November 1995, there was a discussion of these issues on the embnet-general newsgroup and nobody dissented from the proposal that there *was* a problem to deal with. As the initiator of that discussion, I here attempt to summarise these ideas to put them on record. The solutions put forward were many and varied, some with large potential savings in diskspace and some with small, some requiring many days or months of work and some requiring an executive decision.

Partial solution list

1. Merge GenBank EMBL and DDBJ into a really identical database with a rational and mutually agreeable format. Saving 50% - everyone talks about DDBJ, but no-one I know in Europe actually consults it. Perhaps this is because DDBJ has the same format as GenBank so it allows (foolish) people to believe that therefore the information content is identical.

2. Remove the need for creating one database by perfecting the procedures by which data is xchanged between the databases, verified and presented. Saving <50% because there will inevitably be some unresolved entries in an exclusion set.

3. Remove internal redundancy by merging overlapping sequences. Saving ??

4. Reduce internal redundancy caused by multiple determinations of the same sequence. Saving ??

5. Reduce database size by storing only differences in very closely related sequences, such as alleles and mutations. Some savings but major coding problems for software developers.

6. Reduce the redundancy caused by storing identical databases in more than 26 locations in Europe. Saving 80%? Up until now, the local database provision criterion for EMBnet membership has been less expensive than the requirement for a full time salary. This is not going to be so for much longer. It must be emphasised that EMBnet nodes would be doing a valuable job even if all their database access provision was carried out remotely by, for example, SRS and e-mail servers. EMBnet, with its history of co-operation is the obvious entity to negotiate this integration.

7. Separate the annotation from the sequences. This will allow EMBL and GenBank to persist in having different annotation and format while referring to the same sequence. The actual sequences make up about <25% of the database, the other 75% being annotation.

a. Recode the DNA with 4bits rather than 8 bits. (MH) Saving 0.5 x 25% = 12%.

b. Separate the taxonomic information from the annotation to allow researchers access to a taxonomic database that they believed was useful. No savings.

c. Separate the references from the annotation, which might allow access to a remote server for this information.

8. Reduce redundancy intrinsic to the taxonomic and other annotation. The full taxonomy of Homo sapiens is represented 400,000 times in the database. Martin Hilbers created the following table to indicate the potential disk space savings created by reducing the redundancy in various EMBL fields

Field	MB	%
OS/OC	64	5%
R*	134	11%
CC	101	8%
Total	399	24%

In addition he pointed out that removing multiple spaces from the annotation, which “only” improve readability, would reduce the database by nearly 15%, while condensing the displayed sequence to fasta format saves an additional 11%.

9. Remove the /translations from the databases and require software developers to deliver these on the fly. Savings 4%

Next steps

There is no one solution: new ideas and new approaches to networking, software development, storage, and algorithms are required. A number of contributors to these discussions have suggested that the database providers, software developers and EMBnet representatives need to get round a table and decide upon a system of priorities. There is no room for complacency and very little time to sit back a see what turns up. On this one we must all hang together or assuredly we will all hang separately.

With thanks to the following people who contributed (and apologies to any whom I failed to archive!): Matteo di Tommaso, Alan Bleasby, Peter Rice, Peter Stoehr, Reinhard Doelz, Robert Herzog, Martin Hilbers, Hans Ullitz-Moeller, Martin Grabner, Rodrigo Lopez, Frank Wright, Patricia Rodriguez-Tome, Marcella Attimonelli, Jose Valverde,

Sandor Pongor, Jan Noordik, Jack Leunissen, Guy Vaysseix and Graham Cameron.

CpGIsle

Human CpG-island release 4.0 is ready!

Lopez R., Prydz H.

The Biotechnology Centre of Oslo, University of Oslo
Gaustadalleen 21, 0317 Oslo, Norway
Tel: +47-22958754, Fax: +47-22694130
e-mail: hans.prydz@biotek.uio.no,
rodrigo.lopez@biotek.uio.no

Reference: Larsen F., Gundersen, G., Lopez R., Prydz H. CpG island as Gene Markers in the Human Genome, *Genomics* 13:1095-1107 (1992)

BREAKDOWN

This release contains 950 genes and 99 pseudogenes described in release 45 of the EMBL nucleotide sequence database. We have assigned the genes transcribed by RNA polymerase II to two groups on the basis of expression. The first group (widespread) consists of housekeeping genes and genes expressed in a wide range of tissues, and the second group (limited) consists of genes with limited or tissue-specific expression.

Expression of gene	Number	Number associated with islands
Widespread	217	216 (99%)
Limited	719	261 (36%)
All RNA pol.II genes	936	477 (50%)
RNA-pol.I+III genes	14	14 (100%)
Pseudogenes	99	22 (22%)

AVAILABILITY ON THE INTERNET

WWW URL: http://www2.no.embnet.org/
FTP: ftp://ftp.no.embnet.org/cpgisle
ftp://ftp.ebi.ac.uk/pub/databases/cpgisle

CpGIsle is also available for browsing using the SRS-WWW system: http://www.no.embnet.org/srs/srsc

INTERviewNet

This month, Alan Bleasby has kindly agreed to interview Martin Bishop at the HGMP Resource Centre. We can be sure that Alan is not lying through his teeth, but can the same be applied to Martin...

Alan: What the hell are you doing down there?

Martin: err you mean "STRATEGIC OBJECTIVES" I think The Computing Group will:

- Continue to provide and improve its on-line service, user support and training for external users.
- Support the systems, networking and software for the HGMP-RC administration, biology and computing sections.
- Participate in collaborative development projects both in-house and with external groups as the major means of ensuring the continuing growth and viability of the Group.
- Participate in research projects.

Alan: No need to be overly formal, we will pick up on a few of those points later. But why do we need the HGMP-RC as the EBI has publicly stated that it will do everything necessary to support Genome Research for the galaxy?

Martin: I have been asked that before so here is one I prepared earlier.

The European Bioinformatics Institute has three main components:

- Services for data and software
- Research (particularly in protein structure)
- Industrial training

The EBI compiles data and makes it available to other centres. The relationship between the EBI and the HGMP-RC is similar to that of a wholesaler and a retailer. Far from competing, the EBI is making its services available to all the EMBnet modes on an equal footing.

In addition, the HGMP-RC provides two kinds of data to the EBI:

- Sequence data from the cDNA and fugu projects
- Radiation hybrid mapping data from the Genebridge Panel.

The HGMP-RC also collaborates with the EBI on training. A number of EBI staff members (Tom Flores, Rob Harper) teach on HGMP-RC training courses. The HGMP-RC provides access to GDB and expertise on GDB to the EBI.

Alan: Well, that sounds moderately convincing. But I was told that the Sanger Centre is much better and that your services are not needed.

Martin: Yes, in a world of limited resources, everyone has got their own model of how the money is best spent. In a nutshell, we are a service organisation and the Sanger is a research organisation. However, our most successful ventures tend to be R & D based e.g.. The EUCIB Mouse Backcross (http://www.hgmp.mrc.ac.uk/MBx/MBxHomepage.html), and recently Fugu Sequencing
(http://fugu.hgmp.mrc.ac.uk/)

The Sanger Centre is concerned with physical mapping to build sequence ready contigs and with large scale sequencing of yeast, nematode and human DNA. There are proposals for collaboration with the Mouse Genome Centre and HGMP-RC on mouse sequencing.

The Sanger Centre has a large software development group and many of the products are used by HGMP-RC. Sanger staff (Richard Durbin and Sylvia Martinelli) run one of the HGMP-RC training courses (Acedb Workshop). Peter Rice talks about EGCG on our GCG course and a number of other people participated in out Gene Identification Course. Sanger Centre staff have attended HGMP-RC training courses.

Alan: Well it sounds like you are one big happy family down at Hinxton. So why don't you run SEQNET?

Martin: I am glad you asked me that :-) First of all, we don't want to and second of all, no one asked us to.

We are the Human Genome Mapping Project Resource Centre and are specialised in providing biological materials and computing services for the Human Genome Project. We provide a lot of human genetic linkage analysis programs and mirror sites for GDB and MGD which SEQNET do not.

On the other hand there is bound to be some overlap. God appears to have been rather parsimonious in his choice of proteins to perform the genetic functions of living organisms!

The UK national node of EMBnet is SEQNET at the Daresbury Laboratory funded by the BBSRC and the EPSRC. SEQNET is smaller in terms of staff and resources than HGMP-RC. It tends to be used for general molecular sequence and structure analysis by people from universities which lack their own facilities for this purpose. Many universities (notably London, Cambridge, Oxford and Edinburgh) and Research Council Institutes do provide facilities and support staff in this area. There is no recent survey to evaluate the present position with any accuracy and no coherent national policy. But people do use SEQNET because it is easy for them to get what they need in the area of sequence and structure data and analysis. Long may it flourish.

Alan: How far are you from getting the complete genome?
Martin: You mean them not us. For genetic linkage there is a good map based on microsatellites at a resolution of at least 5 cM (which is about 5 Mb). Flavour of the year is radiation hybrid mapping for which you do not need polymorphic markers. It will provide a sound framework. The YAC story is not so happy. We do not have a decent overlapping clone map for the genome yet, except for small
regions. I think BACs will be used to do it as YACs partially delete and form chimaeras all over the place. In conjunction with cosmids the sequence ready clones will be made available.

In terms of the sequence, we may have seen 50 % of the genes already, if you believe in ESTs. How they know what is in them darn 'cDNA libraries' beats me (genomic contamination .. mutter, mutter).

I expect the genes of the human genome to have been adequately characterised by 2002, but not necessarily by the route you might imagine (brute force sequencing).

Alan: What software do you use for collating the data?

Martin: That is a complex question. Please access http://www.hgmp.mrc.ac.uk/ for a partial answer, or send me a ream of A4 paper for the printed version.

Alan: Whom do you collaborate with in this 'project'?

Martin: Just about anyone who is interested. We are an enabling technology centre, rather than doing large scale work ourselves. We have both academic and commercial partners and this side of our operation is where the major growth is at present.

Alan: Can you give us an estimate of when the project will finish?

Martin: If the sequences of all the genes are available in 2002, it will probably take 50 to 100 years to elucidate the structure, functions and pathways relating about 60 000 genes. I am happy to have lived in a period when this is possible but I will be dead before the project is finished. The Human Genome Project which most people talk about is the beginning not the end. The comparative method will be essential, we will be comparing all the genome we can.

Alan: What contact do you have with other genome mapping projects? If so, how do you co-operate
.... do you use similar software?
.... what software do you use?
.... is the software applicable to all the genome projects?

Other vertebrates are the obvious one's - pig, cow, sheep, dog, cat (nice pussy Steve), horse, mouse, fugu. Sadly, the only projects using the same software are pig, chicken and

sheep (based on GBASE and developed at Roslin) http://www.ri.bbsrc.ac.uk/

Everyone, but everyone, uses their own system - man, mouse, fly etc. The only saving grace is acedb which has had about everything under the sun stuffed into it at one time or another with varying degrees of success. Unfortunately, not all genome researchers have understood that you can decouple data storage and maintenance, data update, and user query and display tools. Things are just starting to be built in a sensible, modular, reusable form. But the genomes are different e.g.. polytene chromosomes, hexaploid wheat. Now cereals, that's an interesting story.

Alan: Have you given up mud wrestling yet?

Martin: Fango niente. But in 1995 I paid 50 ECUs for two heaps of shit, unrotted horse (15 ECU) and well rotted bullock (35 ECU).

TIPS from the computer room

Secure e-mail
Rodrigo Lopez, EMBnet Norway

Many find using Email for 'serious purposes' (i.e. user defined 'serious') a problem today. It is relatively simple to falsify, steal and read someone else's Email. Any knowledgeable system administrator can tell you how easy this is and may even tell you some stories of how his users have been affected, either on purpose or by accident, by Email security leaks (i.e. badly set up mailer daemons, badly configured sendmail scripts, etc.).

There are a few techniques which can make Email more secure. The most well known and perhaps the most efficient are based in the so called 'encryption techniques'. The basic idea behind these is that the sender 'signs' his/her messages with a special code (i.e. an encryption key), that contains information on the sender and special character sequences that can be used to check if themessage has been changed in any way in its path across the networks. Another way to protect Email messages is to encrypt the entire message. The message is passed through a special encoder. Only the sender and the recipient have the key (i.e. the code sequence) required to unscramble it. This implies making this key available to those recipients to whom one wishes to send 'secure' mail.

There are indeed two levels of encryption one must know about: Private key-based and public-key encryption. For the public encryption scheme it isn't necessary to know the key to encrypt a letter. Though the opposite is true it one wishes to decrypt it.

Secure Email requires good programs that deal with the actual encryption and signature process. The most known today are PGP (Pretty Good Privacy) and programs based on PEM(Privacy Enhanced electronic Mail) standard. Both work in the same way in that the system ensures that only the intended recipients can read an encrypted message and can trust that the sender that has signed the message is for real. Worth taking notice off is the fact that it is impossible to 'translate' encrypted Email from one system to another without having access to the encryption key.

In PEM, sending Email is designed around a 'trust model' which implies that both sender and recipient are capable of 'keeping a secret'(i.e. the key). Indeed, this 'human attribute', which has defied programmers (and hopefully will continue to baffle them) is the backbone of these systems. One method used to emulate some 'trust' in Email systems is to maintain a 'certifying authority'. This authority makes sure that signatures in secure Emails are from users who really intended to send them. This centralised model of trust provides a pretty high degree of security but, as with anything else, depends on honest and secure management.

In PGP the trust model is also employed but at another level. The idea here is that everybody certifies everybody. In small groups of senders and recipients this works well but once these increase in number one sees the requirement to centralise (i.e. use PEM) instead.

As driving a car, the responsibility and security depend on the drivers and on how well they master their vehicles. The same is true for Email (and anything that has to do with computer networks!). Remember there are ways to steal both keys and Email if the local computer resources are not sufficiently well administered but the ultimate responsibility when sending 'serious' messages across the networks is the users as well as making sure one is well informed and pays attention to system management announcements.

Books and Links

The attention of the readers of embnet.news is drawn to a new book entitled Internet for the Molecular Biologist (Eds. Swindell et al) which was published January 1996. I feel it is an excellent introduction to the internet for the molecular biologist (covers WWW, email, ftp, gopher, ewsgroups, databases, specialist sites etc.). One of the book's strengths is that it was written and published in about eight months so it is as up-to-date as any book can be. In addition, there is a web site:

- (http://www.ccc.nottingham.ac.uk/~mbzsrs/IFTMB.HTML) dedicated to users of the book which

regularly expands and updates the information in the book. So readers of the book get continuous updates at no extra cost! More information on the book can be found at the site:

- http://www.ifrn.bbsrc.ac.uk/gm/lab/docs/iftmb.html

Some other web sites that you might like to draw attention to include:

- http://www.ifrn.bbsrc.ac.uk/gm/lab/docs/molbiol.html
  Molecular Biology Jump Station: A comprehensive collection of USEFUL links for the molecular biologist.

- http://www.ifrn.bbsrc.ac.uk/gm/lab/docs/iftmb.html
  Internet for the Molecular Biologist: A good introduction to the internet for the molecular biologist.

- http://www.apollo.co.uk/a/horizon
  Books for Molecular Biology. Includes full chapter abstracts, book reviews, and ordering information. Check out the new book *Internet for the Molecular Biologist.*

- http://www.ifrn.bbsrc.ac.uk/gm/lab/docs/protocols.html
  Protocols on the WWW. An extensive directory of the sites providing protocols for molecular biology, microbiology, genetics, biochemistry, and cell biology.

- http://www.ifrn.bbsrc.ac.uk/gm/lab/docs/micro.html
  Microbiology Jump Station: A site for microbiology containing links to microbiology institutes, organisations, journals, culture collections, newsgroups, protocols, information, and directories.

- http://www.ifrn.bbsrc.ac.uk/gm/lab/docs/genetics.html
  Genetics Jump Station: A large collection of links for the geneticist.

NODE NEWS

The Austrian EMBnet node

SOFTWARE

Several programs from Oxford Molecular Ltd. are available at the Vienna University now. The Vienna University Computer Center offers within the scope of a campus license the following packages:

- AbM 2.03 (Immunoglobulin Domain Modelling Program)
- Amber 4.2 (Assisted Model Building with Energy Refinement)
- Anaconda 2.01 (Interactive Molecular Surface Comparison)
- Asp 3.11a (Automated Similarity Package)
- Cameleon 3.13 (Sequence Analysis Program)
- Cobra 3.21 (Conformational Analysis System)
- Iditis 3.0 (The Relational Database of Protein Structure)
- Iditis Architect 1.01 (Iditis Data Derivation Suite)
- Tsar 2.4 (Tools for Structure-Activity Relationships)
- Vamp 5.51 (Semiempirical Molecular Orbital Package)

The Operating System of the EMBnet Server was upgraded to DGUX V3.2D-2.

Furthermore the PROCHECK Suite from M.W. MacArthur, R. A. Laskowski, D. S. Moss, J.A.C. Rullmann & J. M. Thornton (J. Appl. Cryst., 26, 283-291) was added to the programs for protein structure analysis.

DATABASES

In 1996 the following databases were integrated into SRS at our site.

- NRSUB a non redundant sequence database of the bacterium Bacillus subtilis in EMBL format.
- TFSITES an index of sites.dat from TRANSFAC
- TFFACTOR an index of factor.dat from TRANSFAC
- TFMATRIX an index of matrix.dat from TRANSFAC
- YPD a reformated version of the Yeast Protein Database

LAN and HARDWARE

Several of the Netware-Server at the Vienna Biocenter were upgraded. The backup system for all Novell etware-Server was switched to Palindrome software from Seagate Software Company and a SureStoreTape 12000e tape drive from Hewlett Packard.

NETWORK

Now the Connection line from Vienna to Graz runs with ATM and a bandwith of 2Mbit. The maximum bandwith for current tests is 4Mbit.

The Swiss EMBnet node

[1] Staff changes
Dr. Nicole Redaschi has been appointed to run the Swiss EMBnet node as a successor of Dr. Reinhard Doelz, who is

[2] Research Report available
The Biozentrum Basel/Switzerland is pleased to announce the availability of its biennial Research Report (covering the years 1993-1995) on the WWW

http://www.ch.embnet.org/biozentrum/

The WWW documentation includes travel information and a list of all people currently working at the Biozentrum. The curriculum and other student information is available in English and German language. All departments are listed with a brief description of each research group and relevant recent publication references. Colleagues who wish to get more information are welcome, as well as students or candidates for pre- or postdoctoral positions, to browse the available descriptions and get more information on this exciting place of research.

The Finnish EMBnet node: CSC

Sequence database reorganisation
http://www.csc.fi/molbio/bionews/1996/8.1.html#REORG
Following the example of major sites, nucleotide databases have been reorganised to exclude EST and STS sequences from default similarity searches within GCG package.

Genetic linkage programs
http://www.csc.fi/molbio/progs/#LINKAGE
A new branch of computational biology is now supported. While the core of the genetic linkage service is the old LINKAGE/FASTLINK combo (thanks to Weiyun Chen at DKFZ for help), the emphasis is on new generation of faster tools: VITESSE, MAPMAKER/HOMOZ and MAPMAKER/SIBS.

The GCG Batch option

The batch option (-BATch) available in many computing intensive GCG programs now uses 'real' batch system NQE rather than putting the program to background. Additionally the load on our two multiprosessor SGI's is checked and the job is automatically submitted to the one with less load and hopefully better performance.

New program: PDBmotif
http://www.csc.fi/molbio/progs/pdbmotif/
Marks PROSITE patterns in PDB files for display in Rasmol.

The Norwegian EMBnet node

Linda Akselberg has taken over duties as manager and administrator of the National EMBnet node in Norway. She replaces Rodrigo Lopez who will be taking a position with the systems groups at the European Bioinformatics Institute.

MPSRCH email server in Bergen, Norway

An experimental email server for biological sequence database searches is now available at the Department of Informatics, University of Bergen, Norway, at the address:

mpsrch@ii.uib.no

through the collaboration of the Biocomputing Research Unit, University of Edinburgh, the Department of Informatics, University of Bergen and the Norwegian EMBnet node in Oslo.

This server gives you access to MPSRCH Version 2.1A, a suite of database search programs for the massively parallel MasPar computer, to aid the rapid identification of novel biological sequence and functional relationships for new sequences. The programs were written by J.F.Collins in the University of Edinburgh, UK, and, when run on the Bergen MasPar's 16384 processors, are the world's fastest implementations of these exhaustive search programs.

MPSRCH uses the Smith-Waterman local similarity algorithm, and performs an exhaustive search of the whole database against the whole of the query sequence.

MPSRCH can be used for both protein as well as nucleic acid searches, backtranslated protein query against nucleic acids database, and nucleic acids query against backtranslated protein database.
The programs we are offering for the time being are:

MPsrch_pp	protein query against protein database
MPsrch_ppa	protein query against protein database, affine gap costs
MPsrch_nn	nucleic acids query against nucleic acids database
MPsrch_nna	nucleic acids query against nucleic acids database, affine gap costs

The programs offer a wide variety of parameters to control the search. All of these have sensible default values, set in an intelligent manner based on long term experience. Therefore, all the user is required to specify for an initial search is the name of the program, and the query sequence. All parameter settings, either defaults or user-specified, are reported back to the user to provide an idea of the nature of the search, and what else might be done.

The databases that are currently available are Swissprot (protein database) and EMBL (nucleic acids database), and we always keep the most recent releases. The daily updates of EMBL and the weekly updates of Swissprot are available as well.

There is an extensive help information system to support the users of this email server. For a start, send a message containing only the word "help" to obtain general information about the mpsrch service. This information system and more about the service is available on the WWW at http://www.ii.uib.no/~linda/bio/mpsrch/mpsrch.html.

For support and advice from a human being, contact mpsrch-support@ii.uib.no.

This service is running on a computer dedicated to research activities, and might on rare occasions be more heavily loaded than desirable. In such cases, please have patience. The service is experimental in that it is now on trial and will be evaluated after some time.

Linda Akselberg
Dept. of Informatics
University of Bergen
Norway
email: linda@ii.uib.no
phone: +47 55 54 40 36

John F. Collins
Biocomputing Research Unit,
Edinburgh University,
Scotland, UK.
jfc@biocomp.ed.ac.uk
+44 131 650 5365

The Hoffmann-LaRoche EMBnet node.

This is the year that Roche celebrates its 'first hundred years' and make many changes. Not only a new (albeit, only slightly modified) logo but also a new postal code! And a new brochure page.

Conferences and Announcements

1. ANGIS advanced user workshop series

July 1996 Call for expression of interest

ANGIS is considering running a series of 'advanced user' workshops in July covering specific topics in detail. This series of workshops would run over a week, from July 15 to 19, and participants could either attend specific workshops, or attend all of them as a course. Each workshop would be given by an expert in a particular field, and would have both lecture and practical components.

This advanced-user course is to be directed at reasonably experienced ANGIS users. For occasional or less experienced users, the one-week ANGIS practical course to be given a fortnight previously (July 1 to 5) is recommended.

The course will contain between 5 and 10 workshops (half a day to a full day each). The topics of the workshops will depend on the expertise of available lecturers and on the interests expressed by ANGIS users. Some potential topics are phylogeny inference, large sequencing project management and fragment assembly, and linkage analysis.

We are interested in hearing from:

- Potential attendees for this workshop series: whether they would be interested in attending such an advanced user course, and what workshop topics they would be interested in.

- Experts in 'biocomputing' who would be interested in teaching a workshop on how to use particular programs and techniques on ANGIS. The emphasis of the workshops should be on biology, and on the use of the programs to obtain solutions to biological problems. The workshops will also in provide enough technical information about the various methods used to explain the limitations and benefits of various approaches.

If you are interested in participating in this course as a student or faculty member, please contact me by email at gaeta@angis.su.oz.au or by phone on (02) 351 7221

2. Molecular Biology of Picornaviruses

A committee from the Institute for Biochemistry organises the IXth Meeting of the European Study Group on the Molecular Biology of Picornaviruses. The congress is held in Gmunden's Toscana Congress Centre from 18-24 May. (Secretariat: Vienna Biocenter, Institute of Biochemistry, University of Vienna, Dr. Bohr-Gasse 9, A-1030, Vienna, Austria, FAX:+43-1-7986224)

3. EMBnet/CNR training course

The Training Course in Molecular Evolution funded by the EMBnet Education and Training Committee and CNR, Italy, will be held in Bari (Italy) on October 2-5, 1996.
Detailed information and the program are available at the following site

http://www.ba.cnr.it/meeting/meeting.htm

4. Mathematical Analysis of Biological Sequences Workshop

Trondheim, August 4-6, 1996.

Please consult our web page: http://www.imf.unit.no/conferences/mabs/

5. GCB'96
COMPUTER SCIENCE AND BIOLOGY
German Conference on BIOINFORMATICS

Leipzig, September 30th - October 2nd, 1996

The German Conference on Bioinformatics is organised on behalf of FG 4.0.2 Informatik in den Biowissenschaften of the German Society of Computer Science (GI) in cooperation with the AG of Computereinsatz in den Biowissenschaften of the German Society of Chemical Technique and Biotechnology (DECHEMA) and the AG Mathematische Modelle in Biologie und Medizin of the German Society of Medical Informatics, Statistics and Epidemiology (GMDS). The conference will take place at the University of Leipzig, September 30th - October 2nd, 1996, and is intended to bring together scientists who are addressing problems in

biosciences and medicine using advanced computational methods including data modelling, simulation, artificial intelligence, computer graphics (visualisation), robotics, combinatorial and stochastic optimisation. The conference is concerned with all aspects combining computer science and biosciences. Topics of particular interest include, but are not limited to:

- Genome Analysis
- Models of Gene Regulation
- Formal Languages and DNA
- Molecular Docking and Recognition
- Molecular Modeling and Protein Design
- Models of Pattern and Structure Formation
- Models in Cell Biology
- Models of Dynamic Biological Systems
- Self-Organization and Complex Systems
- Metabolic Engineering
- Metabolic Pathways
- High Performance Computing
- Biological Database Technology
- Visualisation and Animation of Biological Processes
- Artificial Intelligence and Complex Systems
- Evolutionary Computing
- DNA Computing
- Biological Paradigms in Computer Science

Proceedings

Extended abstracts of all accepted presentations including posters and computer demos will be published in the conference abstract book which will be distributed to all participants. It is planned that full length papers will be invited by the PC for publication by Springer-Verlag "Lecture Notes of Computer Science" after the conference.

Submission Procedures

The closing date for receiving papers (extended abstracts of 3 pages), posters and computer demos (abstracts of 1 page) is May 1, 1996. The conference offers the possibility for the presentation of tutorials (abstracts of 1 page). The decision on acceptance for presentation will be communicated by June 15, 1996.

Authors are urged to specify the category to which they are submitting their paper. Submissions must be written in English and should include title, author's name, mailing address, telephone number, fax number, email address and a list of keywords. They should be sent to:

PD Dr. R. Hofestaedt Tel. 0341 / 9716100
Prof. Dr. M. Loeffler Fax 0341 / 9716109
University Leipzig
Department of Medical Informatic, Statistics and Epidemiology, Liebigstrasse 27, 04103 Leipzig, Germany
email: GCB96@imise.uni-leipzig.de

Organising Committee

R. Hofestaedt (Leipzig & Koblenz, GI-FG, Germany)
T. Lengauer (Bonn, GI-FG, Germany)
M. Loeffler (Leipzig, GMDS, Germany)
D. Schomburg (Braunschweig, DECHEMA, Germany)

Program Committee

J. Collado-Vides (Mexico)
A. Danchin (France)
A. Dress (Germany)
P. Karp (USA)
H. Kubinyi (Germany)
H. Lim (USA)
M. Mavrovouniotis (USA)
M. Mewes (Germany)
J. Shavlik (USA)
S. Suhai (Germany)
M. Vingron (Germany)
E. Wingender (Germany)
H. Zima (Austria)

Important Dates

Deadline of Submission:	May 1, 1996
Notification of Acceptance:	June 15, 1996
Receipt of Camera Ready Manuscript	August 1, 1996

Registration Fees

Full participation	250 DM
GI, GMDS, DECHEMA	200 DM
Student	80 DM
One day registration	90 DM

Includes a copy of the proceedings, tea/coffee at the conference and food and drink at the poster session.

Registration form (to send preferably by email) available from:

GCB'96
Universitaet Leipzig
Institut fuer Medizinische Informatik und Statistik
Liebigstrasse 27
D 04103 Leipzig Germany
Phone 0341 / 9716100
Fax 0341 / 9716109
email: GCB96@imise.uni-leipzig.de

If you need an earlier decision on your submitted paper, so you can make your plans, please let us know. We could provide an early decision, although the partition of accepted abstracts into oral presentation and posters will take place only after June 15. For early acceptance, papers had to be submitted by April 20.

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 (rodrigol@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_1/contents.html***
- ***http://www.be.embnet.org/embnet.news/vol3_1/contents.html***
- ***http://www.ch.embnet.org/embnet.news/vol3_1/contents.html***
- ***http://www.no.embnet.org/embnet.news/vol3_1/contents.html***
- ***http://www.ie.embnet.org/embnet.news/vol3_1/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/***
- ***ftp.ch.embnet.org in the directory pub/embnet.news/***

Back issues.

Online:
- ***http://www.ch.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/***
- ***ftp.ch.embnet.org in the directory pub/embnet.news/***

Publisher:

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

embnet.news
Vol.3, No.1, 1996
10May 1996
ISSN 1023-4144