embnet.news

Volume 2 Nr.3 (ISSN1023-4144)

November 1st, 1995

What a summer ! While the lawns outside were getting parched and the sun was baking the heads of normal people, the embnet.news editorial team were boiling away indoors designing the new EMBnet brochure. It is now finished and available both printed and over the Web. This is our excuse for the delay in producing this issue of embnet.news. Nevertheless you are, as always welcome.

This issue has a brief report on the EMBnet AGM held this year in Madrid. There have been significant changes in the government of EMBnet as well as the addition of a new node. The next AGM is cheduled for Novemeber 1996 in Finland. For bioinformatics theory we have a report on ORF: a program from the Spanish EMBnet node for finding the minimal phylogenetically informative regions in multiple sequence alignments. We also publish a description of MPSRCH: a new method for accessing a maspar sequence similarity server based in EMBnet Switzerland.

The node focus this issue is the Vienna Biocenter which is the Austrian EMBnet node run by Martin Grabner. For the INTERviewNET Rodrigo Lopez talks to Alan Bleasby, the author of OWL, SWEEP, DELPHOS, SMITE, MOWSE, and various other programs and databases he'll disclaim ownership of !
In addition we have tips from the computer room, node news, and all the regular features.

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

Contents

Editorial	1
The EMBnet Annual General Meeting, Madrid	1
SOFTWARE DEVELOPMENT [1] Finding optimal regions for phylogenetic analysis	2
INTERviewNET - Alan Bleasby - SEQNET - UK	3
Tips from the computer room - Indexing GenBank	5
SOFTWARE DEVELOPMENT [2] MPSRCH Rapid sequence similarity searching on massively parallel computers using the HASSLE protocol	6
Node focus - Vienna	7
Node News	8
The EMBnet nodes	10
embnet.news information	11

The EMBnet AGM in Madrid

September 30th - October 1st 1995

The annual general meeting (AGM) of EMBnet took place in Madrid. A scientific meeting was hosted by the Spanish node, "Centro Nacional de Biotecnologia, Universidad Autonoma de Madrid" with support from the "Consejo Nacional de Ciencias". This meeting provided participants as well as the EMBnet oard with unique insight into the "Red Nacional de Bioinformatica", its activities past and present as well as goals for the future. From this meeting it was easy to conclude that there is a strong interest in bioinformatics in Spain which has international experts in fields such as: the use of neural networks for gene prediction, biological data imaging, development of search algorithms for massively parallel computers, just to mention a few.

During the EMBnet AGM some important decisions were made. One of these is the definition of EMBnet's geographical boundaries which are (according to the motion): "'Europe' be limited to Northern Mediterranean countries up to and including Israel, and those West of the Urals". Another decision of importance to the present structure of EMBnet was the addition of an extra member to all Project Committees, which now have four members each. The Project Committees cover Reasearch and Development, Education and Training and Publicity and Public Relations.

Dr. Jan Noordik from the Dutch node CAOS/CAMM, who has been EMBnet's Chairman for the past 4 years stepped down. EMBnet gratefully acknowledges his work and personal efforts in helping mould EMBnet into what it is today. A new Chairman was appointed: Dr. Sandor Pongor from the ICEGB node in Trieste.

EMBnet also has a new node: The Sanger Centre in Hinxton Hall, Cambridge, UK, has joined in and Peter Rice (the coordinator of EGCG - see issue 2.2 of embnet.news) is the node representative.

The new EMBnet brochure was presented to the board during the meeting. The new brochure is a reflection of the organisation and its members. Please contact the EMBnet Stichting office if you wish to receive a copy (see the end of this publication).

SOFTWARE DEVELOPMENT [1]

Finding optimal regions for phylogenetic analysis in multiple alignments of sequences.

María Jesús Martín and Joaquín Dopazo
Spanish EMBnet node

INTRODUCTION

This article presents a heuristic procedure for the detection, within a multiple sequence alignment, of the shortest region whose informational content can still render a phylogenetic reconstruction identical to the one obtained from complete sequences. In recent years, the availability of fast and accurate sequencing procedures along with the use of PCR has lead to a proliferation of studies of variability at the molecular level in populations. Since population studies require the analysis of many individuals, it is often impracticable to examine at the same time many long genomic stretches. In general, the proposed procedure can save costs and effort in any study involving a large number of sequences.

METHOD

The method is based on the comparison of the pairwise genetic distances obtained from a set of reference sequences to those obtained using different windows of variable size and position. The minimum sequence length required for phylogenetic analysis (Martín et al., 1990) has been studied only from a theoretical point of view (Tajima, 1991) and for some tree-reconstruction procedures (Churchill et al., 1992; Zharkikh and Li, 1992). Except for these few attempts, there is no simple procedure for the selection of the optimal length and location of the regions to be analysed. This does not pose a problem when whole genomes (such as those of mitochondria, plastids or small RNA viruses), or specific genomic regions of particular interest are chosen for analysis. However, when large scale molecular analyses are to be undertaken, the optimal allocation of resources should balance the length of the sequence and the number of individuals to be analysed. This is especially true for the case of molecular epidemiology of viruses, where the use of standardised systems for routine sequencing is often necessary. In practice, these short genomic fragments are not always chosen on the basis of very rigorous criteria. For example, the location within the alignment of the fragment to be studied seems to be a key factor that, strikingly, has not been the subject of corresponding studies.

Providing that a set of sequences is assumed to represent the evolution of the organisms under study, the method is designed to find the shortest window (or windows) which in turn can be considered an accurate enough phylogenetic representation of the complete sequences. This is achieved as follows: we start with a given window size (usually 100 nucleotides) and the entries of the distance matrix obtained from the full-length sequence are compared, by means of a simple ChiSq test, to the entries of the distance matrices corresponding to the window length, starting at the first site of the alignment. Since the entries of the matrices cannot be considered to be independent, this is not exactly a ChiSq test and we will call it X2. This window is moved along the entire length of the sequence alignment and a profile of X2 values is plotted. If no values are found below a given cutoff value, the size of the window is increased and a new profile is obtained. The cutoff value can be obtained by means of a computer simulation that relates the X2 values to the probability of finding the true tree (see Martín et al. 1995 for details). The procedure is stopped when stretches with values below the cutoff are obtained for a given window size.

An example with the P1 region of the foot-and-mouth disease virus.

P1 region of the picornavirus (see Rodrigo and Dopazo, 1995) foot-and-mouth disease virus (FMDV) comprises, approximately, one third of the genome of FMDV and codes for the four capsid proteins of the virus. Phylogenies obtained from this protein can be considered as a good representation .of the evolutionary relationships among FMDV samples (Martínez et al., 1992). We have applied the proposed test to the set of FMDV P1 sequences (Martínez et al., 1992). The results obtained have allowed us to define the regions displaying the highest informative content along this region.

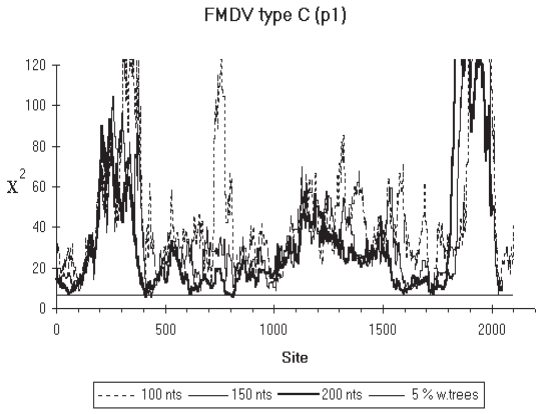

Figure 1. Comparison of local windows to the whole P1 region of FMDV type C. X2 values obtained for each local window starting at the indicated site and spanning 100 (dashed line), 150 (solid thin line) and 200 (solid thick line) nucleotides downstream are shown. Horizontal line represent the cut-off values corresponding to 5% of probability of obtaining a wrong tree

Figure 1 shows the X2 values obtained for window sizes of 100, 150 and 200 nucleotides along the complete VP1 gene. It is worth noticing that one stretch of 150 nucleotides, as well as several stretches of 200 nucleotides have informative contents high enough to render the same phylogeny than that of the complete P1 region, whose size is 2199 nucleotides. The reduction in the amount of sequence necessary for phylogenetic analysis purposes is of one order of magnitude, which is not negligible.

The purpose of the method is to locate those regions showing an adequate level of variability within the set of sequences under analysis. In the example shown in Figure 2, the relationship observed between the X2 index and the values

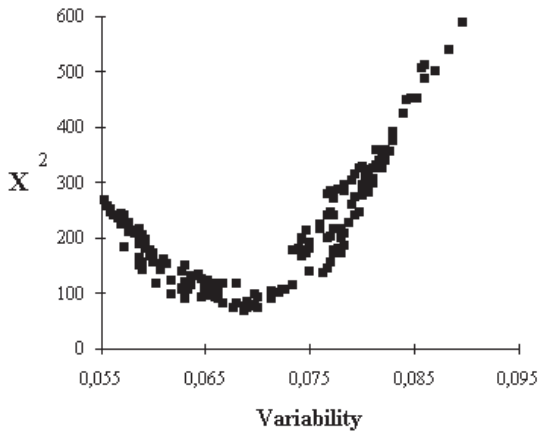

Figure 2. Relationship between the X2 index and the values of variability. The plot was obtained using all the possible overlapping windows of 250 nucleotides long along the VP1 protein gene of type C sequences of FMDV. The variability was calculated as an average of the number of differences per position with respect to a consensus sequence. A minimum value of the plot was found for the adequate level of variability for the set of sequences analyzed.

of variability obtained for all the possible overlapping windows of 250 nucleotides long, along the VP1 protein gene of type C sequences of FMDV, can be seen. The variability value is obtained as the average over all the positions considered (250 in this case) of the proportion of differences from a consensus sequence.

Windows displaying low variability showed high X2 values due to their poor information content. On the other hand, windows having a high variability display also had high X2 values. This is probably because they contain enough noise to disturb the similarities between the total and local matrices. A minimum is found at an adequate level of variability for the set of sequences analysed.

DISCUSSION

The method developed for the selection of the most informative stretches in a sequence can be very useful in molecular screening studies. By monitoring only a representative portion of the loci of interest, instead of studying the whole sequence, a minimal amount of sequencing becomes necessary.

There is classical rule of thumb for the selection of the optimal regions for phylogenetic analysis. If the sequences to analyse show a close relationship, the selected regions must be those showing a high level of variability in which the few nucleotide changes are concentrated. On the other hand, if one is dealing with distantly related sequences, the selected regions should correspond to more conserved stretches, in which the informative changes are situated, discarding other more variable regions in which the mutations mainly constitute evolutionary "noise". The procedure described here provides a quantitative criterion for the application of this rule. Figure 2 shows a clear example of how an adequate level of variability is selected for the set of sequences analysed.

The method proposed finds the best region for sequence analysis on the basis of the set of sequences provided. This leaves the user with the responsibility of choosing the appropriate set of sequences, representative of the expected range of variability. This has the advantage that the selected regions will be those showing the adequate level of variability for the sample used. Indeed, it is important to realize that informative regions are different depending on the degree of variability of the sample analyzed. The method will find the most appropriate regions for each case in a similar way to the rule of thumb described earlier.

The predictions of the method are very consistent provided that the degree of variability of the sequences is similar to the one shown by the reference set and the constraints for fixation of mutations along the sequence do not differ very much (See Martín et al., 1995). A computer program (ORF.EXE), which implements the method, is available from our FTP server (ftp.cnb.uam.es) in the directory / software/molevol. You are also kindly invited to have a look at our software Web page at the URL http://www.cnb.uam.es/www/programas/pag- soft.html, where several pages dedicated to this method can be found.

REFERENCES

Churchill G.A., von Haeseler A., and Navidi W.C. (1992) Sample size for a phylogenetic inference. *Molecular Biology Evolution* **9**:753-765.

Martín M.J., Gonzalez-Candelas F., Sobrino F. and Dopazo J. (1995) A method for determining the position and size of

optimal sequence regions for phylogenetic analysis. *Journal of Molecular Evolution* In press.

Martin A.P., Kessing B.D., Palumbi S.R. (1990) Accuracy of estimating genetic distances between species from short sequences of mitochondrial DNA. *Molecular Biology Evolution* **7**:485-488

Martínez M.A., Dopazo J., Hernández J., Mateu M.G., Sobrino F., Domingo E., Knowles N.J. (1992) Evolution of the capsid protein genes of foot-and-mouth disease virus: antigenic variation without accumulation of nucleotide substitutions over six decades. *Journal of Virology* **66**:3557-3565.

Rodrigo M.J., Dopazo J. (1995) Evolutionary analysis of picornavirus family. *Journal of Molecular Evololution* 40:362-371.

Tajima F. (1991) Determination of window size for analyzing DNA sequences. *Journal of Molecular Evolution* **33**:470-473.

Zharkikh A, Li W-H (1992) Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. *Molecular Biology Evolution* **9**:1119-1147.

INTERviewNET

Rodrigo Lopez interviews Alan Bleasby SEQNET, EMBnet, Daresbury Lab, UK))

RL: Hi Alan. Could you tell us a bit about what SEQNET is, who maintains and supports it and who uses it?

AB: I've been asking myself the same questions for 8 years! SEQNET (sequence network) provides national bioinformatics services throughout the UK. We have approximately 3000 users covering Universities, hospitals and charitable institutions. We are funded by a governmental research council called the BBSRC which we like to call the Biology, Beer and Spirits Research Council. There are two molecular biologists in SEQNET, including myself. We are fortunate to be supported by four people on a help desk, four people in network support, three people in network development and two people in the systems group. Our system is a distributed UNIX one including SGIs, Alphas, MasPar and a Bioccelerator. We've got 50Gb of disc holding all the usual bio-stuff.

RL: There is obviously a lot of virtual (computerwise) biology going on at Daresbury. What kind of things do you all do?

AB: Well we do have the SRS. I don't mean the Sequence Retrieval System which any self-respecting bioinformatics site will have installed from EMBnet nodes; I'm referring to the Synchrotron Radiation Source. People throughout the world wander over here clutching their protein crystals, blast them to bits with high energy X-rays and then join the dots. We have a strong crystallography group here. Also, owing to the SRS we have large data acquistion teams and electronic groups. This resource allows other means of studying biology; for example non-crystalline diffraction from DNA, study of actin/myosin etc. Other biology occurs on the computing front from our Theory and Computational Science group. Everytime anyone blinks there is a new parallel computer in the corridor!

RL: What other services does SEQNET provide to the community?

AB: Well we do have dl.seqnet.jokes but that's only local. We provide email and NEWS services for EMBnet and we are the BIONET/BIOSCI site for Europe, Africa and Asia. We collaborate with our USA colleagues who deal with the Americas and Pacific Rim. It involves a lot of late night email owing to the time differential.

RL: OK. Lets talk about the biological database field. Do you have any projects running that create or exploit, for example, protein sequence information?

AB: Well, we are heavily involved with the OWL (its a hoot) non-redundant protein sequence database and the PRINTS protein fingerprint database. You'll find these on EMBnet anonymous ftp servers (e.g. s-ind2.dl.ac.uk) and at the NCBI. We are also involved with molecular weight fingerprinting of proteins using mass-spec data (e.g. mowse@dl.ac.uk ... send the word help in the message body). Recent avenues of interest include extending OWL and producing a mutations database. Also intron/exon determination.

RL: Users often complain that interfaces are inadequate and are often relics. Does SEQNET have something to alleviate this cry?

AB: Well, as it happens.....
YES YES YES YES YES.
We have things here called CCPs (Collaborative Computational Projects). CCP4 is probably a very familiar one involved in crystallography. I'm involved with CCP11 who's remit is 'protein sequence and structural analysis'. All bioinformatics sites suffer from the problem of how to present data to the users. I can now announce the CCP11 GUI (Graphical User Interface). Using this interface you can drag-and-drop files into a program with automatic file format conversion. You can click on file output in (e.g.) postscript and you'll get a suitable viewer automatically. Help is provided using netscape/xmosaic etc. It is written in tcl/tk and so should be portable from a Cray to a Sinclair

Spectrum. Sorry, I seem to be getting enthusiastic here so I'll stop. The software is available via anonymous ftp in the pub/ccp11 directory
at the s-ind2.dl.ac.uk machine.

RL: What can you do in a place like Daresbury?

AB: First, Daresbury is a tiny village. Its claim to fame is that it is where Lewis Carroll (Alice in Wonderland) was born and lived. There is nothing to do in Daresbury apart from work. It is close to
several major towns though. I think this brings us back full circle to why the SEQNET team get on
so well. We frequently do go-karting or laser-questing together. After such episodes we've removed all frustrations and are too tired to argue. There is also the weekly jaunt out for beer and Indian food.

RL: Why do you put up with all this?
You should have said "Why do EMBnet managers put up with all this" in which case the answer is simple. I think we're all mad.

TIPS from the computer room

Indexing GenBank for GCG by Andrew Lloyd (EMBnet Ireland)

These notes offer a fix for a potential problem associated with GCG. As a disclaimer, let me say that this is NOT a criticism of GCG or its excellent helpdesk but does give a flavour of the sometimes extraordinary demands that GCG users make on the package.

With the databases growing exponentially disc space is always going to be at a premium. Every two months, a new version of GenBank is released, typically 10% bigger than the last one. For GCG users, the GenBank flatfiles have to have a transitory existence on disc as well as the GCG files.

To convert the flatfiles to GCG format you normally issue a command like:

```
% genbanktogcg /data/gb/*.seq -rel=90 -year=95 -mon=08
```

which will work through each of the genbank flatfiles in turn (gbbct.seq, gbest1.seq etc.) and convert them all to GCG format, with GCG style names (gb_ba.*, gb_est1.* etc.).

Now what happens if you haven't got space on the system to store all the uncompressed GenBank flatfiles for a temporary period until the entire database is converted?

The following program reads two name arrays and matches each GenBank name of a compressed *.seq.Z file with the appropriate GCG format names. It then uncompresses each file in turn, converts it to GCG, then recompresses it.

8<——————————————————————————>8

```
#!/bin/csh
#flat2gcg.csh
#works through a compressed database - converts to gcg format
#then recompresses the files and so saves diskspace

#print time check (optional)
date
#set environmental variables
setenv database  /data/GB_flat
setenv gcgbase /data/gcggenbank

#initialise GCG
source /software/gcg/gcgstartup
gcg
gcgsupport

# set up two arrays 1) genbank-flat and 2) gcg names
set gb = (bct est1 est2 inv mam pat phg pln pri rna rod sts syn   \
 una vrl vrt)
set gc = (ba est1 est2 in om pat ph pl pr st ro sts sy un vi ov)

#sets array counter to 1 = first filename
@ k = 1

#a loop to go through all 16 database divisions
while ( $k <= 16 )
      #echoes a name check to the screen
      #the [$k] is an integer variable marking elements in the array
      echo gb$gb[$k]

      #reads the array 'gb' one file at a time uncompress files like  \
          gbbct.seq.Z
      uncompress $database/gb$gb[$k].seq.Z

      #runs genbanktogcg and seqcat on each uncompressed file but
      #renames output to gcg standard name
      genbanktogcg $database/gb$gb[$k].seq -DIR=$gcgbase   \
          -out=gb_$gc[$k] -REL=90 -YEAR=95 -MON=08
      seqcat $gcgbase/gb_$gc[$k].seq -nomon -Summary -Default

      #compresses the flatfile again, but you could rm it instead.
      compress $database/gb$gb[$k].seq

#increments the array counter by one and returns to start of loop
@ k++
end
#print time check (optional)
date
```

8<——————————————————————————>8

Note. lines marked with \ have been warped for layout purposes

Competition: can you write a more elegant solution to this problem? Best entry will be published in the next (Christmas) issue of the newsletter.

SOFTWARE DEVELOPMENT [2]

MPSRCH - a New Method to Access Rapid Sequence Similarity Searching on Massively Parallel Computers using the HASSLE Protocol

R.Doelz, Swiss EMBnet node

• The Searching Method

Sequence searching applications need to balance speed against sensitivity. As described in an earlier issue of embnet.news [1], the 'exhaustive' searching methods are optimised to detect even low similarity in large data collections. Most database searching programs are based on local alignment algorithms, which aim to get a reliable target hit - a region of the query sequence which is best fitting a database sequence.

The Smith and Waterman best local alignment algorithm is currently the most sensitive general-purpose application for database searching. The basic principle of this rigorous method is to compare each character of the query sequence with each character of a databank sequence. To detect remote similarity, insertion of gaps of any size at each position is possible. A sophisticated scoring schema is employed to find the best alignment out of all possible combinations. The handling of gaps can vary. Some implementations, like the one applied in the Blitz E-mail Server at the EMBL in Heidelberg, uses straightforward constant gap penalties. In the implementation of MPSRCH, gap insertion and gap extension penalties vary in numerical values. Once a gap is opened, an extension is less heavily
penalised than the gap insertion itself.

The result of a rigorous database search must always be evaluated for its significance. As lab-bench biology is unknown to computers, statistical methods are employed to discriminate between findings that could have occurred by chance and those which are expected to occur only rarely. A 'probability' or a 'number of expected hits' aid the researcher to judge on the importance of any finding. Parameters which affect the ranking of a hit (i.e. its position in the result list) include the length difference between the query sequence and the database sequence, as well as the specific handling of very short, but well-matched, motifs.

The core sequence comparison method used in the MPSRCH program suite is the Smith and Waterman algorithm, as modified by Goth[2], combined with a prediction of an expected number of results and a ranking function, developed by Collins cand Coulson[3]. The package was developed by John F. Collins and Shane Sturrock at the Biocomputing Resource Unit at the University of Edinburgh, and is distributed by IntelliGenetics, Inc.

• Data for MPSRCH Sequence Searching at EMBnet Switzerland

Both protein and DNA sequence searching is possible with the MPSRCH program suite. Protein searches are, however, recommended, because the results are obtained in few minutes. DNA searches may take hours to complete. EMBnet Switzerland obtains the EMBL, SWISS-PROT and PIR databases from the original database providers. Updates are achieved with HASSLE-based services, which were developed specifically for this purpose. Data updates are synchronised with the originals as closely as possible, and the daily update status is subjected to a quality control system. EMBnet Switzerland also produces TREMBL, a database of all translated reading frames from the EMBL data library. TREMBL is created with the SRS software package (T.Etzold, EMBL). All available protein databases are merged using the 'nrdb' program from the NCBI. After processing, the MasPar searches a total of (currently) more than 150000 protein sequences in a few seconds, even for long protein sequences.

• Accessibility

A subset of the MPSRCH software suite is now available on the powerful 4096-processor MasPar system of the University of Basel. Integration into the widely used GCG software package has been achieved with GCG-like text and WPI-type interfaces, utilising the GCG procedure library interface. The transfer of query data and the execution of commands operate via the HASSLE protocol [4], which is built into the MPSRCH clients and the server at the MasPar front end.

The MPSRCH clients implement a very basic functionality to render sequence searching as easy and as fast as possible for a biologist. The only parameter which can be modified, besides input and output file names, is the set of databases to be searched.

The following databases are supported:

Input	Database
peptide	1) SWISS-PROT + weekly updates
	2) PIR
	3) SWISS-PROT, updates, PIR, TREMBL non-redundant set

DNA 1) Daily EMBL updates
2) EMBL database + EMBL updates
3) non-redundant set, including genbank

The service is freely accessible to all Internet nodes for test-driving and rare access. Unlimited access is available to Swiss academia after registration.

URL for general information:
http://www.ch.embnet.org/hassle.htm

- **Software references**

MPSRCH is Release 2.1D, Copyright (c) 1993, 1994, 1995,John F. Collins, Biocomputing Research Unit, University of Edinburgh, U.K.; Distribution rights by IntelliGenetics, Inc.

A GCG package (Genetics Computer Group, Inc.) is required locally if the GCG-like interfaces are to be used.

HASSLE is Version 5, (C) BioComputing Basel, 1992-1995.

- **Literature references**

[1] Bottu G.: Fundamentals of Database Similarity searching methods, embnet.news 1(1), 1994.
[2] Goth O. An improved algorithm for matching biological sequences. Journal of Molecular Biology, (162), 705-708.
[3] Collins J.F. and Coulson A.F.W. Significance of protein sequence similarities. Methods in Enzymology, 183, 474-487, 1990.
[4] Advanced Computer Network Communication:
Hierarchical Access System for Sequence Libraries in Europe. HASSLE v5 Redaschi N., Doelz R. and Eggenberger F. ; Verlag Dr.U.Doelz, Basel, Switzerland. ISBN 3-905 434-01-6; Electronic documentation available via http://www.ch.embnet.org/hassle.html

NODE FOCUS

Vienna

From its conception in the mid-eighties, the Vienna Biocenter (VBC) was intended to intensify communication in the biological sciences in Austria. Institutes of the University of Vienna, which were previously scattered throughout the city, were concentrated in a purpose-built building next to the non-academic and privately funded Institute for Molecular Pathology (I.M.P.). Following its inauguration in 1992, the VBC has carried out research in the areas of biochemistry, molecular and developmental biology, molecular and classical genetics, plant molecular biology, microbiology, immunology, and virology. The five University Institutes partly from the School of Natural Science and partly from the School of Medicine are still formally independent of each other while relations with the I.M.P. are regulated by a co-operative agreement.

In many respects 1988 represented an important year for biosciences in Austria. EMBnet was initiated and the co-operative contract between the Management of the University of Vienna and the I.M.P. was signed just before construction work started on the VBC. The idea of EMBnet at the European level and basic concepts of the VBC on a national level complemented each other perfectly. It was thus possible to take the development of EMBnet into account during planning of the VBC. Together with the Computer Group of the I.M.P., the Vienna University Computer Center (VUCC) worked hard to obtain the basic contract for establishing Austria's EMBnet node at the VBC.

Modern concepts of information exchange are unthinkable without consideration of all means of electronic communication. Therefore the VUCC was contracted to plan all aspects of the Local Area Network and world wide connection.The basic computer equipment for the Institutes was funded by the Ministry of Science and Research. Five Servers running Novell Netware provide a mixture of PC and Macintosh workstations with the basic software for office work and Internet access. Most of the more than 150 workstations run their own TCP/IP stack and are connected to an IEEE 802.3 standardised ethernet. The aim to provide full EMBnet service for low cost as well as for high end workstations could be realised within the first year of operation. SGI Indigo workstations are used only for special tasks such as molecular modelling.

The main EMBnet Service started on a DEC System 5900 which was financed by the I.M.P. The purchase of this server opened the door to consolidate contracts and join EMBnet officially in March 1993. The quick development and changes in bioinformatics and the computer industry forced us very soon to plan for adjustments. In June 1995 the main service at our node was transferred to a DEC 3900 AXP Server, completely financed by the VUCC. Our users also profit from an improved network connection between the Vienna University Computer Center and its department at the Vienna Biocenter.

At the end of March the bandwidth of the connection line was doubled to 128 kbit/s. This change was significant in EMBnet's network performance monitoring project and encourage us to press ahead with the delayed completion of the connection to the glass fibre backbone of the Vienna University.

During the sensitive phases of integration into EMBnet and the tranition of user services, we stressed user education and training. EMBnet Austria takes care of one of only two teaching class rooms of the VUCC suitable for computer

training. Two courses for Sequence Analysis and Networking were funded by the EMBnet foundation and lead users to efficient use of our main software GCG, EGCG, SRS, Phylip, Molphy, ViennaRNA package and main databases EMBL, Genbank, Swissprot, PIR, ACeDB, AAtDB, SacchDB. Today we serve more than 200 officially registered users throughout Austria with centers of gravity at Institutes of the Universities of Vienna and Salzburg. The activities of the Austrian EMBnet node have been maintained by one person dedicated to be the node manager and system administrator of EMBnet Austria and one person responsible for network administration of the VBC. A further consolidation of infrastructure is planned to enable more energy to be chanelled into EMBnet activities.

Node News

Norwegian EMBnet node

New Databases

1. Transfac/Transterm - These implement nicely with GCG's consensus and findconsensus programs. Recommended to all sites!
2. IMGT/LIGM - The Immnunogenetics database installed under GCG & SRS.
3. SRS-WWW serves 3D structures. Posting on how to do this sent to bionet.software.srs on Wed, 04 Oct 1995.

Swiss EMBnet node (BioComputing Basel)

New staff member

Michael Schmitz joined the BioComputing Laboratory coming from the University of Duesseldorf in Germany, where he worked on RNA secondary structure prediction in the biophysics department with Dr. Steger and Prof. Riesner. Michael has adapted the XSRS1 service to the HASSLE protocol version 5 and develops the network ability of the SRS package.

FastAlert nodes successfully received

The FastAlert system (developed by Florian Eggenberger) has been received very well by the international community. More than 100 sequences have been registered in the last four weeks. To cope with the load, a second compute node was added, and HASSLE redirects incoming requests appropriately.

HASSLE version 5 API fully documented

Nicole Redaschi has completed the HASSLE API and source code documentation. The entire program (.c files) was run through a preprocessor and fed into the JAM formatting system. Interested developers may consult the resulting documentation in LaTEX, RTF or HTML format (On-line version at URL: http://www.ch.embnet.org/hassle.html.

JAM 2.1 released

Improved LaTEX layout and extended HTML indexing are featured in the 'Just Another Metaformat' program JAM version 2.1 which is available by anonymous ftp from ftp.switch.ch. The user guide in JAM is provided as source code and as usual in LaTEX, RTF and HTML versions. (Reference: Doelz, R.: Comput-Appl-Biosci(1995) 11, 224-226.) (Online inspection via URL http://www.ch.embnet.org/development/exposure/jam/JAMINX.HTML).

CAOS/CAMM Center. The Dutch Embnet Node.

The CAOS/CAMM Center has a new addition to its team. Koen Cuelenare has recently joined the Center's staff to replace David Featherston who left to move to Copenhagen where the EMBnet community will probably "meet" him again in an EMBnet sponsored project on Education and Documentation. Koen's primary job is to maintain the Genomics services at the CAOS/CAMM Center. He will also be helping to maintain the collection of Bioinformatics services of our Center, a task he will share with Jack Leunissen and Gijs Schaftenaar. Koen has graduated in Molecular Biology from Wageningen Agricultural University in August 1995. Part of his last year research was performed at the CAOS/CAMM Center, where he assisted Gijs Schaftenaar in the coding of the SRS vs. 4 User Interface (see elsewhere in this Newsletter).

DK node news

At BioBase we have now finished an ascii based menu system for the GCG Rel. 8.1 and EGCG Rel 8.0 programs. The menu system is based on the hypertext browser - hybrow. The functionality is very much like the interfaces known from lynx and tin. Eventually all relevant unix commands and all the other sequence analysis programs available at BioBase will be included as well. All GCG and EGCG commands as well as their options can be used from within the "menu" (without options) and "Menu" (including options) interfaces.