embnet.news

Volume 4 Nr2 (ISSN1023-4144)

July 31, 1997

Editorial

Summertime again ! Well what passes for summer in Ireland, where they refer to 24 hours of constant drizzling rain as "a grand soft day" and have a hopelessly optimistic view of the weather. It turns out that the editorial team was not being hopelessly optimistic when it promised to deliver volume 4.2 of embnet.news within a month of the last, late, issue. We have recently heard that GCG, which for the last several years has provided the de facto global standard for bioinformatic software, has been bought up by or merged with Oxford Molecular. We hope and expect that this rather radical change will generally be a good thing for molecular biologists and bioinformaticians everywhere. On the commercial front we also note the launch of bio1nf0rm a fortnightly newsletter that aims to deliver, for a price, current information about the world of bioinformatics.

Under this assault, embnet.news will continue to deliver its own view of that same world and do this at no cost to its readers. This issue carries a review of a new book about sequence analysis. We would expect that several other volumes on this and related topics will appear over the next few years. There is also an article on methods that can be used to analyse synonymous codon usage in complete genomes, which seems well timed now that complete genomes are being delivered into the public domain almost every month. A nice technique for showing relationships between gene families is presented in the article on GeneDoc. Our regular and irregular features are also present. We try to deliver material that is interesting and informative and welcome contributions.

Enjoy your vacation, we hope that you'll be reading the postscript version of embnet.news under a shady tree in a benign climate.

The embnet.news editorial board:

Rob Harper
Robert Herzog
Andrew Lloyd
Rodrigo Lopez
Peter Rice

Software Development

GeneDoc

Analysis and Visualization of Genetic Variation

Karl B. Nicholas[1], Hugh B. Nicholas Jr.[2], and David W. Deerfield II.[2]

Introduction

GeneDoc provides tools for visualizing, editing, and analyzing multiple sequence alignments of protein and nucleic acid sequences. GeneDoc embeds these tools in an explicitly evolutionary context. This context is most directly expressed as the ability to divide the sequences into groups that reflect the division of superfamilies of genes (and proteins) into distinct families. GeneDoc can analyze and visualize these groups either separately or together. Groups can also be contrasted.

GeneDoc's analysis capabilities include statistical tools that allow users to evaluate explicit biological or evolutionary hypotheses expressed in terms of specific groupings of sequences (Nicholas and Graves, 1983; Nicholas and McClain, 1995). The visualization tools are strongly integrated with the analysis tools and present the analysis results in a form that is easily comprehend and to use in presentations.

1 Bank of America; 315 Montgomery; San Francisco, CA 94127
2 Pittsburgh Supercomputing Center; 4400 Fifth Avenue; Pittsburgh, PA, 15213.

Contents

GeneDoc provides an evolutionary context for alignment editing by evaluating changes to the alignment in terms of explicit evolutionary models. GeneDoc's analysis functions help users discover which sequence residues are important in the structural and functional roles carried out by biological macromolecules.

Editing Tools

GeneDoc's alignment editing features help overcome the current limitations in multiple sequence alignment programs (Nicholas et al., 1995; McClure et al., 1994). Editing can incorporate structural or biochemical information about which residues should be aligned. GeneDoc's alignment scores are based on the accumulated knowledge of evolutionary processes incorporated in the empirical log-odds scoring matrices. GeneDoc provides such matrices for both protein and nucleic acid sequences (Dayhoff et al., 1978; Henikoff and Henikoff, 1992; States et al., 1991, Altschul, 1991). Scores are an objective measure of whether or not specific changes are justified for a given degree of divergence.

GeneDoc offers two different ways to compute a score for any section of your alignment. The first is sum-of-pairs scoring which involves scoring all of the alignments between the independent pairs of sequences and adding these scores together to yield the total alignment score. While sum-of-pairs scoring is less than ideal, it results in alignments that are closer to those produced by superposition of three dimensional structures than do alignments produced by the heuristic methods. The second is weighted parsimony scoring, an alignment criterion that is more biologically desirable but imposes higher computational requirements (Sankoff and Cedegren, 1983). Weighted parsimony will result in an alignment that is most congruent with a user specified phylogenetic tree relating the sequences. Phylogenetic trees for use with weighted parsimony scoring can be imported in either Phylip or Nexus style tree files, or can be built with the graphical tree building interface in GeneDoc. The tree can also be edited in this interface.

GeneDoc has two editing modes that are kept separate from each other to prevent unintended changes in the separate aspects of the alignment. The first mode is alignment editing mode. Characters in one sequence are moved relative to characters in the other sequences in this mode. The overall lengths of the sequences may be changed by either adding or removing gap characters. Gap characters may be added or removed in three ways: in the sequence currently marked by the cursor; to all of the sequences except the one marked by the cursor; or to all of the sequences. Grab and drag arrangement allows sequence residues to be moved without necessarily changing the number of gap characters in the sequence. The second editing mode is residue editing mode in which the sequence residues may be changed from one value to another. This includes changing one sequence character to another and changing gap characters into sequence characters or vice versa. However, no operation that would change the sum of the sequence characters and gap characters is allowed in this mode.

Visualization

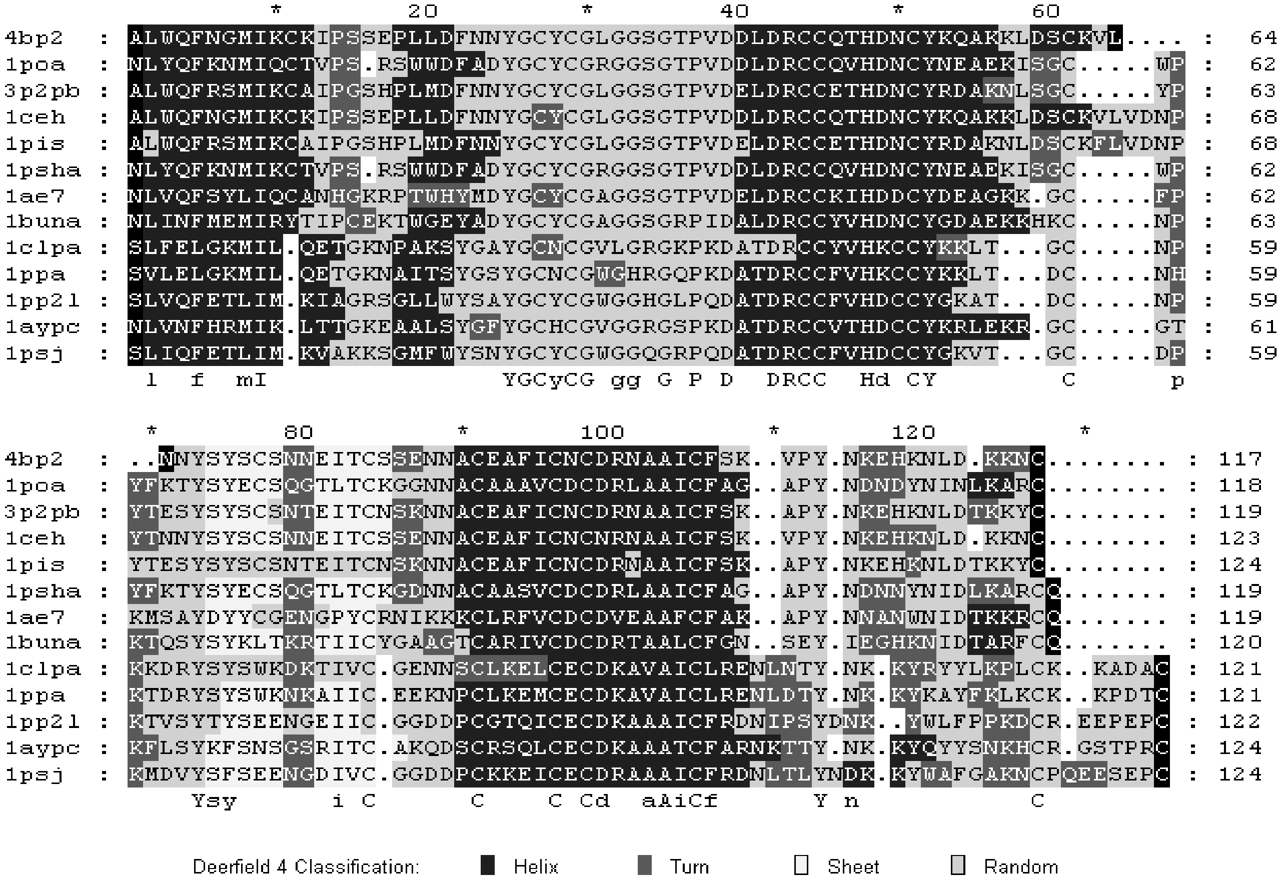
GeneDoc's visualization capabilities are built around two residue display modes and six shading modes. The two residue display modes are to display all residues and to display only those residues that differ from the master sequence. The master sequence is either the consensus sequence for the alignment or for a group within the alignment or the first sequence within the alignment or a group within the alignment. These two residue display modes can be combined with any of the six shading modes.

Three of the shading modes are actually visual displays of widely used analyses of multiple sequence alignments. Conservation mode produces a display that highlights alignment columns that show from 1 to 4 user defined levels of conservation. Quantify mode highlights the 1, 2, or 3 most frequent residues found in each column of the alignment, which focuses attention on the sequence positions that have evolved with a similar pattern of differentiation even though the actual residues at the position may differ. In both conservation and quantify mode the user sets the colors used for the highlighting and determines whether or not to treat conservative substitutions as if they are identical (e.g., I, L, V, M). Physiochemical properties mode analyzes each alignment position in terms of the hierarchical set of amino acid properties similar to those proposed by Dickerson and Geis (1969) and each position is shaded to identify the most exclusive set to which all of the amino acids at that position can be assigned.

The other three shading modes also highlight alignment position according to an analysis. However the analysis is either largely (property shading mode) or entirely (structure and manual shading modes) under the control of the user. The property shading mode allows the user to divide the possible sequence residues into an arbitrary number of sets each assigned its own coloring scheme. The colors can then be applied to those columns where the property identified with the set is conserved or they can be applied to every residue in the alignment.

The structure shading mode allows users to define an arbitrary number of states that the sequence residues may inhabit and assign colors to each state. Users can import information about protein secondary structure or RNA folding and color specific residues in a particular sequence, a group of sequences, or the entire alignment according to that structural information. GeneDoc has provisions for importing state information from the Protein Structure database (PSdb) (Deerfield and Geigel, 1996), DSSP (Kabsch and Sander, 1983), both are derived from Brookhaven PDB files. State information may also be imported from many of the structure prediction programs on the EMBL server, or as user defined values of from the reformatted version of the 3D_ALI database (Pascarella and Argos, 1992) available on the GeneDoc web site. User defined values require a file that assigns the residues of a specific sequence to states defined in a file of user created state definitions. The residues in the specific sequence will be highlighted in the corresponding color. This shading may be extended to the other sequences in the alignment or only to those in the same group as the original sequence. It is possible to shade every sequence in the alignment individually in this manner.

Manual shading allows the user to assign specific colors to individual residues with point and click ease.



```
                  *        20         *        40         *        60
4bp2   : ALWQFNGMIKCKIPSSEPLLDFNNYGCYCGLGGSGTPVDDLDRCCQTHDNCYKQAKKLDSCKVL.... :  64
1poa   : NLYQFKNMIQCTVPS.RSWWDFADYGCYCGRGGSGTPVDDLDRCCQVHDNCYNEAEKISGC.....WP :  62
3p2pb  : ALWQFRSMIKCAIPGSHPLMDFNNYGCYCGLGGSGTPVDELDRCCETHDNCYRDAKNLSGC.....YP :  63
1ceh   : ALWQFNGMIKCKIPSSEPLLDFNNYGCYCGLGGSGTPVDDLDRCCQTHDNCYKQAKKLDSCKVLVDNP :  68
1pis   : ALWQFRSMIKCAIPGSHPLMDFNNYGCYCGLGGSGTPVDELDRCCETHDNCYRDAKNLDSCKFLVDNP :  68
1psha  : NLYQFKNMIKCTVPS.RSWWDFADYGCYCGRGGSGTPVDDLDRCCQVHDNCYNEAEKISGC.....WP :  62
1ae7   : NLVQFSYLIQCANHGKRPTWHYMDYGCYCGAGGSGTPVDELDRCCKIHDDCYDEAGKK.GC.....FP :  62
1buna  : NLINFMEMIRYTIPCEKTWGEYADYGCYCGAGGSGRPIDALDRCCYVHDNCYGDAEKKHKC.....NP :  63
1clpa  : SLFELGKMIL.QETGKNPAKSYGAYGCNCGVLGRGKPKDATDRCCYVHKCCYKKLT...GC.....NP :  59
1ppa   : SVLELGKMIL.QETGKNAITSYGSYGCNCGWGHRGQPKDATDRCCFVHKCCYKKLT...DC.....NH :  59
1pp2l  : SLVQFETLIM.KIAGRSGLLWYSAYGCYCGWGGHGLPQDATDRCCFVHDCCYGKAT...DC.....NP :  59
1aypc  : NLVNFHRMIK.LTTGKEAALSYGFYGCHCGVGGRGSPKDATDRCCVTHDCCYKRLEKR.GC.....GT :  61
1psj   : SLIQFETLIM.KVAKKSGMFWYSNYGCYCGWGGQGRPQDATDRCCFVHDCCYGKVT...GC.....DP :  59
          l  f  mI               YGCyCG gg G P D  DRCC  Hd CY        C      p

          *        80         *       100         *       120         *
4bp2   : ..NNYSYSCSNNEITCSSENNACEAFICNCDRNAAICFSK..VPY.NKEHKNLD.KKNC......... : 117
1poa   : YFKTYSYECSQGTLTCKGGNNACAAAVCDCDRLAAICFAG..APY.NDNDYNINLKARC......... : 118
3p2pb  : YTESYSYSCSNTEITCNSKNNACEAFICNCDRNAAICFSK..APY.NKEHKNLDTKKYC......... : 119
1ceh   : YTNNYSYSCSNNEITCSSENNACEAFICNCNRNAAICFSK..VPY.NKEHKNLD.KKNC......... : 123
1pis   : YTESYSYSCSNTEITCNSKNNACEAFICNCDRNAAICFSK..APY.NKEHKNLDTKKYC......... : 124
1psha  : YFKTYSYECSQGTLTCKGDNNACAASVCDCDRLAAICFAG..APY.NDNNYNIDLKARCQ........ : 119
1ae7   : KMSAYDYYCGENGPYCRNIKKKCLRFVCDCDVEAAFCFAK..APY.NNANWNIDTKKRCQ........ : 119
1buna  : KTQSYSYKLTKRTIICYGAAGTCARIVCDCDRTAALCFGN..SEY.IEGHKNIDTARFCQ........ : 120
1clpa  : KKDRYSYSWKDKTIVC.GENNSCLKELCECDKAVAICLRENLNTY.NK.KYRYYLKPLCK..KADAC : 121
1ppa   : KTDRYSYSWKNKAIIC.EEKNPCLKEMCECDKAVAICLRENLDTY.NK.KYKAYFKLKCK..KPDTC : 121
1pp2l  : KTVSYTYSEENGEIIC.GGDDPCGTQICECDKAAAICFRDNIPSYDNK..YWLFPPKDCR.EEPEPC : 122
1aypc  : KFLSYKFSNSGSRITC.AKQDSCRSQLCECDKAAATCFARNKTTY.NK.KYQYYSNKHCR.GSTPRC : 124
1psj   : KMDVYSFSEENGDIVC.GGDDPCKKEICECDRAAAICFRDNLTLYNDK.KYWAFGAKNCPQEESEPC : 124
             Ysy      i C      C    C Cd  aAiCf      Y n           C
```

Deerfield 4 Classification: ■ Helix ▩ Turn □ Sheet ▢ Random

Analysis

Many of GeneDoc's analyses are Kolmogorov-Smirnov (K-S) analyses of pairs of cumulative distribution functions (Sokal and Rohlf, 1995). K-S analyses provide a rigorous assessment whether two distributions are different. The difference can be either in the location or shape of the distributions. Thus, K-S tests are more broadly based than more common tests like Student's T test or the F test. The K-S tests use distributions of alignment scores or comparisons of sequences in terms of the percentage of identities between a pair of aligned sequences. Probably the most useful test is the analysis of whether the scores for pairs of sequence within the same group are smaller than the scores for pairs of sequences that are in different groups. A positive result for this test indicates that the grouping categories are systematically reflected in the sequences (Nicholas and Graves, 1983; Nicholas and McClain, 1995).

There are two types of contrast analysis that contrast the sequences within one group with those in the other groups on a position by position basis. The PCR contrast highlights sites that meet two criteria. First is that a single residue is completely conserved within the group. Second, this conserved residue does not appear, at that position, in any sequence outside of the group in which it is conserved.

The group contrast analysis is less restrictive within the group than is the PCR contrast analysis. In the group contrast analysis all of the sequence residues at a site are required to have a positive similarity score with each other. Residues outside of the group must have a negative similarity score with every residue from within the group.

Files

GeneDoc GCG's msf file format as its primary file type using the header region to store information about residue display and shading modes along with large amounts of user configuration choices. In addition to the msf files, sequences may be read from or written to ClustalW aln files, Pearson FASTA files, and PIR formatted files. Aligned sequences can also be written to Phylip interleaved files. Graphic results can be sent to the printer or to a Postscript file by using an appropriate printer driver. Highlighted results can also be exported in Windows Enhanced Meta Files or in Macintosh style PICT files.

Summary

GeneDoc is a full featured multiple sequence alignment visualization, editing, and analysis tool. It has an easy-to-use point and click user interface with extensive keyboard mapping for advanced users. In addition to the features described above there are many more features and additional details in the extensive context sensitive help files that comes with the program. Figure 1 shows an alignment of 13 phospholipases A2. The shading for each sequence indicates the secondary structure state of the residue as derived from the three dimensional coordinates taken Brookhaven PDB file that is used to label the sequence. The secondary structure states were computed using the four state PSdb model (Deerfield and Geigel, 1996). The alignment and PSdb files used to create the figure are available on the GeneDoc web site.

GeneDoc version 2.1 runs on any IBM compatible personal computer under Windows 3.1, Windows 95 or Windows NT. It can be obtained at no cost over the World Wide Web at: http://www.cris.com/~Ketchup/genedoc.shtml.

Thanks to Russell Malmberg a version that runs on DEC Alpha workstations under Windows NT is available at: http://dogwood.botany.uga.edu/malmberg/software.html. GeneDoc has benefited from the comments, suggestions, and error reports from a number of early users. Additional feedback is welcomed by KBN at RKetchup@cris.comS.

References

- Altschul, S.F. 1991. J. Mol. Biol., 219: 555 - 565. Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C. 1978. In "Atlas of Protein Sequence and Structure" vol. 5(3) M.O. Dayhoff (ed.), National Biomedical Research Foundation, Washington. pp. 345 - 352. Deerfield, D.W., II and Geigel, J., 1996. http://www.psc.edu/biomed/pages/research/PSdb/PSdbPaper/
- Henikoff S. and Henikoff, J.G. 1992. Proc. Natl. Acad. Sci. USA. 89: 10915 - 10919.
- Dickerson, R.E. and Geis, I. 1969 The Structure and Actions of Proteins. . pp. 16 - 17. Harper & Row Publishers, New York, NY
- Kabsch, W. and Sander, C. 1983 Biopolymers 22: 2577 - 2637.
- McClure, M.A., Vasi, T.K., and Fitch, W.M. 1994 Mol. Biol. Evol. vol. 11: 571 - 592. Nicholas, H.B. Jr., and Graves, S.B. 1983 J. Mol. Biol. 171: 111 - 118.
- Nicholas, H.B. Jr. and McClain, W.H. 1995. J. Mol. Evol. 40: 482-486.
- Nicholas, H.B. Jr., Ropelewski, A.J., Deerfield, D.W. II., and Behrmann, J.G. 1995 Proceedings of the 10th International Conference on Methods in Protein Structure, Eds. M.Z. Atassi and E. Appella, Plenum Press, New York pp. 515 - 525.
- Pascarella, S. and Argos, P. 1992 Prot. Engng. 5: 121 - 137.
- Sankoff, D. and Cedergren, R.J. 1983. In "Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison." D. Sankoff and J.B. Kruskal (eds.) pp. 253 - 263. Addison-Wesley, Reading, MA
- Sokal, R.R. and Rohlf, F.J. 1995 Biometry, 3rd ed. W.H. Freeman & Co. New York, NY.
- States, D.J., Gish, W., and Altschul, S.F. 1991. Methods 3: 66 - 70. Last modified: Mon Jul 21 13:40:25 1997

BITS

Bioinformatics Theory Section

The analysis of codon usage patterns

James O. McInerney, The Natural History Museum, Cromwell Road, London SW7 5BD, UK.

INTRODUCTION

With the exceptions (in the universal genetic code) of ethionine (AUG) and Tryptophan (UGG), most codons have a synonymous alternative, where the substitution of 'silent' nucleotides will not change the encoded amino acid. Since the first nucleotide sequence repositories were being assembled, researchers have observed that not all codons for the same amino acid are used with equal frequency. One often speaks of "the" codon usage pattern of an organism, but rarely is it true that all of the genes exhibit the same codon usage patterns. Within an organism it is normal for there to be significant differences in synonymous codon usage between genes. This observation can have implications for the design of primers and the bioinformatic determination of exon boundaries.

The two greatest influences on codon usage patterns are thought to be mutational bias (where the organism has a characteristic tendency to turn C/G base pairs into A/Ts or vice versa) and translational selection (typically where highly-expressed genes in many bacteria and other organisms preferentially use a subset of 'optimal' codons) see Sharp et al., 1993. Because the datasets involved are typically large and complex, multivariate analyses are essential for any non-trivial investigation of variation in codon usage patterns. Correspondence analysis, which is appropriate for contingency data (a dataset in which the values are not independent) is the multivariate analysis method of choice for studying variation in codon usage patterns (Greenacre 1984).

METHOD

In order to analyse codon usage variation between genes, it is necessary to derive the Relative Synonymous Codon Usage (RSCU) values for each gene (Eqn. 1).

$$RSCU_i = \frac{Obs_i}{Exp_i} \qquad (1)$$

Where RSCU$_i$ is the Relative Synonymous Codon Usage value for codon i, Obs$_i$ is the observed number of occurrences of a codon i. and Exp$_i$ is the expected number of occurrences of codon i. The expected number of occurrences of a codon is calculated according to equation 2.

$$Exp_i = \frac{\sum aa_i}{\sum syn_i} \qquad (2)$$

Where Exp$_i$ is the expected frequency of occurrence of codon i, $\sum aa_i$ is the number of times the encoded amino acid is present in the protein sequence and $\sum syn_i$ is the number of synonyms for the amino acid encoded by codon i. An RSCU value greater than 1 means that a codon is used more often than expected, whilst values less than 1 indicate its relative rarity.

Multivariate analyses of codon usage patterns seek to identify the most relevant trends governing choice of codon in a given organism. It is necessary to use more than a single gene for these analyses and preferably a whole genome. It is possible however, to make a reasonably good assessment of the forces governing codon usage in an organism with as little as 30 or 40 genes. Each gene can be described by the RSCU values for each codon. There are 59 values when dealing with the universal genetic code: AUG, UGG, and the three stop codons are not included. On the other hand, there are 61 in the mycoplasma/spiroplasma genetic code where UGG and UGA are synonymous alternatives for Tryptophan. Correspondence analysis is a method for discovering trends in the dataset and proceeds by plotting all of the codon usage values in an N-dimensional hyperspace (the number of dimensions is determined by the number of synonymously degenerate codons in that particular genetic code). The points on this high-dimensional space can be said to resemble a 'cloud'. In the absence of any pattern of codon usage variation, this cloud will be amorphous and there will be very little difference between the axes of greatest dispersion of the data. Any variation in codon usage between the genes in the dataset will distort the shape of this cloud in a characteristic way and one or a few axes will explain most of the variation in the dataset. Most implementations of Correspondence Analysis will calculate the position of each gene and each codon on each of the first few axes through the dataset.

A graphical display and examination of the relative position of genes on the first and second axes (those which explain the greatest proportion of the variation in the dataset) can be a powerful tool. Outliers, for example, can be readily identified. It is also useful to plot the positions of the genes on the axis, or axes, of greatest dispersion (or greatest explanatory power) against other measures of codon usage bias. One such measure is the G+C base composition of the third position of codons for which there is a synonymous alterna-

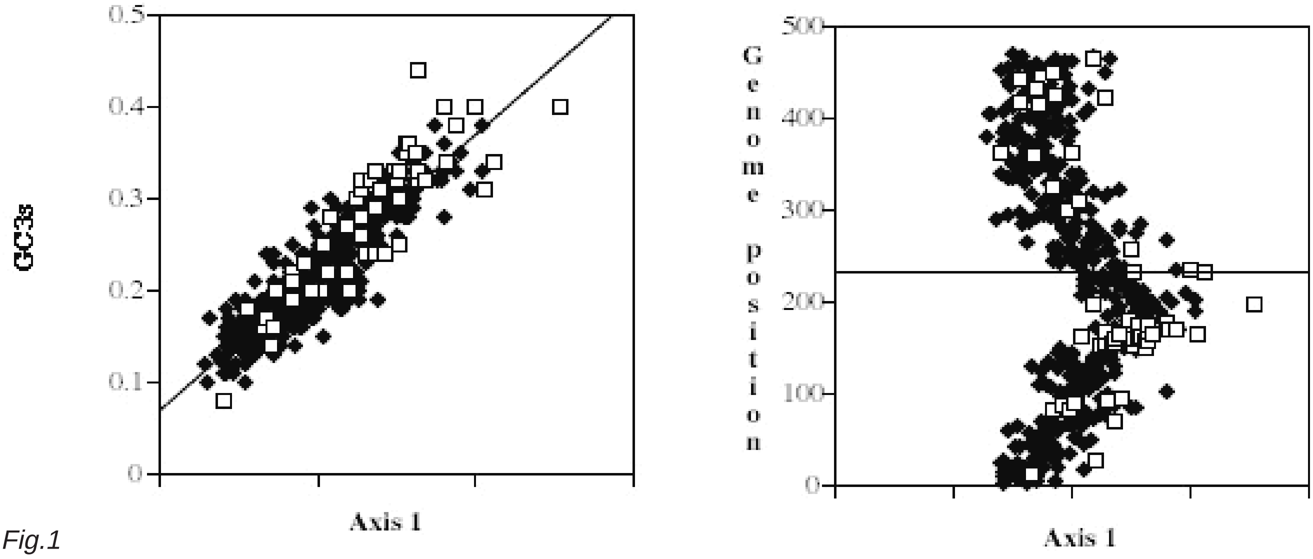


Fig.1

tive (GC's). An obvious correlation between position on the first axis and GC's would indicate that variation in base composition is an important contributor to codon usage. It should also be possible to identify genes of high (elongation factors, glycolytic enzymes, ribosomal proteins) and low (regulatory proteins) expression level in the dataset and investigate any correlation between position on the graph and level of expression. It is known for a number of organisms that the highly expressed genes tend to utilise a small subset of 'optimal codons'.

An example using the Mycoplasma genitalium genome

The M. genitalium genome sequence was reported to be 580,070 bp in length (Fraser et al., 1995), The analysis of codon usage variation in this genome revealed that mutational bias is very important. This was deduced by carrying out a correspondence analysis on the RSCU values for each gene and plotting the position of the genes on the axis of greatest dispersion against the GC's values for these genes. The relationship is very strong, with a correlation coefficient of 0.901 (fig 1, preceding page).

In contrast to many other bacteria, there was no detectable difference in M. genitalium codon usage as a function of expression level. A very surprising observation in this organism was that codon usage variation (and GC's) is associated very strongly with position on the genome (McInerney, 1997a). This phenomenon has never been observed in a bacterium before and remains unsolved. However it does appear to be a powerful evolutionary mechanism.

Software Availability

- Software for performing detailed analyses of codon usage variation are available from a number of sites.
- The program GCUA: General Codon Usage Analysis (McInerney, 1997b) is available from ftp://ftp.nhm.ac.uk/pub/gcua.
- A web site dedicated to multivariate analysis of DNA sequences is available at http://acnuc.univ-lyon1.fr/mva/coa.html.
- A program, CodonW is being released by John Peden at ftp://www.molbiol.ox.ac.uk/cu.
- FORTRAN source code for a variety of programs is available at ftp://acer.gen.tcd.ie/pub/cod/ the file README contains a brief description of each file.

References

- Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M., Fritchman, J.L., Weidman, J.F., Small, K.V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T.R., Saudek, D.M., Phillips, C.A., Merrick, J.M., Tomb, J.-F., Dougherty, B.A., Bott, K.F., Hu, P.-C., Lucier, T.S., Peterson, S.N., Smith, H.O., Hutchinson III, C.A.and Venter, J.C.(1995). The minimal gene complement of Mycoplasma genitalium.Science 270(5235): 397-403.
- Greenacre, M. J. (1984). Theory and applications of correspondence analysis. London, Academic Press.
- McInerney, J. O. (1997a). Prokaryotic genome evolution as assessed by multivariate analysis of codon usage patterns. Microb. Compar. Genomics 2(1): 1-10.
- McInerney, J. O. (1997b). GCUA: General Codon Usage Analysis. London, The Natural History Museum; ftp://ftp.nhm.ac.uk/pub/gcua.
- Sharp, P. M., Stenico, M., Peden, J. F. and Lloyd, A. T. (1993). Codon usage: mutational bias, translational selection or both? Biochem. Soc. Trans. 21: 835-841.

TIPS from the computer room

Executing FTP commands in batch mode

JR Valverde and AT Lloyd

This tip would be probably useless in an ideal world, but network lines being as they currently are, you will surely find yourself having trouble downloading files because of high network traffic. In these circumstances it is better to defer the transfer until there is less network congestion - usually late in the night or over weekends.

But who wants to stay overnight if the computer can do it all by itself?

The easy way to transfer a file after hours is to execute the command in batch mode. The problem with this is that the very first step, authentication, must normally be done interactively. This is a problem since you won't be there and FTP won't usually take a user name and password from standard input unless it is a terminal.

And how do we manage to run FTP in batch mode? Easy, just modify the following script to suit your needs:

Clip between the lines and save the script somewhere in yourhome hierarchy with a suitable name, such as "batchftp.sh"

1.Use a text editor, such as vi or emacs, to change the files
2.Substitute "ftp.server.on.the.net" by the name of the server you wish to access
3.Replace the lines "dir" and "other-commands-here" with whatever commands you want to execute
4.Save the file and exit the editor
5.Make the file executable with the command

```
      % chmod u+rwx batchftp.sh
```

6.Run the script or schedule it at a given time: you can run it immediately (e.g. for testing) with

```
      % ./batchftp.sh
```

or submit it for later processing with 'at':

```
      % at -m username -f batchftp.sh hh:mm
```

where hh:mm is the time when it should be run in hours (24 hour clock), and minutes. The '-m username' option ensures that you will receive a notification message by e-mail once it has finished.

-------------8< cut here 8< ----------

```
#!/bin/sh
#
# VERY SIMPLE script to automate FTP.
# Jose R. Valverde EMBnet/CNB 5-Feb-1997
# <--lines beginning with the # sign are
# not read by the computer
# you might want to issue a 'cd' commands to
# position yourself in the right directory
# to receive the incoming file
# cd ~/xfer
# do the actual FTP
ftp -n -v ftp.server.on.the.net << FTPCMDS
user anonymous myname@embnet.org
dir
# other-commands-here, such as:
# cd pub/data/ change directory on
# remote server
# lcd /data/incoming change directory
# back home
# get readme.doc get one file
# mget *.Z get several files together
#
bye
FTPCMDS
```

--------------8< cut here 8< ------------------

The '<< LABEL' at the end of some lines tells the shell to use the following lines in the script (up to the a line beginning with LABEL) as the standard input for the command.

The line beginning 'user anonymous...' only works because of the -n on the previous line. If you leave out the -n, the ftp protocol searches for a file called .netrc in your home or login directory. You can create '.netrc' to contain login information for one or more machines, and can be useful if you use ftp to download files from many different computers on a regular basis.

The basic information must be stated as follows:

```
machine1 host.domain.net
login username
password somepassword
machine2 otherhost.domain.net
login username
password somepassword
```

Note that you have to enter the password, so it is NOT advisable that you use this method for other than anonymous logins on public servers. If you still use it for real accounts, make sure nobody else can read this file and remember that even so, it is a bad idea to store real passwords on it.

Software Developments

Webin and Sequin

New Sequence submission systems at the European Bioinformatics Institute

Katarzyna Kruszewska, Guenter Stoesser

Introduction

Most journals require submission of DNA sequence information to one of the international nucleotide sequence databases prior to publication. When the data submission is received, database staff provide the author with a unique database accession number to identify the sequence. The database accession number is included in the manuscript, preferably as a footnote on the first page of the article, or as required by individual journal procedures. This allows the community to retrieve the data upon reading the journal article.

Recently, a new generation of sophisticated sequence submission tools have become available from the EBI - allowing authors to submit sequence data to the EMBL nucleotide sequence database in a simple and user-friendly way, either via WWW forms (Webin) or via a multi-platform (Mac/PC/Unix) stand-alone software tool (Sequin). Database entries created by the new submission systems and submitted to the EMBL nucleotide sequence database at the EBI will be exchanged and shared among the International Collaboration of Nucleotide Sequence Databases (DDBJ/EMBL/GenBank). Therefore it is only necessary to submit to one of these databases regardless of where the sequence data will be published.

How to submit data to the EMBL nucleotide sequence database.

Webin - New WWW Sequence Submission Tool

Webin is the new WWW Sequence Submission Tool for submitting nucleotide sequence data and associated biological information to the EMBL nucleotide sequence database at the European Bioinformatics Institute (EBI).

To access WebIn at the EBI please use the following URL: http://www.ebi.ac.uk/submission/webin.html or click on icon WEBIN

WebIn guides the user through a sequence of WWW forms allowing the submission of sequence data and descriptive information in an interactive and easy way. All the information required to create a database entry will be collected during this process:

1. Submitter information
2. Release date information
3. Sequence data, description and source information
4. Reference citation information
5. Feature information (e.g. coding regions, regulatory signals etc.)

Webin also allows multiple sequence submissions from the same author in a quick and convenient way. Webin's copy facilities allow replication of information (submitter details, source and citation), thus significantly speeding up the submission process for multiple submissions.

Webin is written in Perl using the CGI.pm module. This tool works with all browsers, but performs best with Netscape 3.0 (and higher) due to built-in JavaScripts.

Sequin - New Stand-Alone Sequence Submission Tool

Sequin is the latest multi-platform (Mac/PC/Unix) stand-alone software tool developed by the NCBI for submitting entries to the EMBL, GenBank, or DDBJ sequence databases. Sequin is an interactive, graphically-oriented program based on screen forms and controlled vocabularies that guides the submitter through the process of entering sequence data and providing biological and bibliographic annotation. Sequin is designed to simplify the sequence submission process, and to provide increased data handling capabilities to accommodate very long sequences, complex annotations, and robust error checking. The Sequin program, along with detailed downloading and installation instructions plus general information is available from the EBI via WWW browser and anonymous FTP.

http://www.ebi.ac.uk/subs/allsubs.html
ftp://ftp.ebi.ac.uk/pub/software/sequin/
Sequin is replacing Authorin as the stand-alone submission tool. Authorin is no longer available from the EBI. For the time being though, we will continue to accept and process Authorin submissions.

Further submission information

For further information on submission of sequence data to the EMBL nucleotide sequence database please access:

http://www.ebi.ac.uk/ebi_docs/embl_db/ebi/authorinfo.html
or contact database staff at:
EMBL Nucleotide Sequence Submissions
e-mail: datasubs@ebi.ac.uk
telephone: +44-1223-494499
telefax: +44-1223-494472

Software

WWW2GCG

A Web interface to the GCG package

Marc Colet, interviewed by Robert Herzog, Belgian EMBnet Node

What is WWW2GCG ?

As its name implies, WWW2GCG is a web interface to the full set of GCG programs. It is presently geared to version 9 of the package, but it can work with version 8, given a few simple modifications. In contrast to the early releases, several months ago, most GCG programs work flawlessly under this interface. The main gray area remains the sequence editor, which is presently being handled by the development of a Java sequence editor.

How it works

WWW2GCG is based on html pages which are generated by a set of scripts on the basis of the ".cmd" description files of the GCG system[1]. Initially, these html pages were regenerated at each installation of WWW2GCG. This produced many problems, as some of the .cmd files are not totally consistent with GCG's own syntax. By incorporating pregenerated and verified html pages in the distribution,

[1] The system can in fact be used for any program that expects a command line and for which a description of the input parameters has been created in the proper format. Egcg has been included in the set of WWW2GCG commands in this way

installation on the client side is much simpler. The html pages are presented to the user from a main menu page, the structure of which closely matches the chapters of the GCG manual. This is a simple compact tabular page with a set of selectable areas for the various programs inside a given chapter. The web browser on the client side presents the user with the main page of the chosen GCG program, and the user can interact with it, by giving values to one or more of the parameters (either the obligatory "prompted" parameters, or the "optional" parameters). The user sends the modified page to the server where the changes determine various actions, such as reading the length of a given sequence and showing it on a new page forwarded to the client.

As seen from the user's side

The first action after asking for WWW2GCG is authentication of the user, who needs to be registered as a valid UNIX user on the computer. At a later stage, the GCG identification mechanism will make sure that he is among the registered users of the programs. After being given the access, the user can ask for the help at any level. The mechanism behind this is the standard HTML GCG help. It opens in a new window and can stay on the screen as an handy reference. Incidentally, this help system is only visible to validated GCG users.

The basic philosophy of the WPI or the SeqLab GCG interfaces, when dealing with sequence files, is to refer to them in so-called "list files". This concept has been adopted here, including such valid options inside a list file as begin:, end: etc.. The management of list files is integrated in the package as a simple file editor. In addition, any file with a ".pep", a ".seq" or a ".gap" extension in the current directory can be accessed as well.

Whenever a program page is opened, it appears in the first instance with only the prompted parameters and data file. Access to the optional parameters needs an additional click. When all options have been selected and validated, a click to the main action button (marked with the large red arrow) sends a command to start the GCG program itself. After the few seconds of its execution, a new html page provides an opportunity to look at the results of the analysis. When the initial GCG program generates output tailored for another program, the latter is presented as an optional "piped" program. Another click and either the text output appears, or the proposed piped program can be started. The classical piping examples are compare-dotplot or mapsort-plasmidmap. At this time, graphics are presented through the Java applet fig2java[2]. This applet reads the "figures" format generated by GCG and produces the corresponding picture in a separate window. The output is not completely flawless at this time, but will be improved with the future developments of Java and the applet [3].

The "WebShell"

In addition, WWW2GCG provides what could be called a "Webshell". Indeed, most usual UNIX commands (ls, cp, more, mv, rm, cd, mkdir, etcI) can be executed from the main page, or from within any of the program pages, by clicking on the "Directories" area at the top right of the page. The power user can even start any complex UNIX command from an editable text area at the bottom of this page Batches and Perl scripts are only a few keystrokes away. Navigating back and forth between the Webshell to any GCG program is also a matter of a mouse click.

The GCG manager's perspective

The installation of WWW2GCG obviously requires the previous installation of a fully functional GCG package. It also needs a functional httpd server running on the same computer. But WWW2GCG is really very easy to install and configure. There are in fact two major phases to the installation:

- compile a C program, called getit
- declare two logical links, to the root of the GCG package and to the root of the httpd server

The compilation of the C program works on all platforms that have been tested, both with the native C compilers and the public domain GNU C compiler. The first function of getit is to change the owner of the process that gets started through WWW2GCG by the httpd deamon. This process starts out owned by "nobody" [4]. The first thing getit does is to test whether "nobody" is indeed the owner of its own process. If it is not the case, getit exits with no action whatsoever. Otherwise, it will go on checking the various requirements, and especially the identity of the user who identified himself, and terminate after giving the user the right to start the requested program. As "nobody" should be highly constrained in its privileges on the computer, making sure that getit initially runs as nobody is indeed an important safety measure.

In order to do this, getit has to run as a "setuid" process, a prospect that will have many UNIX managers shivering in terror. Indeed, if you consider that giving telnet access to your machine is dangerous, then don't use WWW2GCG either. Both are at the same level of (in)security. But as giving access of your computer to your users is the inevitable

[2] Nobuyuki Miyajima, Kazusa DNA Research Institute Department of Genome Informatics, http://pka3.kazusa.or.jp/java/fig2java/

[3] With the recent availability of Acroread on OSF/1, the initially chosen pdf alternative will probably be revived.

[4] Nobody is generally chosen as the owner of the httpd process. But any user may have been chosen. This will not impair the functioning of WWW2GCG

fate of a GCG manager, the issue is irrelevant. A distinctly different issue is whether you accept exposure of your GCG computer to the Internet. Controlling it with some firewall is a healthy safety precaution, and the httpd calls have to be restricted to those users and those addresses you consider acceptable.

If your httpd server is installed in a classical way, it will search for a htpasswd file if htaccess is turned on. As a consequence, this htpasswd file must contain the necessary information for all the authorized WWW2GCG users and be kept up to date. As this may seem rather cumbersome, an alternative way can be chosen, where the httpd server is told to read the /etc/passwd file[5]. In this way, any valid UNIX user is placed in a position to use WWW2GCG. On computers where a shadow password system is installed, some appropriate file has to be provided. This file must contain the uid:password couples and should reside somewhere, to be readable only by "nobody". This is probably quite a good protection of the system.

Getit, the brain and soul of WWW2GCG

The central function of WWW2GCG, i.e. reacting to the commands composed by the user on the web page, is accomplished by the second part of getit [6]. This is in fact a parser program that will analyze the streams coming from the httpd server, where it acts as a cgi program. Depending on what it got from the httpd server, getit will either start one of a 20 odd Perl scripts designed to accomplish one of the many management chores, like changing the current directory, getting the length of the sequences, looking for any file more recent than a certain time, etc. or finally, when all options and parameters have been set, the particular GCG program that was called. Incidentally, a pause of 1 second had to be incorporated so that any file created by a GCG program in the home directory of the user will be time-stamped at least 1 second "younger" than the process that started the program.

The management of a virtual WWW2GCG "session"

In order to work properly, GCG needs a set of global variables, known as "names" and "symbols". These variables and the management of them is a legacy of the VMS operating system on which GCG was initially developed. On a UNIX computer, these variables are stored in so-called "shared memory segments", which are areas of the computer memory that will keep these variables during the whole course of a user session. With the web mechanism, the concept of session is irrelevant, as each call to the httpd server is a separate process. For WWW2GCG to work, a mechanism had to be invented to simulate a "session". When a call to WWW2GCG happens, a file is created in the home directory of the user, with the name .gcgnid. This file contains the reference to a small segment of shared memory that is reserved for the user. The date of creation of this file is vital for the following steps of the mechanism. This memory area contains the "names" and "symbols" and the GCG variable gcgnid, which are vital for all GCG programs to run. This setup will exist for the next 8 hours. Each time getit is restarted, ...gcgnid gets a UNIX "touch", so that it will keep the system up for another 8 hours. This remains valid only as long as the current web browser is connected. So the virtual GCG "session" lives on the userUs screen as long as the browser is open. Once closed, the web authentication has to be reinitiated. The shared memory segment will be liberated by the usual GCG mechanism, using their shared memory management.

To conclude

WWW2GCG is a nice piece of software engineering, completely tailored to accomplish a set of well defined tasks, provided that the programs to be controlled are command line oriented. It is surprisingly light and simple to install on a computer already configured with GCG and an httpd server. The size of the complete compressed code fits on a single HD diskette. The user interface it provides is highly appreciated by the GCG end user, as it makes everything much more intuitive than the naked command line. GCG offers good user interfaces like WPI and SeqLab which may be a better choice, provided that the users are fully equipped with X-windows, although such configurations are much more demanding on server resources.

WWW2GCG has been installed several hundreds of times on all UNIX flavours supported by GCG. Its recent enhancements with the "Webshell" allow virtually all UNIX tasks, like file editing, file and directory management to be executed without leaving the web browser. A few tasks related to single and multiple sequence editing (seqed, lineup, gelassemble, etc.) require the use of a telnet client, but their replacement with ad-hoc editors running on the client, like the soon to be released "roasted" Java sequence editor should improve the situation in the near future.

WWW2GCG was first presented at the 1995 WWW meeting in Darmstadt. All the essential elements of the web client-server mechanism were already settled at that time. Other web interfaces to GCG have been build. The most well known is probably w2h, a production of Martin Senger, presently at the EBI.

[5] This needs a minor edit to the source code of the httpd deamon

[6] In his development of W2H, Martin Senger designed a separate program to accomplish only the task of changing the owner of the process.

Availability

WWW2GCG is available as a compressed file on the server of the Belgian EMBnet Node, at ftp://alize.ulb.ac.be/pub/www2gcg. A few html documents at the same URL will help you in installing the software. A demo of WWW2GCG can be viewed at http://alize.ulb.ac.be/demo/www2gcg.

References

Colet, M. and Herzog, A.. WWW2GCG, a Web interface to the GCG biological sequences analysis software. Comput. & Graphics (1996) 20, 445-450

Martin Senger, W2H, WWW interface to the GCG Sequence Analysis Software Package,
embnet.news (1997) vol4.1, 8-9

INTERviewNet

Andrew Lloyd interviews Frank Wright.

Statistician/Computational Molecular Biologist at BioSS (Biomathematics & Statistics Scotland).

AL-1: Hi Frank, you work for a group called BioSS, but... What is BioSS and what is its remit in biocomputing ?

FW: Yes, I work for a biomathematics group called Biomathematics & Statistics Scotland (BioSS). BioSS was established ten years ago and is funded a block grant from the Scottish Office Agriculture Environment and Fisheries Department (SOAEFD) and contracts from other public bodies and industry.

The group consists of about 25 statisticians and mathematicians, and computing specialists, plus me - I'm more of a geneticist than a statistician :-). The main activity of BioSS is working with agricultural, biological and environmental scientists based at six research organisations in Scotland that are funded by SOAEFD. This involves collaborative research, and the provision of advice and training covering a wide range of statistical and biomathematical applications.

Biocomputing in BioSS: that's my main role :-) I advise and provide training in molecular sequence analysis to about 200 scientists based at the institutes that we support. I also get involved in collaborative projects requiring considerable sequence analysis.

AL-2: I know you do a fair amount of teaching. Can you tell us about your molecular sequence analysis training courses?

FW: My courses are currently based on software (mainly GCG) running on Unix machines like the UK SEQNET service. This is not unconnected with the fact that the majority of my users have SEQNET accounts :-) I train between 25 and 50 people a year in general molecular sequence analysis. I have a one-day introduction to molecular sequence analysis, which is mainly biocomputing, sequence database usage and simple sequence analysis tasks. After that, there is a one-day follow-up course which concentrates on general sequence analysis methods like alignment, database similarity searching, and protein secondary structure prediction. My third one-day course covers the use of the Web for sequence analysis (most of the material in this course will be integrated into the first two courses). I also present a 2-day course covering phylogenetic tree construction from molecular data (also presented to MRC HGMP users).

Ideally, I would like to spend longer on each topic but it's hard to drag scientists out of their labs to do "computing" all day (charging for courses probably makes this even harder!).

On top of that, most people that I train know little about Unix, X-windows, etc (some describe their pre-course computing experience as "Word6") so some time has to be spent on that. Recently I've been emphasising the use of Netscape for Web-based sequence analysis (as you EMBnet guys have put some pretty good stuff online). In my darker moments, I've even considered that maybe we should be doing some sequence analysis tasks in the MS-Windows environment...

AL-3: What kind of a relationship do you have with SEQNET ? Are they pleased to have you field some of the problems that would otherwise finish up on their helpdesk?

FW: SEQNET provide the main computing environment for the scientists that I collaborate with and train. Alan Bleasby and his co-workers provide an essential service by the provision of both up-to-date data and have also been very quick to load new software. I've tried to deal with most queries originating from "my" 200 SEQNET users but I have occasionally approached Alan. I'm happy to help with the few questions (usually phylogeny) that are passed on to me.

More generally, I feel Biocomputing support professionals have much to gain from sharing expertise.

AL-4: How often do you think about commercialising the product ?

FW: I presume that you mean charging for training courses? We at BioSS have started charging modest fees for training, but at the moment have no plans to charge the full cost of developing courses.

AL-5: You have, over recent years, become Joe Felsenstein's

European ambassador, that is to say that you mirror his PHYLIP site at ftp://ftp.bioss.sari.ac.uk/pub/phylogeny/phylip How did that relationship come about ?

FW: As you can imagine, FTP-ing PHYLIP from Seattle to Europe can be slow-going (understatement!). Joe has set up a quick-to-load Web page on his machine:

http://evolution.genetics.washington.edu/phylip/getmeeurope.html

with hyperlinks to the appropriate files on our FTP server in Edinburgh.

Joe and I first met when he was on sabbatical in Edinburgh while I was a PhD student. I was working on codon usage bias patterns, while he was working mainly on phylogenetic methods. I recall thinking that I should avoid phylogeny as a research area as reconstructing phylogenetic relationships looked a pretty daunting endeavour! Faster computers and new methods (including some developed by Joe and available in PHYLIP and also PAUP*) have since transformed phylogenetic analysis. I've increasingly become involved in phylogenetic data analysis and also phylogeny training. We were happy to load an up-to-date copy of PHYLIP on our Website to make life a little easier for PHYLIP users in Europe (and in thanks to Joe for advice on phylogeny matters).

AL-6: Do you get any time for research ?

FW: BioSS has increased its research output in recent years. Most of my research-related work used to consist of molecular sequence data analyses (mainly molecular systematics) on behalf of collaborators. Now I have a Ph.D. student working on a method of detecting mosaic sequences (due to recombination) in phylogenetic datasets. I'm hoping to spent more time on methodology, but I still find analysing real data is my main interest. I've recently moved my work location to the BioSS group based at the Scottish Crop Research Institute in Dundee, so I will shift emphasis to plant genomes (although not exclusively).

A***L-7***:In the recent general election, the Scots voters rejected every conservative candidate. How long do you suppose it will be before there is an application to join EMBnet from a Scottish Node ?

FW: If the Conservatives had won the last election then my view is that Scottish Independence would have eventually happened! (anything to escape from the stranglehold of what is essentially an English National Party). My guess is that the Scots will be content with a devolved parliament (similar to that of regions of Spain), so HaggisNet is not imminent :-)

Book Review

DNA and protein sequence analysis; a practical approach.

NO. 171 in the Practical Approach Series.
Edited by Martin Bishop and Chris Rawlings.
publ. Oxford University Press 1997.
ISBN- 0 -19-963464-5 Hbk #60.00, 0 -19-963464-7 Pbk #29.95.

Reviewed by Andrew Lloyd, EMBnet Ireland, July 1997

Bioinformatics is widely acknowledged to be one of biology's growth areas. With several formal courses (http://www.ie.embnet.org/other/tut.html) in the subject, both virtual and real, some leading to formal qualifications such as MSc, it continues to surprise me that nobody has written the textbook. Whatever the publisher's intention and with an excess of wishful thinking, I had hopes that Bishop and Rawlings might have delivered it. It is inevitable that, however distinguished and disciplined the editors, a multi-author volume is likely to be a curate's egg - good in parts. However, it requires more discipline than found here for a couple of dozen authors to adopt a uniformly practical approach to promoting the understanding of their subject. For example, several authors refer to Altschul's chapter on sequence comparison, but he does not reciprocate. So his rather good list of effective ways to handle homology searching is compromised by failing to point to the discussion of low-complexity masking in the following chapter.

The most explicit acknowledgement of the subtitle and perhaps the best chapter of the book is given by Nick Goldman on Phylogenetic Estimation. He reviews the methods, points to the pitfalls, discusses the statistical and theoretical assumptions and tries to gauge the success of different ways of deriving phylogenetic trees. The take home messages - never draw trees from proteins, always do global rearrangements, always use maximum likelihood for preference, never darken my doorstep if you use parsimony - are clearly and explicitly stated and should give phylogeny apprentices important guidelines to begin their work. They can question Goldman's certainties when they have more experience.

Other chapters are not so well directed and have less thought for who the intended readers are. One appears to be a practical approach for using a pan-selectionist road-roller to crush a neutralist butterfly. Chapter 6 is a good review of software for the Macintosh but there is no equivalent chapter for PC software. As it is widely accepted that Macs are going fast down the same tube that took VMS, this bias will inevitably bring forward the sell-by date of the book.

And this identifies a key problem with printed books (which Goldman again explicitly states) that they are rather less timely when read than when they are written. If it takes more than a year between the editors signing off the Preface and bound copies appearing in book shops, then it is doubly important, in such a fast moving field, that things are current at the time of writing. There are several failures on this count. Christian Burks bemoans the difficulty of making queries across multiple databases without mentioning SRS - although this software is cited (briefly) elsewhere in the book. One wonders how much of Burks' LiMB 3.0 (now 5 years without update) is still of any value. Gary Williams tells us that "Email servers tend to be the result of academic projects that ... may cease to exist" and then directs us to a source of Una Smith's now well outdated "Biologist's Guide to Internet Resources" which has indeed ceased to exist. Better go to ftp://sunsite.ucl.edu/pub/academic/biology/ecology+evolution/bio guide/. Stephen Altschul says that " the user is generally advised to eschew [the] outmoded methodology" of sliding windows but is happy to commend, as are other authors in this book, clustalV which was superseded by clustalW two and a half years ago.

With my expectation of finding a text book I like the protocols for using specific programs in Chapter 5 (using Staden's xbap as a DNA sequencing tool) and Chapter 12 (using GCG to predict functions from RNA sequences). These step by step instructions are neatly boxed and are a valuable guide to seeing what the software, not famed for the clarity of its documentation, is capable of. In Chapter 11, the issue of identifying genes is approached with a broader sweep by comparing several different programs and reporting the results of a competitive analysis of two test datasets. The tactical and strategic approaches are both valid and useful but to have them in adjacent chapters in the same book can leave the poor reader swamped floundering to readjust their level of understanding.

In the last book I reviewed on these pages

http://www.embnet.org/embnet.news/vol3_3/books.html the publishers had made a brave attempt to stave off the obsolescence of the printed word by pointing to a website for updates. Ironically, the information on the present book that can be found on the OUP website http://www1.oup.co.uk/cite/oup/smj/books/mrktng/nbb0197/Life_Sciences/Biochemistry/DNA_and_Protein_Sequence_Analysis/ is even more out of date than the book itself. The bottom line is - should graduate students spend 45 euro to buy this book - and the answer is probably no. But there is enough sound practice and solid guidance in it that they can and should make a strong case for the departmental library to purchase it.

Courses and Workshops

SRS 5 Workshop

Martin Grabner EMBnet Austria

The workshop took place at the EMBL-EBI in Hinxton, Uk from 22-25 April 1997. Its main intention was to teach programming in the new Icarus language and maintaining an SRS server. Emphasis was also given to using the new features in the World Wide Web interface, in particular the calling of application programs from within SRS such as FASTA, BLAST or CLUSTALW.

Here is a participants point of view

Who does not feel queasy before shifting to a new version of software, which he cannot afford to ignore? Will changes be well documented? Will the format of old data be readable? Will handling be reasonable and user-friendly? ... Questions and fears, which agonise not only the end user, but also system administrators. This is especially true, if an administrator expects to shift from version 4 of SRS to SRS 5, an upgrade that has involved a lot of reorganisation.

Thus the considerable interest and the rapid uptake of places at the SRS 5 workshop, organised within the scope of the EMBnet education and training program at the end of April, was not unexpected. The organisers (Thure Etzold and his SRS-Team) intended to motivate SRS-administrators to integrate special databases into the SRS-library-network. There was also another agenda: - to facilitate and speed up the switch from SRS version 4 to version 5.

Like the SRS-manual in version 4 the new version is tailored to the use of SRS-administrators rather than to end users needs. The SRS5-workshop was based on the excellent online manual of version 5, so that participants did not have to cope with heavy bags filled with paper and could move around without restriction. Furthermore this manual is part of the SRS 5 distribution and can be inspected on every SRS 5 Server over the WWW.

On the first course day old foxes in SRS drew a deep breath, because the installation routine was as easy as with SRS 4. Every course participant was able without difficulties to setup his personal SRS5 WWW-Server for his personal database experiments. Thus theparticipants quickly created their own platform for next days' exercises.

During the following course days participants were familiarised step by step with Icarus (the new SRS parser that is significantly less wing-compromised than its namesake). Its concepts, syntax and programming techniques captivated all who were willing to learn over the next couple of days. Those skilled with regular expressions were at an advantage. Sure, it was no child's play but - compared with the possibilities in SRS 4 - flexibility in programming and design was greatly increased.

It was especially nice to see the debugging commands which should help all autodidacts after the course. Maybe we will unearth unexpected complexity when creating new views and tables. But consider, each sleeping beauty has hidden secrets and powers. Because icarus scripts can be run with the icarus interpreter, the writing of scripts was almost fun.

```
#!/bin/env icarus
$Print:"hello Hinxton\n"
$Print:"Thank you SRS-Team\n"
```

Node Focus

INN

Leon Esterman - EMBnet Israeli node
ISRAELI NATIONAL NODE (INN)
Weizmann Institute of Science
Rehovot, Israel

Introduction.

The Israeli National Node (INN) was created by the efforts of Prof. Marvin Edelman and Leon Esterman from Weizmann Institute of Science (WIS). The Biological Computing Division had long standing experience (from 1978) in serving large groups of biologists both in the Weizmann Institute and in the wider Israeli scientific community. The site was equipped according to specifications of EMBL in Heidelberg concerning the hardware, software and staff. The Ministry of Science and Technology authorized the Israeli National Node (INN) of EMBnet in 1990 and has supported it ever since. From its outset, INN has aimed to provide and foster a national infrastructure for molecular bioinformatics in Israel. This includes providing ready access to up-to-date data banks, centralized equipment purchase and instructive assistance to the scientific, medical and biotechnological community throughout the country. In addition, INN aims to integrate Israeli bioinformatic activities with those of the international scientific community.

INN's activities

National

INN's current national activities include:

1.formatting and channeling incoming data packets from international databanks to its affiliates around the country
2.providing hands-on assistance in getting started to outlying sites
3.instruction and trouble-shooting services for all sites
4.a manned 'hot-line' service during working hours;
5.workshops and courses in databank utilization.

INN services approximately 250 research groups and more than 1500 scientists around the country. INN interacts with and supports other national molecular-biological infrastructural activities, such as: the recently-established Bioinformatics and Genome Resources Center at the Weizmann Institute of Science; the Center for Human Biodiversity at Tel Aviv University, the Center for Protein Structure at the Hebrew University and 11 INN sub-nodes in all the higher educational universities and organizations.

International

In addition to being Israel's gateway to EMBnet and the EBI in Europe, INN also administers an international effort for cooperation in bioinformatics in Poland and East Europe which is supported by UNESCO. Likewise, INN has been awarded a recent EMBnet grant to facilitate international use of Israel's Compugen LTD (Petah Tikva) novel BIOCELLARATOR computer.

User Support and Hotline

Support is provided by electronic mail and telephone. When necessary, support is also provided on-site at the Central Facility or at the user's location. In addition, INN organizes training courses in the use of molecular biology databases and software.

Training program

One of the most important basic function of INN is the training programs for the community of Israeli Molecular Biologists. We are running introductory training programs and workshops:

a. Introduction to the Internet
b. Introduction to the World Wide Web
c. Introduction to UNIX
d. Introduction to molecular biology databases
e. Introduction to GCG

f. Introduction to the BIOCCELERATOR

We have access to a well equipped training computer room and well equipped lecture rooms for the formal lectures on the WIS main campus. Recently, our program has been granted the status of a Msc and PhD training course by the Feinberg Postgraduate School in Weizmann Institute of Science. We have also been running our training sessions on the INN sub-nodes.

INN's personnel

Leon Esterman: manager of INN's Central Facility, Head of the Biological Computing Division
Irit Orr: UNIX System Manager and DAPSAS, INN Scientific Consultant
Steven Becker: INN Scientific Consultant
Rhoda Frydman: UNIX System Administrator, DAPSAS and INN Scientific Consultant

DAPSAS

All programs, servers and databases are running under the DAPSAS (DNA And Protein Sequence Analysis Services) system.

Servers:

WWW http://dapsas1.weizmann.ac.il/bcd/inn.html
SRS http://dapsas1.weizmann.ac.il/srs5/index.html
Bioccelerator http://sgbcd.weizmann.ac.il/

Programs:

GCG, EGCG, Asset, Blimps, Emphasize, Gibbs, Image, Fasta package, Frep, Magick, MakeLogo, MSU, Matindt & Matinspector,Phylip, Primer, Promotor Scan, Readseq, SAPS, Signal Scan , RasMol, Smite, Seqpup, Tmap

Databases:

EMBL, EPD, GCR, GenBank, GenPept, NRL_3D, OWL, PIR, PKinase, ProDom, Prosite, Rebase, SwissProt, TFD, TMbase,
TRANSFAC, IMGT, TREMBL.

Hardware Equipment:

Having started with the IBM - PC (64KB, 10 MB Disk in 1984), our new constellation on which the DAPSAS1 environment runs consists of:

Dec AlphaServer 8200 5/300, Digital Unix 4.0 Operating System ,2x300 MHz DECchip 21164, 1 GB memory, ATM communication interface, 40 GB Tape Loader for backup,Storage of 100 GB (GigaBytes Disk Space) RAID, A very fast processor for ultra-rapid Homology Search consists of: Biocellerator of 320 MMPS (Million Matrix Per Second)

Webwanderer

In this issue we launch a new section which might be subtitled 'Web Resources for Computational Biologists'. We'll try to keep you informed about sites on the WWW that will inform, serve and amuse you.

BIO

The Biotechnology Industry Organization (BIO) would like to invite you to visit our new, improved web site at http://www.bio.org.

BIO is the official industry organization serving and representing the emerging biotechnology industry in the United States and around the globe. The organization's leadership and service-oriented guidance have helped advance the industry and bring the benefits of biotechnology to the people of the world. The BIO web site offers the BIO Marketplace (a searchable database and communications system for over 12000 vendors supporting the industry), BIO Purchasing, the BIO Job Center with current postings of available biotechnology jobs, Biotechnology News resources, Laws and Policies, BIO meetings, Biotech Company profiles and contacts, BIO Education resources and much more useful information for the scientific community.

Joe Bumgarner, Jr. jbumgarner@sciquest.com
SciQuest - http://sciquest.com

ProAnWin

The new version of ProAnWin (Protein Analyst for Win 3.11/95) is now publicly available from IUBio as
ftp://iubio.bio.indiana.edu/molbio/ibmpc/paw.exe
ftp://iubio.bio.indiana.edu/molbio/ibmpc/paw.readme You will need UUDECODE to get the archive with the program.

ProAnWin makes multiple sequence alignments, threads multiple alignment onto known 3-dimensional structure, imports data in all major formats (SWISS-PROT, PIR, FASTA, GCG, Clustal), imports protein 3D structure from Protein Data Bank files (PDB format), transforms protein/peptide activity values (log (A), ln (A), A/K, A+k, etc.), searches linear and spatial sites, conservative and variable changes of specified physico-chemical properties (for example, helical hydrophobic moment), searches linear and spatial sites, having high and low values of specified physico-chemical properties (for example, Kyte-Doolittle hydrophobicity), searches linear sites in multiple protein alignment and spatial sites in protein 3D structure influencing protein activity/property, plots average physico-chemical profile for the family of sequences, plots physico-chemical profiles for protein 3D structure, simulates protein-engineering experiments and predicts protein activity. It has options for automatic mutant generation (to increase or decrease protein activity) and for manual mutant generation, makes protein 3D pictures (mono and stereo) with sites highlighted, has more then 400 amino acid physico-chemical properties; and numerous other features for protein engineers and molecular biologists.

You can use ProAnWin for such analyses as:

- protein structure-function and structure-activity investigations
- designing proteins and peptides with improved activity
- making multiple protein alignments and getting sense from it
- studying phenotype-genotype correlations
- preparation of protein 3D pictures with sites highlighted
- protein features analysis
- comparative protein sequence analysis.

Dr. Alexey Eroshkin eroshkin@vector.nsk.su
Institute of Molecular Biology, Koltsovo, Novosibirsk

ProMSED2

ProMSED2 is a MS Windows application for both automatic and manual DNA and protein sequence alignment, editing, comparison and analysis. ProMSED2 is the enhancement of ProMSED benefitting from users' remarks and suggestions. The program reads many sequence formats and performs automatic alignments, alignment visualization and editing and it allows sequences to be aligned interactively leaving unchanged previously aligned regions. The program has an user-friendly interface. Manual alignment and sequence analysis are facilitated by coloring schemes reflecting amino acid similarity in mutational, physico-chemical and other properties. Although ProMSED was targeted at protein sequences, it can be used on DNA sequences as well. The program provides flexible tool for sequences alignment, analysis, visualization, edition and presentations.

Available from:
ftp://ftp.ebi.ac.uk/pub/software/dos/promsed or
ftp://iubio.bio.indiana.edu/molbio/ibmpc/promsed2.exe or
ftp://iubio.bio.indiana.edu/molbio/ibmpc/promsed2.readme
The program has many features including: loads several sequence families in different windows; adds sequences to existing alignment, combines sequences from various files; makes presentation quality color and black-and-white prints of complete alignment or any selected block; calculates sequence similarity of complete sequences, of any selected sequence subset or of marked block in % and in PAM250 units; prints sequence similarity matrix; displays conserved and semiconserved positions; has many amino acid coloring schemes aimed to facilitate

Dr. Alexey Eroshkin eroshkin@vector.nsk.su
Institute of Molecular Biology, Koltsovo, Novosibirsk

ProAnalyst

ProAnayst: DOS version of ProAnWin with additional functionality (single and multiple sequences analysis, profiles analysis, combinatorial libraries; design of protein engineering experiments)

ftp://iubio.bio.indiana.edu/molbio/ibmpc/panalys1
ftp://ftp.ebi.ac.uk/pub/software/dos/proanalyst

The many useful functions of ProAnalyst include: data conversion from several protein sequence formats (FASTA, SWISS-PROT, PIR, CLUSTAL).; flexible visualisation of protein 3-D structures with sites highlighted; multiple linear regression analysis of Structure-Activity relationships, discriminant analysis and ANOVA; alphabetical and physico-chemical analysis of protein features variations (in 1D and 3D structures; investigation of physico-chemical factors related with activity or property changes in mutant proteins; searching motifs in Combinatorial libraries (peptide, phage-display libraries, etc.) with motif mapping on the target protein; mapping results on 3D structure and sequences.

Dr. Alexey Eroshkin eroshkin@vector.nsk.su
Institute of Molecular Biology, Koltsovo, Novosibirsk

bio1inf0rm

A new commercial newsletter, called bio1nf0rm, for workers in the field of bioinformatics has been launched. It is due out 24 times at a cost of 585 USD a year. You can register for six weeks worth of free issues by filling in the form on their web site:

http://www.bioinform.com/

Bioinformatics Courses

A re-organized a listing of syllabi of bioinformatics courses offered worldwide can be found at any of:

- http://www.techfak.uni-bielefeld.de/bcd/Curric/syllabi.html
- http://merlin.mbcr.bcm.tmc.edu:8001/bcd/Curric/syllabi.html

- http://www.biotech.ist.unige.it/bcd/Curric/syllabi.html

Georg Fuellen fuellen@techfak.uni-bielefeld.de
Research Group in Practical Comp. Science, Univ. Bielefeld, DE

Sequin

A new version of Sequin (v2.14), the sequence submission/ editing tool from NCBI for all platforms is now available from: http://www.ncbi.nlm.nih.gov/Sequin/

From here you can find the latest developments, new Frequently Asked Questions, and the most recent version of the help documentation, as well as a new and much improved version of the Sequin Quick Guide, a step-by-step description of how to prepare your Sequin submission.

Network Entrez, an NCBI tool for accessing bibliographic, sequence, and structure records is now fully integrated into Sequin. PowerBLAST, a version of the client for the popular BLAST software for sequence comparisons, is now available from within Sequin. You can have Sequin calculate and display the alignment between a sequence in the record and another sequence in a file. Records can now be viewed with two new Display Formats. Summary format shows the range of any sequence alignments in the record. Sequence format shows the sequence(s) in the record along with any associated features.

B.F. Francis Ouellette francis@ncbi.nlm.nih.gov
NCBI http://www.ncbi.nlm.nih.gov/

New BLAST Service

A new BLAST service for the recently released microbial genomes from The Institute for Genomic Research (TIGR) is now available from the BLAST home page at:
http://www.ncbi.nlm.nih.gov/BLAST/
or directly from:
http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-tigrbl

These unfinished genomes are not yet available in GenBank or Entrez. The genomes are searchable via TBLASTN (a user's protein query vs. a 6-frame translation of the microbial DNA sequences) using the new gapped BLAST algorithm (Altschul et al., 1997, submitted).

B.F. Francis Ouellette francis@ncbi.nlm.nih.gov
NCBI http://www.ncbi.nlm.nih.gov/

TRANSFAC 3.2

TRANSFAC 3.2 is available now at: http://transfac.gbf.de. TRANSFAC is a database of eukaryotic transcription factors and their binding sites. It consists of six cross-linked tables: - SITE - CELL - FACTOR - CLASS - MATRIX - GENE

It is also cross-linked with TRRD (Transcription Regulatory Region Database) and COMPEL from the ICG, Novosibirsk (N. A. Kolchanov, A. E. Kel). It contains numerous cross-references to external databases such EMBL, SWISSPROT, PIR, FLYBASE, EPD, and PROSITE. It also cross-references to PDB. And has a transcription factor classification sytem: http://transfac.gbf.de/TRANSFAC/cl/cl.html

On the TRANSFAC server, you will find also the sequence analysis programs - PatSearch - MatInspector - SaGa - FastM and Thure Etzold's SRS5 with a large collection of databases.

Edgar Wingender
Thomas Heinemeyer thh@gbf.de
http://transfac.gbf.de/Staff/thh.html
Ges. f. Biotechn. Forsch. mbH, Braunschweig, DE

The Virtual Lab

http://www.novo.dk/vl/index.htm

The Virtual Lab is a free educational resource on the internet, sponsored by Novo-Nordisk. It teaches more about genetic engineering and how actual researhers design new medicines and proteins. It uses shockwave to provide unique multimedia lessons geared to upper level high school students and college students. If you don't have shockwave or hate browser plug-ins, the virtual lab offers a pure HTML version which can be viewed on any browser!

Tor Kristensen tor@araneum.dk
Araneum A/S http://www.araneum.dk/

The pdf format

Robert Herzog
Belgian EMBnet Node

Our regular readers may have noticed that, in addition to the usual html and PostScript formats, the June issue of embnet.news was also made available as a pdf file. This format will from now on be part of the "official" channels of distribution of our newsletter. Back issues of embnet.news will also be readable in pdf format from most distribution sites. If you are not familiar with the many scientific publications now readable online, where the pdf format is established as a de-facto standard, the pdf format may need some introduction.

The PDF format has been conceived by the Adobe company. The full description of the PDF format is a book of 394 pages. It can be found (as a PDF file!) at http://www.adobe.com/supportservice/devrelations/PDFS/TN/PDFSPEC.PDF. If you are working on a Macintosh or printing on a departmental printer, it is very likely that your printer supports PostScript. PostScript is well established as the standard of reference in the professional pagesetting and printing circles. Faced with the tremendous growth of the Internet, the Adobe company came up with a format that allows the presentation of all kinds of publications on the screen exactly as they get printed. This was conveived as kind of an extension of the PostScript concept.

The "Portable Document Format" was born. Like PostScript it supports scaleable fonts and graphics objects in a nice compact format, and is widely used as a web readable format for many well known scientific publications like Science, Journal of Molecular Biology, EMBO Journal, Journal of Biological Chemistry and a multitude of others. It is also imposing itself as the standard for high quality online documentation. Take a look at the web site of Adobe to convince yourself [1].

The production of the pdf formatted files can happen thanks to various software modules produced by Adobe [2]. Adobe Writer can serve rather like a printer driver, to produce pdf files directly from within any microcomputer application that can send its output to a printer. Adobe also produces software to convert other formats into the pdf format. The first and the most common of these converters if Adobe Distiller. This software takes any PostScript file and produces the corresponding pdf file. As a free copy of Adobe Distiller came with the PageMaker software we use to create embnet.news, the conversion of the PostScript file and its exposure on the web is only a matter of minutes of additional work. Among the nice aspects of this transformation is the size of the end product : the pdf format is distinctly more compact than PostScript itself.

Producing pdf files needs some commercial software, but reading it can be done with freely available high quality software coming from Adobe. You need the software named Acrobat Reader or Acroread. It is currently available for all environments, and for many UNIX flavours (Solaris, HP-UX, IRIX, OSF/1 [3], Linux and others)..

In order to read a pdf file, the simplest technique is to declare Adobe Acroread as a helper application in your favourite web browser. Alternatively, you can download the pdf file with the classical ftp technique and load it off-line in the Acroread program. Using the Acrobat Viewer plug-in or helper application, the documents can be displayed on the screen at any comfortable size, and it can printed at the best quality level, even on systems that don't have the PostScript functionality. High readability and high printed quality for everyone anywhere, isn't this nice modern technology ? In its most recent versions, Acroread incorporates a "find" function [4], which puts you in the unprecedented position of being able to read a paper as if it were printed, but with an additional assistant to find any word instantaneously..

The pdf format is with us to stay, as it blends beautifully with the web technology. While the production of pdf needs commercial tools, its reading and the production of high quality printed output is free for anyone to use. If this looks like a commercial, pardon me; it is not my intent. Embnet.news is proud to use this kind of technology, but has no financial support whatsoever from the Adobe company.

NODE NEWS

Switzerland

Starting in the spring of 1996, we have been faced with the difficult task of following Reinhard Doelz. We would not expect anyone to forget his contribution, and are at least trying to restore the services he was providing to the Swiss community and establish a fully functional EMBnet node.

After a year of borrowing disk space and CPU time on a small Sun Ultra1 server belonging to the Computer Department of the Swiss Federal Polytechnic in Lausanne, we should be ready to run GCG on a more appropriate machine (a 4-processor SGI Origin 2000) in July. This machine is operated in collaboration with the Schools of Chemistry and Pharmacy of the University of Lausanne, and will provide services for molecular modelling as well as sequence analysis. We are looking forward to a fruitful collaboration, which should benefit researchers from the entire country.

On an experimental basis at first, and now in full production, we have a BLAST server running the Washington U. (Warren Gish) version of this popular search algorithm.

[1] http://www.adobe.com

[2] Public domain software to convert PS to pdf is available, like the ps2pdf converter, part of the Aladdin Enterprises Ghostscript package. Look at most ftp sites distributing Ghostscript.

[3] The version for Acroread for OSF/1 is brand new. It appeared on the ftp server of Adobe on July 16 this year. See the following URL: http://stkwww.fys.ruu.nl:8000/~hogeveen/digital_acrobat.html for an activists page of OSF/1 users insisting in getting Acroread for their platform.

[4] Look for the "Acroread with Search" software, available for most microcomputer platforms.

[5] A public domain PDF reader is under developent : xpdf. You can get a copy for many unix platforms at ftp://ftp.andrew.cmu.edu/pub/xpdf

Unlike the NCBI version, WU-BLAST produces gapped alignments, and its sensitivity is thought to be slightly higher. BLAST was implemented on a dedicated dual-processor Pentium Pro machine operating under Solaris x86, and accesses weekly updated protein (Swiss-Prot and our own non-redundant database) and nucleic acid (EMBL, non-redundant, dbest, dbsts, REPBASE) databases. The server can be reached through our Web pages (EMBnet-CH or ExPASy), or by any standard BLAST client as "pcisrec-d402d.unil.ch".

We ran our first EMBnet sequence analysis course in early March with overwhelming success: we had over 120 registrations, and were able to accommodate only 45 students. This first group seems to have been quite satisfied with the course, and many of the students are now using our services on a regular basis. The one-week, full-time course will be repeated in September to absorb the backlog. We hope that from next year on we will also have time to offer more advanced courses in topics such as molecular modelling, phylogenetic analysis, genome analysis, or modern techniques in sequence comparison (HMMs, profiles and the like). Help from the EMBnet ET Committee will be much appreciated!

The Swiss EMBnet node is still very much a work in progress, and I am sure that more news (mostly positive, I hope..) will be reported in upcoming issues. Stay tuned!

Germany

The German EMBnet node, GENIUSnet, at the German Cancer Research Centre (DKFZ) in Heidelberg continues to develop its resources and services. Here is an update on some of the latest changes.

Hardware

The GENIUS computer has recently been upgraded to a CONVEX SPP 1600 configuration with 16 processors (up from 8). In addition, a 40 MB WiN connection was installed at DKFZ, and has significantly shortened response times in network connections to or from our site. We are currently testing a COMPUGEN bioccelerator which we expect to speed up various database search programmes available here.

WWW site

The GENIUSnet homepage at: http://genius.embnet.dkfz-heidelberg.de:8080/menu/ has been brushed up. From this page, all visitors have access to our WWW interfaces for the Genomic Database (GDB) and the Online Mendelian Inheritance in Man database (OMIM), as well as to the recently installed Sequence Retrieval System (SRS5) covering most of the databases held at DKFZ. In addition, all registered users can access WWW2HUSAR, the WWW interface to our HUSAR sequence analysis package. HUSAR news, a WWW2HUSAR demo session, and information on our local services are available for everyone.

Software

New programs and algorithms are continuously being added to the GCG/HUSAR environment established at our site. Information on the latest entries can be found in the above mentioned HUSAR news pages.

ICGEB

In the last 6 months or so, we have installed 3 web-based servers for DNA analysis. They include plotting various DNA parameters along the sequence and, more importantly, the calculation of curvature propensity based on either bendability data, or static dinucleotide geometry. They are found on the newly reorganized ICGEBnet menu page on the updated ICGEBnet WWW home page, at the following site, under "DNA tools": http://www.icgeb.trieste.it/net/

We have chosen rather patriotic names for the servers, such as "bend.it" and "curve.it". There are a number of explanatory files involved, i.e. the servers are meant to contain the necessary background materials and references in hypertext form. The new ICGEBnet page (installed in June) has pointers to our older but still on-going projects, like SBASE and the P450 directory. A new addition to the page is the GSM, WWW menu developed at CAOS CAMM. This was installed during a recent EMBnet tech manager visit by Harco de Hilster and Jack Leunissen.

On the hardware side, we have recently expanded the memory on two of our servers to 128MB and added 18 GB of disk space. Something else that has been keeping us busy is a hacker attack we had in February. We were closed down for a week and have been re-registering our users since then, via snail-mail. It was the kind of fun we could do without.

Apart from this, we are enthusiastically preparing the forthcoming EMBnet course 1-7 September, 1997.
http://www.icgeb.trieste.it/net/netcourse.html

SEQNET (UK)

Howard Sherman, the head of both the SEQNET and Chemical Databank Service has recently retired. We wish him well. All SEQNET management matters should now be directed to Alan Bleasby (ajb@seqnet.dl.ac.uk), the node manager.

A perl script has been developed to perform a fast search of a sequence against PROSITE. This can be downloaded from http://www.seqnet.dl.ac.uk/prositesearch

Note that you will need both perl 5 and SRS 5 to run this.

Finland

Although the node manager in Finland has changed from Heikki Lehväslaiho to Erja Heikkinen, her address and the fax number remain the same as Heikki's, however Erja's telephone number has changed to +358-9-4572433 The Finnish EMBnet WWW server is at http://www.csc.fi/

EMBO, as requested by the Academy of Finland, evaluated the quality of the Finnish molecular biology and biotechnology research last year. While the evaluation committee (in their report in January 1997) found the quality by international standards high in general and even excellent in some cases, they recommended a considerable investment in developing the areas of bioinformatics and structural biology research. Finland has been lacking education and expertise in both disciplines and it is a great challenge to start building the facilities. CSC is a natural participant in this development as the provider and maintainer of a metacomputing environment and a fast national network. This is a wonderful opportunity to create and vitalize connections with academia. The only limits are set by resources both human and financial.

As the biological/biomedical research culture has moved from descriptive molecular research towards understanding the function of molecules both computational and data management have needed to evolve. So far CSC has got by with only one bioscience application specialist, but now it has become obvious that more are needed to properly serve all our biocustomers in academia (eight universities) and private institutes. We also organize courses on various computational topics (including specific courses tailored to meet customers' wishes) and more personnel would broaden the options in this area. With the current speed of data accumulation it is impossible for one person to master all fields classified under biosciences!

CSC is owned by the Ministry of Education and the current science policy in Finland is dynamic so there is hope! Three project researchers in structural biology (at least one of them a post-doctoral fellow) will start their research at CSC by the end of the year. Even though not directly in customer service they will help strengthen the bioscience support available at CSC. All in all, bioscience activities are expanding at CSC. Our new scientific director, Olle Teleman, is the head of a structural chemistry and biology research group at Technical Research Centre of Finland (at the section of Biotechnology and Food Research).

Poland

The number of users increased to more than 500. Our Web page (http://www.ibb.waw.pl) has been completely rewritten to facilitate quicker access to the most frequently used Internet services and to provide on-line manuals to the software packages. An SRS 5.0 server was installed and is freely available from our Web page. Updated versions of TREMBL are now available to our users. The "Protein Sequence Analysis" Web browser based on GCG programs has been developed and is at the testing stage.

The EMBnet Nodes

National nodes:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Special nodes:

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

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

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

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

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

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

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

Associate nodes:

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

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

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

Dear reader,

If you have any comments or suggestions regarding this newsletter we would be very glad to hear from you. If you have a tip you feel we can print in the Tips from the computer room section, please let us know. Submissions for the BITS section are most welcome, but please remember that we cannot extend space beyond two pages per article. Please send your contributions to one of the editors. You may also submit material by Internet E-mail to:

emb-pub@dl.ac.uk

You are invited to contribute to the LETTERS TO THE EDITOR section.

If you had difficulty getting hold of this newsletter, please let us know. We would be only too happy to add your name to our mailing list. This newsletter is also available on-line using any WWW client via the following URLs:

The Online version, (ISSN 1023-4152) :

- ***http://www.uk.embnet.org/embnet.news/vol4_2/contents.html***
- ***http://www.be.embnet.org/embnet.news/vol4_2/contents.html***
- ***http://www.no.embnet.org/embnet.news/vol4_2/contents.html***
- ***http://www.ie.embnet.org/embnet.news/vol4_2/contents.html***

A Postscript version (ISSN 1023-4144) is available. You can get it by anonymous ftp from:

- ***ftp.uk.embnet.org in the directory pub/embnet.news/***
- ***ftp.be.embnet.org in the directory pub/embnet.news/***
- ***ftp.no.embnet.org in the directory pub/embnet.news/***
- ***ftp.ie.embnet.org in the directory pub/embnet.news/***

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

- ***ftp.uk.embnet.org in the directory pub/embnet.news/***
- ***ftp.be.embnet.org in the directory pub/embnet.news/***
- ***ftp.no.embnet.org in the directory pub/embnet.news/***
- ***ftp.ie.embnet.org in the directory pub/embnet.news/***

Back issues are available at most of these sites.

Publisher:

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

Editorial Board:

embnet.news
Vol.4, No.2, 1997
31 July 1997
ISSN 1023-4144