

Editorial

Welcome to embnet.news. In the first 1998 issue of issue we have most of our usual features, with perhaps a broader geographic range of contributions than usual. The InterviewNET section is a chat across the width of Europe from NW to SE, whilst the EMBnet Node-In-Focus is SANBI from the other hemisphere. Gonzalo Claros from Malaga has written about a graphic tool for enzyme kinetics, whilst Koen Cuelenaere from the CAOS/CAMM Centre in Nijmegen has written "Yet Another GCG to WWW Interface". Perhaps particularly significant is the report on the imminent shutdown of GDB - the Human Genome Database. This is coupled with an article by Bruno Gaeta, from ANGIS in Australia, on the consequences of this decision for genome researchers world-wide.

Bioinformatics is often notable for the way in which the beneficiaries, such as academics and the pharmaceutical industry, are able to decouple themselves from the necessity to fund the resources on which they rely. In 1996 it transpired that in a very short time scale, Swissprot, the arguably best annotated, most widely integrated and most carefully curated protein sequence database would effectively stop being produced. After a petition from hundreds of users worldwide, each of whom recognised the value of the resource and relied on it for their research, SwissProt was reprieved. However, it is not certain if a long term plan is in place for ensuring its continued existence. Even EMBL, the DNA database, is largely dependent on funding from the European Union and must submit a project proposal and compete for scarce resources with other bioinformatics projects in a peer review process.

Those node managers, who are obliged from policy or necessity to charge users for access to bioinformatic services, can testify to how difficult it can be to get academics to translate how much they "value the resources provided" into a cheque. Much of the best bioinformatics software has been written to fulfil a locally compelling need and has then been made publicly and freely available. Perhaps this generosity gives a false "there *is* such a thing as a free lunch" message about bioinformatics in general. One positive aspect about software going commercial is that users can then legitimately demand that a user-friendly interface and effective documentation is provided.

For better or worse things in our bioinformatics world things are changing as the accountants begin to weigh in. It's no longer so easy to help run a training course outside of one's

own institution, or to work up some first rate software on company time and then give it away, or to rely on a resource which is wholly supported by a funding agency in a single country.

Let us, in the future, try to ensure that the central resources (we all have a good idea what they are) have a long term future and a coherent development plan for all users regardless of their nationality, affiliation or competence.

The embnet.news editorial board:

Alan Bleasby
Rob Harper
Robert Herzog
Andrew Lloyd
Rodrigo Lopez
Peter Rice

Termination of GDB Project

Dear Colleagues,

The Genome Database project (GDB: <http://www.gdb.org>), which provides human gene mapping data to human genetics researchers from its base at Johns Hopkins School of Medicine, will soon be ceasing its operations. This action is a consequence of the decision by the project's primary funder, the U.S. Department of Energy's Office of Energy Research, to discontinue GDB funding in order to focus its informatics resources primarily on the results of the sequencing phase of the Human Genome Project. Termination of the GDB project is expected to be completed by July 31, 1998. The

Contents

Editorial	1
Termination of GDB project	1
R.I.P. GDB ?	3
The Joy of Sed and Awk	4
GCG Plugin for WWWGMS	6
Lines & Kinetics, a graphical tool for linear regressions and enzyme kinetics	7
Crackers and other junk	10
Upcoming Conferences and Web sites	11
INTERviewNET - Sandor Pongor of ICGEB	12
Node Focus - SANBI, EMBnet in South Africa	13
DEJAnews	14
Node News	19
The EMBnet Nodes	21
embnet.news information	22

database will continue to be made available to the scientific community after that time, but further change to its content will cease.

GDB was initially created at Johns Hopkins in 1989 by the Howard Hughes Medical Institute to capture the data from the Human Gene Mapping Library project at Yale. In 1991 the responsibility for funding the project was assumed jointly by the Department of Energy, the National Institutes of Health, and the Japan Science and Technology Agency. A series of mirror sites in many countries helped to ensure international access to the data.

One of the first major accomplishments of the project was to capture, in electronic form much of the information about human genetics and gene mapping accumulated by the scientific community during the two decades prior to the Human Genome Project. These data were reviewed and edited by a worldwide group of volunteer scientists, to assure quality, under the auspices of the Human Genome Organization (HUGO). The data included information on human genes, probes, clones, and allele frequencies, and represented the first significant attempt to collect the data necessary to get the Human Genome project under way. During its lifetime GDB provided informatics support for a succession of HUGO-sponsored Human Gene Mapping meetings and Single Chromosome workshops, hosted in a variety of locations around the world. These meetings brought together researchers from many countries to piece together integrated chromosome maps from numerous separate experimental results; the results were then made available to the research community via GDB.

GDB pioneered the use of the World-Wide Web as a tool for dissemination of bioinformatics data to the community, which has now become routine. The project was also one of the first to deploy a Java client application (Mapview) to graphically display database query results. In recent years novel algorithms were developed to integrate multiple maps into a single comprehensive chromosome model which could be searched and displayed.

By the mid-1990's small-scale mapping was giving way to high resolution whole-genome mapping at a small number of centers, as part of the mapping phase of the Human Genome Project. More recently the focus has shifted again to high-throughput sequencing. As the heyday of traditional mapping has faded, so has the perceived need for a large community database project focussed on maps. The decision has therefore been made to terminate the GDB project. The database will continue to be made available to the scientific community at the same Web address for the foreseeable future, though most curation and data acquisition activities will cease before long. It is expected that the international mirror sites will make their own decisions regarding the longer term maintenance and supply of the final GDB release.

Human Gene Nomenclature will continue to be curated by Dr. Sue Povey at the University College of London (<http://www.gene.ucl.ac.uk/nomenclature/>). The OMIM database, which, since 1995, has had no formal connection with GDB, will not be affected and will continue to be available through NCBI (<http://www.ncbi.nlm.nih.gov/omim>).

The Computational Bioscience Section (<http://compbio.ornl.gov>) of the Oak Ridge National Laboratory has agreed to maintain the servers for access to the current copy of GDB after the project ends at Johns Hopkins University. This Section is headed by Dr. Ed Uberbacher, the developer of the widely-used GRAIL eukaryotic gene-finding program. Oak Ridge is the coordinating site for the Genome Annotation Consortium (GAC, <http://compbio.ornl.gov/CoLab>) project, which is developing software and data systems to assist in assembling and annotating the results of human genome sequencing on an ongoing basis.

Having GDB available locally will facilitate the GAC's integration of map data with sequence to produce the annotated reference genomic sequence, which is the ultimate goal of the Human Genome Project.

I would like to take this opportunity to applaud the GDB project staff and management for their hard work and creative effort over the 8 year history of the project. I would also like to offer my deepest thanks to the past and present members of both our Quarterly Review Committee and our International Scientific Advisory Committee for their advice and support over the years and to thank our host institution, the Johns Hopkins School of Medicine.

We hope that this decision does not cause any great inconvenience to members of the scientific community. Please send comments or concerns to comments@gdb.org, or fill out the comment form at <http://www.gdb.org/shutdown/shutdown.html> which allows comments to be anonymous. Please be as specific as possible in describing any use of GDB in your research for which no equivalent resource currently exists. These comments will be forwarded to myself and the funding agencies.

Funding agency representatives can be contacted directly at the following addresses:

Department of Energy:

Daniel.Drell@oer.doe.gov, Marvin.Frazier@oer.doe.gov,

National Institutes of Health:

Lisa_Brooks@nih.gov

The GDB staff can be contacted collectively at gdbstaff@gdb.org.

Sincerely,

Stanley Letovsky, Ph.D.

Director, Genome Database

letovsky@gdb.org

R.I.P. GDB?

GDB shutdown: the view from an Australian node

Bruno Gaëta, Australian National Genomic Information Service, <http://www.angis.org.au>

The US department of Energy recently announced the termination of its funding of GDB, the Human Genome Data Base, in order to concentrate on the funding of the sequencing phase of the human genome project. This decision is likely to result in a shutdown of GDB by the middle of 1998. The database will remain available over the Internet, but no longer be updated.

Database history

GDB was intended as a public repository for human genomic mapping data and supporting information. As such, it provides access to most of the data accumulated on human mapping markers, their organisation into genetic and physical maps, their known polymorphisms, the experimental evidence for their localisation, as well as related citations and contributing laboratories and researchers. GDB does not contain any sequence information, instead relying on links to Genbank and other sequence databases to bridge the gap between mapping and sequencing data.

The first versions of GDB used a text interface built around a series of menus and forms and accessed over the Internet through telnet. This version of GDB suffered from a number of shortcomings. Many users found the text forms difficult to use, and in the absence of specific training, were discouraged from tapping into the wealth of information available. The fully text-based database was especially limited in its representation of maps. Another problem was that any changes to the database, including the addition of new markers, had to be processed by database administrators. This often resulted in substantial delays between the submission of new data and their appearance in the database.

Version 6.0 of GDB, introduced in 1996, saw a radical change in database format. The information was organised into a more object-oriented model better suited to the storage of biological data and relationships. Furthermore, the new version introduced community curation, therefore allowing researchers to contribute information directly to the database and decreasing submission time. But the most important change as far as many users were concerned was the introduction of a more user-friendly worldwide web interface, which made GDB more accessible to 'casual' browsers, and allowed the inclusion of graphical representations of maps,

originally through a helper application, and subsequently through a Java applet. GDB version 6.0 still had to be queried through complex forms, but subsequent versions saw the development of simple query tools such as 'Search by keyword' and 'Find a gene'.

Database structure

GDB uses a relational database model. This can deal more economically with its wide range of data types than the flat file format used by Genbank and other major sequence repositories. Since the complexity of this model makes a simple distribution of the database impracticable, a network of specialised mirror sites (nodes) has been set up to replicate the full database structure around the world in order to minimise network delays for local users. Database updates were originally sent on tape, but GDB switched to a network-based automatic incremental update strategy which minimised the amount of network traffic required. Unlike the major sequence databases, which have to be downloaded in their entirety several times a year, new data in GDB need to be sent to a node only once where it is merged into the rest of the database.

The relational database approach allows the formulation of complex queries which a flat file database would be unable to process, but with a significant trade-off in the speed of information retrieval. As the amount of data stored in GDB has increased, the long time taken to obtain the answer to a moderately complex query has been a major reason for dissatisfaction with the database and an impairment to its use.

GDB in Australia

ANGIS, the Australian EMBnet node, has been a GDB node since 1992 providing access to the database for a diverse community of users around Australia, New Zealand and beyond. Another Australian GDB node is located in Melbourne at the Walter and Eliza Hall Institute of Medical Research. ANGIS has not only maintained a mirror of GDB, but has also produced documentation and educational materials to assist in the use of the database, it included the querying and browsing of GDB in the training courses run for the users of the service. ANGIS is accessed by scientists working in all fields of molecular biology, not just human genetics. The impact of a potential shutdown of GDB on this user community is difficult to predict. An informal survey of some ANGIS users, working on the human genome and genetic diseases revealed that most of them considered the information stored in GDB essential for their work, but found its retrieval slow and complicated.

The ANGIS users who supported GDB viewed the database as an essential resource bringing together a whole range of diverse information. Typical uses included looking up a

particular author's laboratory, and retrieving information about markers identified there, finding out the allele frequencies and standard sizes for a set of markers prior to a genetic linkage experiment, identifying a marker cited in the literature by looking up its aliases, summarising all the known information about a particular human gene, including its chromosomal location, its polymorphisms, its mouse homologs and relevant literature citations and other uses. Several users found GDB invaluable in the preparation of genotyping experiments involving microsatellite markers and RFLPs, especially for retrieving PCR conditions and primer sequences. In many cases, GDB was perceived as the sole source for the required data. In other cases, when the information was known to be available from an alternate source such as Genethon or CEPH, GDB was still the preferred option because all the data were available in one place and cross-referenced. This centralisation of the information in one database was seen as a real timesaver.

The major criticism of GDB was that the database organisation and user interface were confusing and catered more to a computer scientist's view of the data, rather than to the type of questions asked by biologists. The 'simple search' and 'Find a gene' tools were by far the preferred method of querying the database. The more complex query forms for retrieving specific types of data, which used to be the only way of accessing the database, were deemed too complicated to use by all but the most experienced users. The slowness of the database also came under criticism, especially the time needed to get maps from the database and display them. For this reason some users had given up on using GDB for most tasks. They only searched it for data unavailable anywhere else, such as polymorphism and gene nomenclature information.

Towards the future?

The decision of the DOE to terminate the funding of GDB reflects the change of trends in human genome research, away from genetic linkage analysis, and towards the ultimate goal of a complete genome sequence. However there is still going to be a need for mapping data, especially in the study of human genetic variation and disease. Such research is prevalent in Australia, and GDB is often perceived as essential support in these quarters. If the database is no longer maintained, alternative sources of genome mapping information would have to be well publicised. It would have to adequately replace the centralised repository provided by GDB.

From a database technology point of view, GDB is one of the most advanced public databases currently available to the biologist. Its architecture allows a large amount of data of different types to be stored efficiently and to be searched using complex criteria. The strategy used to update nodes is sensible and minimises the amount of network traffic

required. However, this technological advantage has come at a price when it comes to the usability of the database by biologists. The organisation of the information into 'objects' and 'classes' makes better sense to a database administrator but may not reflect the way in which biologists think about the data. This has been a significant impairment to the use of GDB in the past, and one that the database creators have been addressing in recent versions with the introduction of simple query tools.

The GDB interface is still a work in progress. Its creators have been steadily improving access to the database and building a more user-friendly front end. Paradoxically, these constant improvements may have been a hindrance to regular users who have had to keep up with the changes in the system. The changing interface also made more difficult the production of documentation and teaching materials to educate new users. It is a shame that GDB may be shut down before these changes are complete and the more user-friendly database has a chance to gain fuller acceptance among the genome community. It would be a greater shame if the demise of GDB results in more difficult access to the data it contains, on which a lot of research of medical importance depends.

Note: A WWW page has been set up at <http://www.gdb.org/shutdown/shutdown.html> to collect user feedback on the closure of the database and forward it to the funding agencies. It can be used to register comments or concerns, especially if your research depends on GDB.

Bruno Gaëta is in charge of education services and user training at ANGIS. He has been writing documentation and providing user support for GDB users on ANGIS since 1993. He has written several tutorials documenting the use of GDB, some of which were made available to the international GDB community. He has also co-written a chapter on GDB in the ANGIS Bioinformatics Handbook which is used for training courses all around Australia.

The Joy of Sed and Awk

Colin Semple, Genetics Department, Trinity College, Dublin

The rapid growth of bioinformatics research has increased the demand for programs that recognise patterns and manipulate text in UNIX files. The specialised, and often rather involved operations, required by the researcher are not catered to by the standard repertoire of UNIX commands. For those of us in this position the solution is usually to write our own software in our favourite programming

language, assuming that learning Perl is still near the bottom of our things-to-do list. However, another sometimes ignored alternative is to use the commands of the UNIX utilities sed and awk.

Sed is a non-interactive text editor which can perform the same tasks as interactive editors such as vi or ex. Awk is a programming language which can do everything sed can do. It also does floating point arithmetic. Both sed and awk commands can be entered at the command line or in an executable file or script.

Sed commands can be useful if you want to perform some kind of text editing task on a number of different files. If you are processing files with a shell script that executes several programs serially it may be necessary to amend the output of one or more of the programs. For example, if the first program executed precedes its output with the lines "Output from first program.\n Written by J. Bloggs 1997.\n" you could remove the troublesome first two lines by inserting the following sed command between the commands to execute the first and second programs.

```
sed '1,2d' outputfile1 > otherfilename
```

This may look familiar as sed addresses text in the style of the text editor ed. However scripting in sed can be rather less straightforward, precisely because it is syntactically quite different from computer languages. This is not a problem with awk which follows similar conventions to C for loops, conditionals and built-in string manipulation functions and resembles Perl in its informal variable declarations. When scripting in sed one also has to wrestle with its primitive storage scheme when using scripts that need more storage.

The conventions of awk commands are the same whether entered at the command line or in a script. These require that the command begins with the word awk, followed by the awk command(s) enclosed in single quotes (or the name of the script which contains the commands) and the name of the input file. Commands always refer to patterns and/or actions. Patterns are descriptions of text that awk will recognise and are enclosed in forward slashes. They may also include logical and arithmetic operators. Actions refer to the operations to be performed on the input and are enclosed in curly brackets. If an action consists of more than one statement they must be separated by semicolons. Unless otherwise instructed, awk assumes that the data in a file are held in columns, called fields, delimited by white space and rows, called records, delimited by a carriage return or newline. Fields are known to awk as a dollar sign followed by a number above zero. The entire current record is referred to as \$0. Thus the following command asks awk to print the first two fields in the current record (by default to standard output) if it finds the string "RNA":

```
awk '/RNA/{print $1, $2}' filename
```

I find command line awk invaluable for certain formatting problems I regularly encounter. For example, I have had no success in persuading the vi editor to substitute newlines for tabs. The following awk command achieves this effortlessly, writing the output to a file called otherfilename:

```
awk '{gsub(/\t/, "\n"); print}' filename > otherfilename
```

A more general application of command line awk is in standardising irregularities in spacing between fields. I am often faced with files where the first few data fields are separated by tabs but the ones following are separated by some number of tabs or spaces. As a result one can have problems processing the file when using programs that are fussy about the format of their input. The following command converts all field separators composed of spaces and/or tabs to single tabs (which as literal strings must be enclosed in double quotes).

```
awk '{print $1"\t"$2"\t"$3"\t"$4}' filename > otherfilename
```

The arithmetical capabilities of awk can also be very handy. When I want to exclude/include file entries satisfying a simple expression, say that a number in the first field is greater than zero, I have a couple of alternatives. I can either write a one loop program, compile it and run the file through it or use the following awk command which evaluates the expression and writes all three fields, separated by underscores, to another file.

```
awk '$1 > 0 {print $1"_"$2"_"$3}' filename > otherfilename
```

As awk is a fully fledged language it allows loops to perform iterative tasks with built in functions such as getline, which is used to read the next line of input. In the following example getline is used to count the lines in a file. The command states that it is to return the next line from the file called filename until its return value is greater than zero. It returns 1 if it was able to read a line, 0 if it encounters the end of the input file or -1 if there is some kind of error. Each time the function reads a line the variable n is incremented by one. This number is then written to standard output.

```
awk 'BEGIN{while (getline < "filename" > 0) n++; print n}'
```

Once you start to get the hang of it you rapidly find yourself dabbling in awk scripting. The simple program tidy.awk below is one I wrote which reformats files containing sequence data to FASTA format.

Lines starting with a # are comments.

```
# Tidy.awk: Takes input file of two line
# records separated by blank lines; the first
# line of each record containing the sequence
# name and the second the sequence. Gives
# FASTA formatted output.
# Enter tidy.awk filename at command line.

BEGIN {FS = "\n"; RS = ""; width = 60}
# Expect records separated by blank lines and
# consisting of fields separated by newlines,
# variable called width set to 60

{
#do the following to every line
  name = $1
# assign the contents of the first field to
# the variable name
  x = length($2)
# assign the length of the string in the
# second field to the variable x
if (index(name, ">") != 1)
  print ">"name"\t\t"length($2) > "tidy.out"
else
  print name"\t\t"length($2) > "tidy.out"
# if name does not start with a > character
# add one and write it to file tidy.out,
# if it does then write it to the same file
print "name = "name"\tlength = "x
  if (x <= width) print $2 > "tidy.out"
  else {
    a = 1
    while (x > 0) {
      bit = substr($2, a, width)
      print bit > "tidy.out"
      a += width
      -= width
    }
  }
# if length of sequence string is less than
# required output width then write0 to
# tidy.out, if not then take consecutive
# substrings of the required width
# from the sequence and write them to
# tidy.out
}
}
```

Useful sites:

A comprehensive awk manual and awk source code are freely available at http://w4.lns.cornell.edu/public/COMP/info/gawk/gawk_1.html

Some information (UNIX man pages) on sed is available at <http://www.softlab.ntua.gr/cgi-bin/man-cgi?sed+1>

The Perl homepage is at <http://www.perl.com/>

Useful books:

Quigley, E. 1997. UNIX shells by example. Prentice-Hall PTR, New Jersey.

Dougherty, D. 1992. Sed and Awk. O'Reilly & Associates, Inc., California.

GCG "plugin" for WWWGMS: yet another GCG to WWW interface?

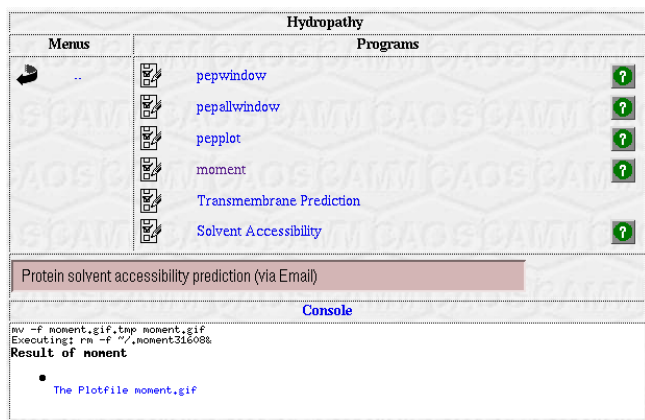
For quite some years now we, at the CAOS/CAMM Center, have been using GMS (General Menu System). With this menu system we offer our users not only uniform access to all the programs and databases available at the Center, but also to e.g. network programs, information services and file and queue managers. As described in EmbNet News volume 4.1 GMS comes in three flavours, the most intuitive one being the Web version, WWWGMS. However, in actually running programs from the Web menu, a telnet or X-windows connection is being set up. For some of our remote users, hidden behind a locally installed firewall, this means that Web menu usage is denied because of firewall blocking of X-windows connections and certain telnet sessions.

This situation encouraged us to make as many of our programs as possible accessible through the WWW. The first step in reaching this goal was the development of the GCG plugin for WWWGMS. This plugin contains two main Tcl procedures:

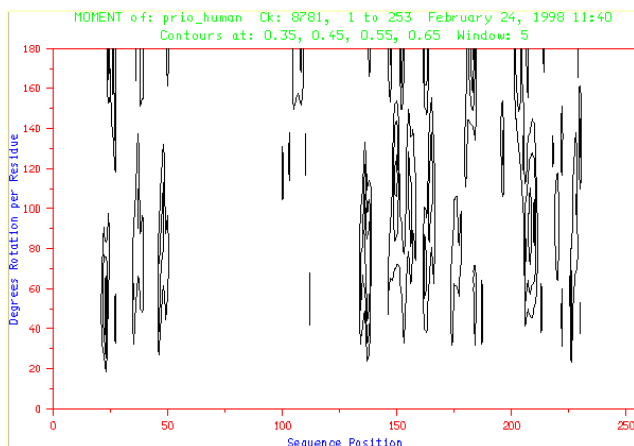
- a procedure that reads a specified GCG/Seqlab configuration file from which it produces GMS calls. These calls are processed by the GMS system and result in a form (a WWW form or an X-windows form, depending on the the GMS version).

Example of a WWW form, dynamically generated while reading the Seqlab configuration file of the "moment" program

- a procedure that parses the form and executes the specified GCG program with the arguments supplied. In the WWW version the procedure creates hyperlinks to the files produced



After execution of the program, hyperlinks to the results of the "moment" run are generated in the Menu console. Of course, these results can also be accessed with the GMS filemanager: the results file is stored in the home directory.



By clicking at the hyperlink, the "moment" output is transferred to the WWW browser and shown. Because this transfer uses "http", firewall problems are circumvented.

So is this "plugin" just another GCG to WWW interface?

Well, yes and no. Yes because interfacing is the basic function of the plugin; it creates WWW forms which a user can complete and submit.

But no because the plugin performs this function by using GMS calls. So in fact, GMS generates the Web forms. Hence this GCG plugin cannot be used as a standalone application to generate HTML forms for the (E)GCG programs like other GCG WWW interfaces such as W2H and WWW2GCG. It has to be part of a GMS implementation.

Still a beta version

Most of the plugin has been finished and executes error free but parts are still under development to upgrade it to full

functionality. E.g. when specifying an input sequence, the sequence's size should be calculated and shown in the forms where e.g. the allowed values for the begin and end value for an analysis are displayed. We also plan to use some Javascript to make use of the rules that can be defined in the Seqlab config files.

For more information, please contact Koen Cuelenaere at the CAOS/CAMM Center.

Lines&Kinetics: a graphic tool to deal with linear regressions and enzyme kinetics.

Manuel G. Claros, Francisco M. Cánovas

Laboratorio de Bioquímica y Biología Molecular, Facultad de Ciencias e Instituto Andaluz de Biotecnología, Universidad de Málaga, 29071 Málaga, Spain.

Introduction

Many experiments in laboratories require a correlation analysis based on a straight line (for instance enzymatic activity and binding, exponential decays, microbial culture growth, calibration curves, ELISA, protein concentration...). There are several ways to accomplish this by means of large, expensive, commercial programs that have features which are, for the most part, unneeded. Moreover, the researcher has to use the software with blind faith; since the experimental data are analysed by a computer, one assumes that the results must be meaningful and correct. On the other hand, there are also many small utilities that can perform linear regressions, but their lead to significant lacks, such as inability to do data interpolation or maintain the interpolated data on the screen, inability to choose what values should be used to construct the line, no simultaneous plot given, and lack of intuitive evaluation of the fit (Straume and Johnson, 1992). On the contrary, the construction of linear regressions with logarithmic data is very important to calculate the doubling time for a microbiological culture (Stanier et al., 1995), datum that is not readily provided by any software package. Finally, enzyme kinetic that follows the Michaelis-Menten equation can be studied by means of a straight line since the equation has been transformed to convert it into a line. With this in mind, we have designed a small, inexpensive and easy to use application that fills the requirements for the experiments mentioned at the beginning.

The linear regressions

Lines&Kinetics is a HyperCard stack programmed in HyperTalk language for Macintosh computers. It requires HyperCard Player 2.1 or later. We have used the Gauss-Newton method for the least-squares fit because it is the best choice for linear regressions (Johnson and Faunt, 1992). The least-squares fit assumes several considerations: (1) that all the experimental uncertainty can be attributed to the y variable, (2) that the experimental uncertainties of the data can be described by a Gaussian distribution, (3) that no systematic error exists in the data, (4) that a straight line is the most fitting, (5) that the data are independent of each other, and (6) that there are enough data points to provide a good sampling of the experimental uncertainties. Usually the first five conditions have to be assumed by the experimenter, being the last condition the most empirical. Although the minimal number of independent data points is equal to the number of parameters being estimated, experimental data contain uncertainties that increase the number of data points needed (Straume and Johnson, 1992). Unfortunately there is not an a priori way to predict it, but Lines&Kinetics estimates the minimal number of data (Mn) that can estimate the y value with an error of 5%, and a confidence limit of 95%, following the equation

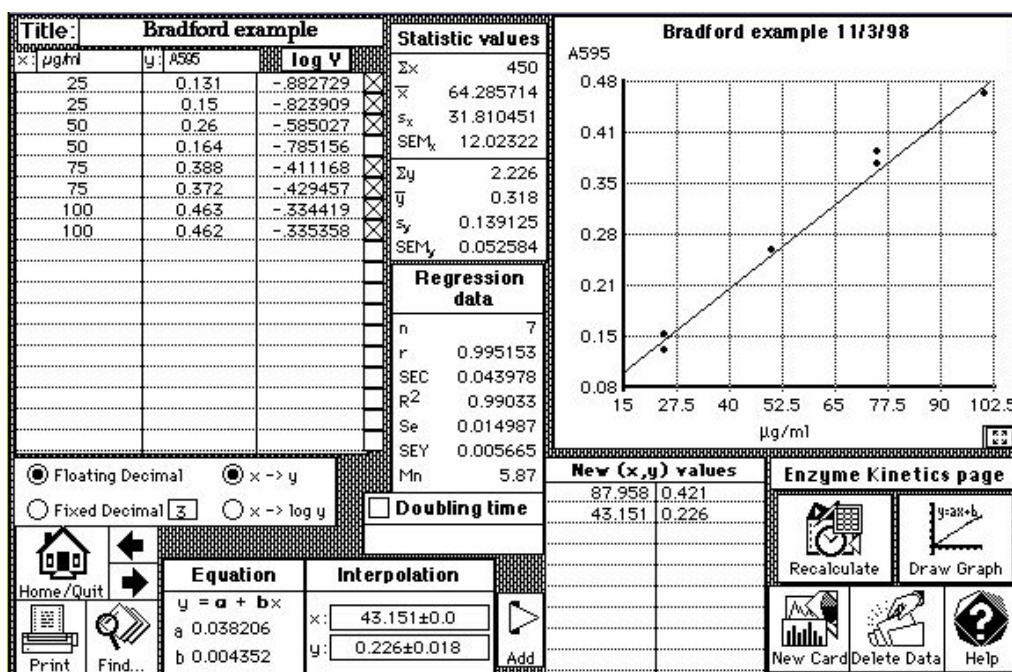
$$M_n = \left(\frac{t_{0.025; N-2} \times S_e}{0.05 \times \bar{y}} \right)^2$$

The accuracy of the fit can be evaluated (Hassard, 1991) by means of correlation (r) and regression (R2) coefficients, the standard error of correlation (SEC), the error standard

deviation (se), and the standard error of f(x) (SEY). The confidence limits of interpolated values, instead of the standard error, are displayed with every interpolation. Statistical estimates, such the standard deviation (s) and the standard error of the mean (SEM) that would permit hypothesis tests (DeGroot, 1986), are provided by Lines&Kinetics.

Sometimes a point(s) appear(s) to be totally different from the pattern displayed by the majority. These points are known as outliers (Draper and Smith, 1966) and may affect the least-squares fit. One way to minimise this effect is excluding it from the analysis; although it is an arbitrary choice of the experimenter, one possible criterion is that removal of the outlier(s) should substantially decrease the standard deviation, so it diminishes clearly the minimal number of data estimated to provide a reliable result. For example, Figure 1 displays a protein concentration correlated with the absorbance at 595 nm (A595) following the Bradford protocol (Bradford, 1975). Without considering the outlier, the least-squares fit provides a regression of 0.99, a SEY of 0.006 and 6 the data points number estimated to construct the same adjustment with an error of 15%. However, when the outlier is taken into account, the regression diminishes to 0.93, SEY increases to 0.014 and the minimal number of data points suggested is set to 37 (data not shown). This and the confidence limits for interpolations indicate that the pattern line obtained is very precise.

Figure 1: Typical default display of Lines&Kinetics. The calculation of protein concentration of two A595 values is shown. Note that an experimental couple of data points has not been considered for fitting (see text for details).



Logarithm calculation permits the use of a straight line when data are correlated by an exponential function, as for example microbial cultures or exponential decays; we have implemented the estimation of the doubling time since in biological laboratories it is frequently used and requires time consuming calculations (Claros et al., 1995).

The compiled arguments indicate that the use of Lines&Kinetics involves several advantages regarding other similar software, like simultaneous plotting of data points and equation, retention of interpolated values, the intuitive way to assess fitting goodness and the handle of outliers. Values of x and y can be obtained as well as keyboard inputs, from text, tab delimited files. All results can be exported as a tab-delimited text or a graphic (PICT) file (Figure 1).

The enzyme kinetics

Another way to interpret x , y values are as substrate concentration ($[S]$) and reaction rate (V) respectively, for an enzyme kinetic analysis. Knowledge of the kinetic parameters limiting rate (V_{max}) and Michaelis constant (K_m) is very useful for enzyme characterisation and biochemical purposes. The problem for the novice is the choice of a method for calculating these valuable parameters. Statistical arguments and criteria have settle some problems such as weighting the points (Di Cera, 1992) and use of least-squares fitting techniques. The popular methods of Lineweaver-Burk ($1/V$ against $1/[S]$) and Eadie-Hofstee (V against $V/[S]$) are the most inaccurate ones and should never be used to calculate K_m nor V_{max} for publication. Although Lineweaver-Burk gives the best-looking straight line when fitted by eye, it gives a grossly misleading impression of the experimental error and the points are not distributed homogeneously. Eadie-Hofstee does not fill at all the conditions for a least-squares method since the error containing variable (V) is in both axes, provoking an angular distortion of the error. The third classical method, the Hanes-Woolf plot ($[S]/V$ against $[S]$) avoids all these problems, even unweighted, and should be preferred over the other straight-line plots for most purposes (for a detailed comparison, see (Henderson, 1992; Cornish-Bowden, 1996)). Despite this, Lines&Kinetics calculates the K_m and V_{max} using the three methods and the least-squares fit, giving the correlation and regression coefficients for each method. In cases where a co-operativity is present, calculation of the Hill coefficient is also available.

The three methods mentioned above are parametric i.e. they assume the mentioned characteristics for the least-squares fit, which is not always obvious in enzyme kinetics. A non-parametric method should be desirable, such the Direct Linear plot ($-[S]$ against V), in which the median values of a series of estimates of K_m and V_{max} are shown to give the best result. This plot avoids validating the normal distribution of errors, shows insensitivity to outliers, and

provides the best estimates for K_m and V_{max} (Henderson, 1992; Cornish-Bowden, 1996). This method can be accomplished by hand but requires hard calculations to provide the median estimates. Lines&Kinetics takes into account all the considerations that this analysis requires and assesses the punctual estimates of K_m and V_{max} and displays the best estimate of them.

Availability

This free software is available for Macintosh computers by E-mail to claros@uma.es and by FTP at ftp://ftp.rediris.es/software/incoming/science/. Very soon it will be available too at ftp://ftp.ebi.ac.uk/pub/software/mac/ and ftp://sumex-aim.stanford.edu/sci/

References

- Bradford, M.M. (1975). A rapid and sensitive method for quatitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.*, 72, 248-254.
- Claros, M.G., Perea, J., Shu, Y., Samatey, F.A., Popot, J.L. and Jacq, C. (1995). Limitations of the in vivo import of hydrophobic proteins into yeast mitochondria. The case of cytoplasmically synthesized apocytochrome b. *Eur. J. Biochem.*, 228, 762-771.
- Cornish-Bowden, A. (1996). *Fundamentals of enzyme kinetics*. London, Protland Press.
- DeGroot, M.H. (1986). *Probability and statistics*. Reading, Addison-Wesley Publishing Company, Inc.
- Di Cera, E. (1992). Use of weighting functions in data fitting. *Methods Enzymol.*, 210, 68-87.
- Draper, N.R. and Smith, H. (1966). *Applied Regression Analysis*. New York, John Wiley & Sons, Inc.
- Hassard, T.H. (1991). *Understanding biostatistics*. St. Louis, Mosby-Year Book Inc.
- Henderson, P.J.F. (1992) Statistical analysis of enzyme kinetic data. In Eienthal, R. and Danson, M.J. (eds), *Enzyme assays*. IRL Press, Oxford, Johnson, M.L. and Faunt, L.M. (1992). Parameter estimation by least-square methods. *Methods Enzymol.*, 210, 68-87.
- Stanier, R., Ingrham, J., Wheelis, M. and Painter, P.R. (1995). *The microbial world*. Englewood Cliffs, Prentice-Hall, Inc.
- Straume, M. and Johnson, M.L. (1992). Analysis of residuals: criteria for determining goodness-of-fit. *Methods Enzymol.*, 210, 87-106.

Crackers and other junk

Why there are so many security holes in the Internet community?, asks Markus Sadeniemi.

Last year was the year of junk mail. Of course we have been receiving unwanted mail all the time, but lately the amount of it has exploded. If you write news articles or if your address appears on some web page, you probably end up in the same address databases that the nice folks sending all these "make money fast" letters use.

In Finland where every fourth person carries a mobile phone, junk mail has become a problem even for them. Nowadays you can send a short message to a GSM phone. Most people are not so happy to see a "call me back, honey" message on their display.

Another continuous nuisance are crackers who try to break into your computer. There seem to be a lot of people who randomly test computers on the network to see whether they happen to have some door open for intruders. In a site like CSC we see attempts like this daily, but luckily most of them are quite trivial.

Good old days

Certainly there have been crackers around for a long time. Finland - or FUNET to be more precise - connected to the Internet in 1988 just months after the infamous Internet worm caused havoc in the network and just a few weeks after connections to some US computers were cracked using Finnish computers as turning points.

I still long for the good old days. Then these incidents were rare. Organizations connected to the net were almost exclusively universities and other research organisations. They had no reason to protect their users in case they misbehaved. Nowadays a commercial network operator is not necessarily willing to protect outsiders against its own paying customers. And if you want to send SPAM mail, say an advertisement in ten million copies, you can always go the next internet provider if your account is closed. In the old days a misbehaving student would not change his university once a month.

Legalities and technicalities

In the US it is illegal to send unsolicited fax messages. Many countries, including Finland, are considering whether to prohibit unsolicited electronic mail in a similar vein. This could help a little. By far most of junk mail I have seen has originated in the US. We may be pretty sure, however, that if mass posting become illegal in the US then most of it simply moves to other countries with less strict laws.

Cracking in its many forms is illegal in most countries but this hasn't stopped the activity. As long as the probability of being caught is fairly small, it will go on. The probability is not zero, many people in Finland have been caught and found guilty in court of justice. But the probability is still too low.

This has its technical background in the history of the Internet. People designing the basic structure of the Internet supposed that equipment can fail and made the network self-correcting in case of failure. But the basic assumption was that people behave in a friendly co-operative mode. Security just was not an issue.

It is good manners, for example, to use your own name when sending messages. But the Internet doesn't really force you to that. It is not very complicated to pretend to be someone else. On the one hand we need a mechanism for user identification, but on the other hand we also need well defined mechanisms allowing users to remain anonymous in certain circumstances. The Finnish privacy ombudsman has said that there should be a right to remain anonymous, if you only browse through public web pages. The web provider should not be able or allowed to collect information on you and use it himself or sell it further. Finnish privacy legislation gives you some protection, but not enough.

Politics

The Internet community has been aware of security requirements ever since the Internet worm ten years ago. During the last five years, work has been quite intense. Still new security holes are found in computer operating systems every month. It is not only old systems that are insecure. A surprisingly large amount of new software is being made without regard for security.

Part of the problems would disappear if cryptography was made an integral part of all computer systems. Today all data, even passwords and mail, is transmitted in clear text over the network. Actually all data should be encrypted by default.

Why has this not been done? The main reason is that the US government has strongly discouraged the use of cryptography and it is even illegal to export cryptographic software or hardware from the US. The simple reason for this is that US intelligence organizations want to retain their ability to listen to traffic everywhere in the world.

Of course, as a Finn, I am happy that the Finnish company Data Fellows has a good market for their security products. It is only a small nuisance that I have to delete ssh software from my portable PC before flying back home from the US (and re-install it at home). But I still feel that security and encryption should be an integral part of every computer and

communicating device.

It's your responsibility!

So we do have buggy operating systems and we do have people who want to break into our computers and send mail in a million copies. What to do about it?

First of all, we - the systems administrators at CSC - try to patch the bugs in software as soon as we find them. We hope that we notice them before crackers do. We also do our best to ensure that at least our mail servers are not used for mass postings.

But for you, a user: First of all delete junk mail, don't try to answer it. Don't let it disturb your peace of mind. If you are technically minded, use procmail to filter out some of the mail you don't want to see, but be careful not to filter out too much.

Secondly, use a good password: something that is not a word in any reasonable language or an easy transformation of a word. Adding a number to the end or changing an "i" to a "1" doesn't help. Your password should never be transferred over the net, even on local network, in clear text. It should be always encrypted. For this you need special programs, telnet or ftp won't do. Use ssh instead.

If your data is sensitive, encrypt it. If you send it by mail, use PGP or some other means of encryption. Remember that Finnish laws in certain cases even require you to handle your data with care. This is the case, for example, if your data contains sensitive information about people. In the UK the Government is trying to pass legislation regarding PGP

And remember: You have to keep your password confidential even if there is nothing confidential in what you do! You must not be the one who lets the cracker into the system in the first place.

Markus Sadeniemi
Director of FUNET
Center for Scientific Computing
Markus.Sadeniemi@csc.fi

Upcoming conferences, and WEB site announcements.

Conferences

1) The First International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'98)

Novosibirsk - Altai mountains, Russia

This conference will be the first in the series and run from August 31 to September 8, 1998. We expect a shift of experimental activity from direct sequencing to investigation of genome functioning and regulation in nearest future and, therefore, accumulation, analysis, and recognition of genomic regulatory sequences should soon become major Bioinformatics' targets.

URL: <http://bgrs.bionet.nsc.ru>

mirror in UK: <http://genomic.sanger.ac.uk/bgrs.html>

2) Protein Domain Workshop.

Announcing a Workshop on protein domain analysis at the Sanger Centre, May 18th-20th inclusive.

<http://www.sanger.ac.uk/~birney/PDW>

with an on-line registration form and more information.

The workshop is a 3-day course of both theoretical and practical work. The aim is to train motivated molecular biologists and computer scientists in the latest domain finding techniques

3) Wellcome Trust Genome Campus Inaugural Symposium

15-17 June 1998 Wellcome Trust Genome Campus Hinxton Hall, Hinxton, Cambridge

<http://www.wellcome.ac.uk/wellcomegraphic/a4/al2index.html>

4) Wellcome Trust Summer Schools

In 1985 the Governors of the Wellcome Trust agreed to the establishment of a Summer School programme of advanced residential laboratory-based courses. The primary aim is to provide scientists with 'hands-on' training in advanced state-of-the-art research techniques, directly applicable to their current research interests. Training is in sufficient depth to allow the participants to transfer the technology to their own laboratories.

For further information go to the following URL

<http://www.wellcome.ac.uk/summerschools/>

5) Courses, Conferences and Workshops at the EMBL Heidelberg

The full range of EMBL and EMBO courses and workshops

can be found at the following URL

<http://www.embl-heidelberg.de/collected/Courses.html>

5) First Internet-Extended Bioinformatics Conference sponsored by Virtual Environments International (VEI) at <http://www.vei.co.uk/bioinfiec1/> April 14 to April 24 '98

FREE REGISTRATION

We are pleased to announce the special offer of free registration for the First Internet-Extended Bioinformatics conference to be held on the Internet (<http://www.vei.co.uk/bioinfiec1/>) from April 14 to April 24. Although access to the conference materials and presentations will be free, participants must register through the registration facility available at <http://www.vei.co.uk/bioinfiec1/register/front.html>. A delegate profile will be created and a userid and password will be issued to the registrant for conference access.

WEB sites

1) An interactive image-based key to gymnamoebae (Rhizopoda, Gymnamoebia) is now available via internet on the following address:

<http://rem.ifmo.ru/dox/amoebae.htm>

The current version of the key includes data on all systematically valid species of the family Thecamoebidae and is illustrated with 240 microphotographs, TEM photographs and line drawings. Original species descriptions and references to all corresponding literature complete the information on every species.

2) The first articles of In Silico Biology (ISB) have been published.

In Silico Biology (ISB) is a new online journal appearing at

<http://www.bioinfo.de/isb/>

In an attempt to bridge the gap between experimental scientists and the community of computational biologists "In silico Biology" (ISB) has been established now as an online journal, from which a printed version will be available as well, published by IOS Press, Amsterdam.

3) Bioinformers Events Calendar

This event calendar serves as a clearing house for all kinds of events in the bioinformatics and related areas. Whether you are looking for a big international conference to spend some of your travelling money

on, a workshop dedicated to a specific problem, a small seminar in one of the local institutes or a course on the topic you always wanted to master: you'll find it here.

<http://bioinformers.ebi.ac.uk/Events/>

4) Genome Navigator

Genome Navigator is located at

<http://www.mpimg-berlin-dahlem.mpg.de/~andy/GN/>

A few more microbial genomes have been added for your browsing pleasure. There are 9 organisms in total at the moment.

INTERviewNet

Andrew Lloyd of INCBi interviews Sandor Pongor of ICGEB.

Q. What and where is ICGEB ?

ICGEB is an international organisation which promotes the safe use of biotechnology world-wide, especially in the developing world (<http://www.icgeb.trieste.it/>). As an organisation, we have 60 signatory countries world wide, including Russia, China, India, most Latin-American and East-European countries. From the European Union Italy, Spain and Greece are members. In addition to research, ICGEB is active in organising training courses and workshops, provides about 70 one-to-two year fellowships every year, and runs a grant system. In the practical sense ICGEB is a research institution, doing molecular biology for practical applications, We have two laboratories, one in Trieste, where the headquarters are and one in New Delhi. At these two locations we have 12 research groups with 10-15 members each.

Q. But like all the top organisations do you have an outstation elsewhere?

We have no outstations but there is a system of about 30 Affiliated Centres. These are molecular biology research institutions in our member countries. They serve as local ICGEB contact points which help to distribute our course materials and which are often the recipients of our research grants.

Q. How did an expert on Z-DNA come to work in an organisation nearer the beginning of the alphabet ?

I heard about ICGEB when the organisation started. This was in 1984 while my wife and I were still at Cornell. I

wrote a polite inquiry and 4 years after I got an invitation for an interview.

Q. How long do you think it will be before a significant portion of the third world gets access to bioinformatics ?

Bioinformatics was one of the first areas ICGB started to work on. At present it seems that the main problem of the developing countries is networking, followed by a lack of information on the subject. As networking improves, students and young researchers soon learn how to access the Internet resources, but a systematic knowledge is still hard to get.

Q. How can ICGB most effectively promote this revolution ? And can EMBnet help ?

In 1990, ICGB established the first login service for developing countries. This service (ICGBnet) was modelled on the EMBnet nodes and we soon joined EMBnet itself as a special node. For a long time our main activity was to provide the databases and GCG programs for the users that logged in through X.25 and the Internet. I find equally important the biocomputing courses of which we run one to three every year. Over the years we have provided hands-on training to over 700 students. EMBnet was of great help in the last five years, both as a co-sponsor and by providing the lecturers. Jack Leunissen, Dave Judge, Amos Bairoch, Martin Bishop, Rob Harper have been regularly participating as teachers and will also come this July. Before that we will have two theoretical courses on structural biology, one of them will be a NATO workshop on structural biology and genomics (<http://www.icgeb.trieste.it/net/netcourse.html>).

Q. Anyway Trieste is only a couple of borders away from home in Hungary. Do you go there often ?

Yes, once every two months.

Q. Because of its history as a free port, I have Trieste filed in my mind along with Tangiers as being filled with smugglers, mercenaries and spies. Is it really any more louche than other north Italian cities ?

I have not counted the spies in other cities... Trieste is certainly special in the sense that it has strong Austro-Hungarian and Southern Slav influences which certainly improved the local cuisine. Culturally, the people of Trieste are proud of a Central European atmosphere, which I could best define as something between headache and constipation but I may be wrong in that..

Q. Do all the bars have romantic guitar pickers with bandanas on their heads ?

I have not seen any of these; perhaps we should ask Jack or Dave..

NODE FOCUS

SANBI EMBnet Node under Focus

SANBI was founded by Win Hide in 1996 as a centre for the growth of expertise in Bioinformatics and Genomics in Africa. Through strong support from the South African Foundation for Research Development, GlaxoWellcome Pharmaceuticals, and Silicon Graphics, the node has actually managed to survive and thrive under the harsh legacy of the former government in South Africa. SANBI is sited at the historically disadvantaged University of the Western Cape, though located at a nature reserve within the shadow of the Table Mountain it is situated in a bleak light industrial area.

SANBI started as a little office with an SGI Indigo 2, donated by Juli Nash at SGI, and a borrowed macintosh. The office was shared by Win Hide, a tiny internet connection, and Rob Miller, an American who had trained in Janet Thorntons' Lab at UCL in the UK.

Via a project funded by the US Department of Energy, we have built a database of all publicly known human gene transcripts (ESTs) and have clustered these ESTs by identity to produce consensus sequences. Now we are in the process of linking these sequences to the genome and also proteins that they code for.

Rob was horrified to learn that he was funded by the US government to work with ESTs. As we all know, these are the most violently despicable of all know types of sequence as they really are of low quality. Rob moved to Durban in disgust and, via a remote link, managed to work in Japan, the USA and Cape Town, to produce the database now known as STACK <http://techno.sanbi.ac.za/stack/>

STACK can be accessed at SANBI and now also at the US National Centre for Genome Resources

<http://www.ncgr.org/cgi-bin/SANBI/maestro/front.pl>

for alpha account

or <http://www.ncgr.org> for further access information.

It differs from most EST databases in that it contains all alignments, references and consensus sequences. It is organised according to tissue type as opposed to whole body clusters.

SANBI has some interesting staff, students and projects. We'd like to share a little of that with you.

Research at SANBI:

SANBI's focus is in Genome Bioinformatics. The work revolves around the epidemic tuberculosis in the surrounding city, the genome information appearing from Sanger and TIGR and the population information appearing from labs in South Africa and the rest of the world. We are attempting to link genome data, population data and evolutionary analysis to provide insight into the virulence of TB. We are funded by GlaxoWellcome and the South African MRC for this work. To date we have discovered that there is really very little variation indeed between the genomes of TB that have been sequenced by TIGR and Sanger. Weird. TB seems to like to vary mostly by insertion element variation.

In other projects our node manager and senior applications programmer, Andrey Ptitsyn, hails from Siberia to bring expertise in word frequency analyses and distance measures. Naturally, military-style we put this to good use in making him manage the GCG 8.0 package and integration for the future arrival of the WebAngis suite into our local site.

Other PhD students from South Africa and Ghana are working on Neural net promotor prediction, genome evolutionary rate analysis and human genome transcription - alternate splicing. Projects are based here in South Africa, but we use supercomputers at the NCSA in the USA and also at Tsukba DNA Research Institute in Japan. We regard the Irish-Japan Axis as our base as Win Hide was trained in the same evolutionary tree as Andrew Lloyd and Takashi Gojobori.

Training

Having the good will of WebAngis has really helped, as it allows us to focus our fixed resources on training and development of South Africans in Bioinformatics. To that end we have started with the Southern Hemispheres first. Summer 1998: Object-Oriented bioinformatics Programming with Java Course (<http://techno.sanbi.ac.za/javacourse/index.html>) provided by Brian Karlak, a hang-glider pilot and bioinformaticist from Berkely in California. The glider is based at the House of Bioinformatics in Cape Town, where most of the visiting Post Docs tend to stay.

Service

Our service provision has been pretty limited, mostly because we have had a series of technical problems in installation of WebAngis. These are now en route to becoming history, so we look forward to going online in mid 1998 with GeneKraal our own biltong-flavoured version of the Australian WebAngis system.

TecSpec:

SANBI is a National Institute and a proud member of EMBnet. We have 10 workstations (SGI/SUN/LINUX), an SGI Origin 2000 4 processor, a 64KB dedicated link.

Future:

SANBI is internally linked with full ATM connectivity and an external ATM that links between the major universities in South Africa. This experimental network will be our focus in terms of developing remote training services. SANBI also has a dedicated link direct to Vienna and Virginia in the United States.

We wish to develop training links with the USouth Africa and Europe to provide the community with powerful databases and, in turn, to house visiting researchers and trainers. We welcome solicitations to visit and will attempt to host numerous visitors in the years to come.

Meetings:

SANBI is supporting the bioinformatics session of the INTERNATIONAL CONFERENCE ON SYSTEMS, SIGNALS, CONTROL, COMPUTERS in Durban, South Africa from Sept 22-24 1998.

<http://nsys.ntech.ac.za/iaamsad/SSCC98.html>

DEJAnews

The Plague

The plague of the serious research worker is SPAM. That is to say unsolicited Email on topics that are of no interest at all. The networks, newsgroups and mailing lists are being flooded with messages which promise to make you a millionaire or satisfy your desires. Now I wouldn't mind being a millionaire, and I have nothing against satisfying your desires, but at my age, it has all become a bit tiresome, a bit boring and besides none of it ever works!!!

The Bionet newsgroups

Research scientists need to communicate with each other; they need to exchange information and ideas. In the past the Bionet newsgroups have been the forum that scientists have used to discuss matters related to biology. However many people have recently been put off by the senseless waste of bandwidth and off-topic postings which means that they have to filter through loads of dross and get to the gold. The noise to signal ratio has increased significantly.

The Bionet Web site
















Attempts have been made to provide quality control in newsgroups by making them moderated and ,at a last count, 47 of the bionet newsgroups now have moderators, who filter the messages posted to newsgroups and will only sanction "on-topic" material that is pertinent. Some of the newsgroups in the Bionet archives have been storing messages since 1990 or thereabouts and all of them are WAIS indexed and are also available for reading via hypermail. This is a valuable resource and should be more widely publicised and supported.

Dejanews Server

It used to be that the only access a scientist had to the internet was by Email and, if you mentioned a newsreader or newsgroups then, they were completely dumbfounded. But now with the advent of the web it would seem that everybody wants to be famous for their 15 minutes and the "home-page" is now the premier method of delivering information. However, it would be nice to revisit Usenet and the Bionet newsgroups and see how Dejanews has repackaged Usenet for the point and click browser freak.

The Dejanews server indexes nearly all of the Usenet newsgroups and a limited number of the bionet newsgroups. They have made a selection of those bionet newsgroups that they think are the most valuable.

Using Dejanews is a convenient way to read and post to Bionet newsgroups. If you decide to browse the bionet heirarchy then the presentation of newsgroups is as follows

Browse Groups Results			
 You are in the bionet newsgroup hierarchy			Help
<u>Group</u>	<u>Browse</u>	<u>Read</u>	<u>Post</u>
bionet.agroforestry		 809 articles	Post
bionet.announce		 244 articles	Post
bionet.biophysics		 110 articles	Post
bionet.cellbiol		 274 articles	Post
bionet.drosophila		 131 articles	Post
bionet.general		 325 articles	Post
bionet.genome	 1 branch		
bionet.immunology		 257 articles	Post
bionet.info-theory		 180 articles	Post
bionet.jobs	 2 branches		
bionet.metabolic-reg		 148 articles	Post
bionet.microbiology		 403 articles	Post
bionet.molbio	 3 branches		
bionet.neuroscience		 405 articles	Post

By clicking on the "244 articles" in the bionet.announce newsgroup you can gain access to all of the most recent postings. Each posting is listed by an article number, the date on which it was posted, the subject of the posting and also the author of the posting. Since Bionet.announce is a moderated newsgroup then the quality of the postings are usually good and most of the SPAM has been filtered out.

As can be seen from the Subject lines bionet.announce is a newsgroup that deals for example, with announcements of new databases (TRANSFAC 3.3) or upcoming conferences (SMS Meeting) etc.

Clicking on any of the subject lines will bring up the full text of the original posting... so if you want to keep up with what is current in biology then it is a good practice to take some time out to read a newsgroup such as bionet.announce on a weekly basis.

Quick Search Results

Matches 151–175 of exactly 244 for search:

~g (bionet.announce)

Find

- Help
- Power Search
- Interest Finder
- Browse Groups

	Date	Scr	Subject	Newsgroup	Author
151.	98/02/13	025	Urgent RECOMB 98 registrati	bionet.announce	Eugene Kolker
152.	98/02/13	025	= CGG Web Genomics analysis	bionet.announce	Victor Solovyev
153.	98/02/12	025	2nd Intl. Conf. on Transgeni	bionet.announce	Li Birong
154.	98/02/12	025	Recomb98: Deadlines & Pr#1/2	bionet.announce	Eugene Kolker
155.	98/02/12	025	Recomb98: Deadlines & Pr#2/2	bionet.announce	Eugene Kolker
156.	98/02/12	025	USGS Earth Sciences Intern -	bionet.announce	DORSEYB
157.	98/02/12	025	TRANSFAC 3.3-Announcement	bionet.announce	Thomas Heinemeyer
158.	98/02/12	025	11th Congress of the FESPP.	bionet.announce	ttsonev
159.	98/02/12	025	WELLCOME TRUST GENOME CAMPUS	bionet.announce	Gary Williams
160.	98/02/09	025	Information Technology Appli	bionet.announce	SLaxminara
161.	98/02/09	025	SMS Meeting	bionet.announce	Simon Burton
162.	98/02/09	025	Recomb98: student financial	bionet.announce	Eugene Kolker
163.	98/02/05	025	New Cyanobacterial Bibliogra	bionet.announce	Mark
164.	98/02/05	025	Postgraduate Studentships av	bionet.announce	Frank Norman
165.	98/01/18	024	Career Discussion Forum, 199	bionet.announce	Dave Jensen
166.	98/01/18	024	Plant Proteins in Abiotic St	bionet.announce	Plant Protein Clu
167.	98/01/18	024	An Introduction to Structura	bionet.announce	Plant Protein Clu
168.	98/01/18	024	Plants, Proteins and Express	bionet.announce	Plant Protein Clu
169.	98/01/18	024	BRI: Online Course about Bio	bionet.announce	Christian Frosch
170.	98/01/18	024	PIR-International Protein Se	bionet.announce	Christopher R. Ma
171.	98/01/18	024	J.D. Hanawalt Powder Diffrac	bionet.announce	SQuick2653

Find me info on PFAM

If you are like me, then I often read an article and at the time it it does not seem improtant. But six weeks later it is a matter of life or death to track down the original posting and this is where Dejanews PowerSearch comes in real handy. Let's say that your boss wants to know where to get information on PFAM and it has to be on his desk on the afternoon or it's your head on a platter. Well you could use a web search engine like Altavista and track it down that way, but often you get lots of hits on your keyword. It is far better to get the word straight from the horses mouth. That is to say, what information is available from the original authors?

Dejanews Powersearch Form

The way to do this is to use powersearch within Dejanews. There are a few fields that you can fill in so you can narrow down your search. The first dialogue box allows you to fill in your keyword e.g. pfam. From the ARCHIVE you can select (complete, standard, adult or jobs.). From KEYWORDS MATCHED you can select (25, 50, 100). From the RESULTS FORMAT you can select (Concise, Detailed, Threaded.). From the SORTED by you can choose (Score, Group, Author, Subject, Date.). In the GROUPS dialogue box type bionet.* so you will only search the bionet groups. Here is what the filled-in form should look like.

The Leader in Internet Discussion

Search
Post Message
my
My Data News
Help

Power Search

Search for: Find

Example: ufo AND (sighting OR abduction OR alien)

- Help
- Quick Search
- Interest Finder
- Browse Groups

Archive: Group(s):

Example: alt.tv.x-files or *x-files*

Keywords matched: Author(s):

Example: demos@dejanews.com

Number of matches: Subject(s):

Example: FAQ or (Frequently Asked Questions)

Results format:

Date from: To:

Sorted by:

Example format: Apr 1 1997

PowerSearch PFAM results

The results are then returned to you in the following format and, as you can see, they are nicely sorted according to author. Also, since you have only selected the bionet newsgroups then you are not snowed-under with useless information from newsgroups in which you have no interest. The most interesting hit would appear to be Pfam 2.1 Release by Ewan Birney.

Stop using **YOUR** newsreader

[Click here to start using My Deja News.](#)

Power Search Results

11 Matches for search:

pfam

Find

- [Help](#)
- [Quick Search](#)
- [Interest Finder](#)
- [Browse Groups](#)

- [installing PFAM database from Sanger on SGI ???? - Victor Lu](#) 1997/01/21
- [Installing PFAM database from sanger on SGI ????? - Victor Lu](#) 1997/01/21
- [Installing PFAM database from Sanger on SGI ?? - Victor Lu](#) 1997/01/21
- [Bromodomain Update - Francois Jeanmougin](#) 1997/08/19
- [Pfam 2.1 Release - Ewan Birney](#) 1997/10/31
- [New Release 2.0 - ProClass Protein Family Database - Cathy Wu](#) 1997/12/02
- [New Release 2.0 - ProClass Protein Family Database - Cathy Wu](#) 1997/12/04
- [New Release 2.0 - ProClass Protein Family Database - Cathy Wu](#) 1997/12/04
- [New Release 2.0 - ProClass Protein Family Database - Cathy Wu](#) 1997/12/08
- [Re: Finding consensus seq in my protein](#)
 - [Sean Eddy](#) 1997/12/09
- [Wise2.0 beta release - Ewan Birney](#) 1998/02/27

Pfam 2.1 Release posting

You should note that DejaNews exhorts you to stop using your newsreader and start using MyDejaNews. You will need to register if you want to post via Dejanews. They are also making strenuous efforts to filter out SPAM so it looks like it will develop into a very useful service.

Clicking on Ewan's name will bring up the full text of the posting. A nice feature is that each article has a special header which allows you to do various things like:

- Post a NEW message into a Bionet newsgroup
- Post a REPLY to an article you have been reading
- Email a PERSONAL REPLY to the original author of the article
- BOOKMARK an article for later recall
- Obtain an AUTHOR PROFILE

Author Profile Results

Author: [Ewan Birney <birney@sanger.ac.uk>](#)

▪ [Help](#)

- 9 unique articles posted.
 - Number of articles posted to individual newsgroups (slightly skewed by cross-postings):
 - [7 bionet.software](#)
 - [1 bionet.announce](#)
 - [1 bionet.software.gcg](#)
-
- Our Author Profile is a great way to get insight into an author's Usenet presence and find out what he/she is interested in. Indexing errors however, though rare, can occur. Because of this, the newsgroup names/counts may not always be completely accurate, and as our [disclaimer](#) states, we are not liable for said inaccuracy. You can always check the actual article numbers by clicking on each newsgroup link.
 - By your continued use of our service, you agree to be bound by the terms expressed in our [disclaimer](#), and agree not to hold DejaNews responsible for the contents of the Usenet database, or any inaccuracies in the information provided.

Give me an author profile

I think that the author profile is a very neat enhancement. On Usenet especially for the newcomer there is the tendency to believe that if it has been written down then it must be true. The newbie does not discriminate. He believes everything. So, if you click on the author profile then you will get the following display, which shows that Ewan has also been posting messages to bionet.software and bionet.software.gcg. If you are then interested in finding out what Ewan has posted to bionet.software then simply click on the the hyperlink and the articles will be displayed.

Author Profile Results

Author: **Ewan Birney** <birney@sanger.ac.uk>

[Help](#)

- 9 unique articles posted.
- Number of articles posted to individual newsgroups (slightly skewed by cross-postings):

- [7 bionet.software](#)
- [1 bionet.announce](#)
- [1 bionet.software.gcg](#)

-
- Our Author Profile is a great way to get insight into an author's Usenet presence and find out what he/she is interested in. Indexing errors however, though rare, can occur. Because of this, the newsgroup names/counts may not always be completely accurate, and as our [disclaimer](#) states, we are not liable for said inaccuracy. You can always check the actual article numbers by clicking on each newsgroup link.
 - By your continued use of our service, you agree to be bound by the terms expressed in our [disclaimer](#), and agree not to hold Deja News responsible for the contents of the Usenet database, or any inaccuracies in the information provided.

Postings to bionet.software

Once the articles have been displayed you can see that Ewan has got interests in quite a few areas of bioinformatics... a real renaissance man... and that far from promising to make you a millionaire, he actually posts messages related to biology. And, by the way, there is no need to type in those mysterious keywords that are in the dialogue box since they appear automagically when you click the bionet.software link.

Power Search Results

7 Matches for search:

- [Help](#)
- [Quick Search](#)
- [Interest Finder](#)
- [Browse Groups](#)

	Date	Scr	Subject	Newsgroup
1.	98/02/27	042	Wise2.0 beta release	bionet.software
2.	97/10/25	037	Re: Multiple Alignment of St	bionet.software
3.	97/07/16	036	Re: readseq in C++/Java - co	bionet.software
4.	96/10/29	035	Re: Mulitple alignment edito	bionet.software
5.	96/10/25	035	Mulitple alignment editor	bionet.software
6.	96/10/07	035	PairWise and SearchWise new	bionet.software
7.	96/10/07	035	Re: Blast search help needed	bionet.software

As you can see, Dejanews has indexed articles from Ewan from as far back as 1996 and his latest posting is only last month. As an exercise I will leave it up to you to find out what Wise 2.0 beta release is. You never know, it might finally be a scheme for making money that really works... so go ahead and satisfy your desires and find out what it says!!!

Rob Harper

Node News

EBI - European Bioinformatics Institute NEW BLITZ

Rodrigo Lopez

MPSRCH at the EBI is no longer available. Hardware failure on the MasPar regrettably prevent the EBI from continuing to offer this service. It is now replaced by a Bic2 from Compugen and is complemented with scans from Geoff Barton who has recently joined the EBI. The two 'methods' (Compugen's and Barton's Smith and Waterman implementations) provide a comprehensive S&W scanning service to the community.

The services can be reached over the WWW on the following URLs:

http://www2.ebi.ac.uk/bic_sw

<http://www2.ebi.ac.uk/scanps>

The blitz email server is still in operation and permits the user to choose either of the S&W methods. Full instructions on how to use the blitz server can be obtained by sending an email message to blitz@ebi.ac.uk with the word 'help' in the message body.

Interactive use of the Bic2 is restricted since it is a serial device and cannot handle multiple requests simultaneously. Therefore, when the device is busy, interactive submissions are sometimes automatically redirected to the email server for processing.

Node news from BEN

Guy Bottu

Good news : the Belgian Federal Office for Scientific, Technical and Cultural Affairs has recently renewed the financial support of BEN for a period of three years (until December 2000).

Hardware/software/lifeware extensions

BEN has acquired two new computers, a SUN Enterprise 450 equipped with 4 250 MHz processors, 512 megabyte memory and plenty of storage space, aptly named "bigben" and a PC Alpha carrying an AXP 533 Mhz with 512 megabyte memory named "babyben". They will soon replace the DEC Alpha 3000/500 which was becoming seriously overloaded. The BEN staff has been reinforced with a third collaborator, David Coornaert. His main task will be to

implement a series of new services in the domain of "genomics". He will start by implementing a BLAST/FASTA server for searching a sequence against completely sequenced genomes. BEN has improved its protein databank and offers now to its users a non-redundant and weekly updated databank, composed (in that order) from SwissProt, PIR, GenPept and TREMBL. It is available under GCG, BLAST and SRS. Among the databanks and software added to the collection available at BEN during the last two years let's mention CUTG, IMGT, dialign, LALNVIEW, Sequin and XGRAIL.

WWW2GCG development

The development of WWW2GCG, a Web interface for the GCG software, goes on without interruption under the direction of Marc Colet, manager of BEN. Amongst the additions of the last two years, let's mention:

- the "Webshell", an interface to UNIX commands which allows the user to inspect, create and delete files and directories in his personal account in the remote computer.

- BOV (Blast Output Viewer): a file parser which transforms BLAST and FASTA outputs on-the-fly into hypertext. BOV was developed by Philippe Alard, collaborator with BEN. BOV outputs allow you to retrieve sequence entries by hyperlinking to the SRS wgetz command on the remote computer.

- the possibility to automatically open GCG output files using the MIME concept. From the BEN site, MSF files are send with content type chemical/msf, e.g. for opening with the sequence editor Genedoc (see Embnet News 4(2)). Modified SIM outputs are sent as chemical/x-aln for opening with LALNVIEW (see <http://expasy.hcuge.ch/sprot/sim-prot.html>)

- ROASTED: an "over the Web" sequence editor that offers the same functionality as seqed but avoids the need of a telnet session. ROASTED is a Java Applet, developed by Pierre Spegelaere, collaborator at the Université Libre de Bruxelles and working in close association with BEN.

Node news from Spain

We are glad to learn that Sonia, a Computer Scientist working at EMBnet/CNB has found a better job offer in a private company. She is leaving us in March 1998. During her stay at EMBnet/CNB she has been doing a great job as a System Manager and developing new services for our users and the network community. We all wish her the best of luck in her new career.

As a side effect EMBnet/CNB has now a position available,

which we hope to fill in as soon as possible, for a computer scientist interested in the field of Bioinformatics and we are starting the selection process.

Node News from India

CENTRE FOR DNA FINGERPRINTING AND DIAGNOSTICS (CDFD), HYDERABAD, INDIA

The Centre for DNA Fingerprinting and Diagnostics (CDFD), an autonomous centre registered as a society, has now been set up in Hyderabad by the Government of India under the Department of Biotechnology, Ministry of Science & Technology. At present CDFD is housed in the East wing, 3rd floor of the Centre for Cellular and Molecular Biology (CCMB) and will be shifted to new site when the building is constructed. The centre provides DNA Fingerprinting services to various investigating agencies as well as to the general public. The centre also offers services for molecular diagnosis, carrier detection and genetic counselling for various genetic disorders. The Centre carries out research in areas such as the human genome, biodiversity, wild-life conservation, silkworm genome, genetic disorders etc.

BIOINFORMATICS

The bioinformatics facility at the centre is geared up to cater to the research needs of the country's scientific institutions like DBT (Department of Biotechnology), CSIR (Council of Scientific and Industrial Research), various universities (Central and State) etc., which pursue research in areas of biology and biotechnology. Once the EMBnet node is fully set up and functional the centre will be able to serve the entire country as well as the global research community. To set the ball rolling the centre will shortly have its homepage installed at VSNL's (Videsh Sanchar Nigam Limited) web server, the only ISP at present in the country.

RECENT SYMPOSIUMS & TRAINING COURSES

The Centre in association with CCMB, organised an International Hands-on training course on DNA Fingerprinting during 6-19 November 1997 sponsored by the Federation of Asian Scientific Academies and Societies (FASAS), Centre for Science & Technology of the Non-aligned and other Developing Nations. 21 participants from 8 countries (Bangladesh, Egypt, India, Nepal, Malaysia, Pakistan, Srilanka and Zambia) participated in the training course. The Centre, in association with ADNAT (the Association for the promotion of DNA Fingerprinting and other DNA technologies), has organised a two-day symposium and a two-week residential hands-on training course on DNA technologies: Forensic and other Applications from 23rd February to 10th March 1998.

Update from UCL

Cinema V2.1

For those of you who'd be interested in such a thing, I wanted to let you know that we're making CINEMA available from our ftp site. The small print says: "The source code is now available under the GNU General Public License. For more details, please contact Julian Selley using the feedback form."

(see <http://www.biochem.ucl.ac.uk/bsm/dbbrowser/CINEMA2.1/kit.html>)

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@ted.ie
Irish National Centre for Bioinformatics
Dublin , Ireland
- [IL] INN lsestern@weizmann.weizmann.ac.il
Weizmann Institute of Science
Rehovot, Israel
- [IT] CNR marcella@area.ba.cnr.it
Consiglio Nazionale delle Ricerche
Bari, Italy
- [NL] CAOS/CAMM embnet@caos.camm.nl
Caos/Camm Centre
Nijmegen, Netherlands
- [NO] BiO linda.aksberg@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 Nacional 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@inndama.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/vol5_1/contents.html
- http://www.be.embnet.org/embnet.news/vol5_1/contents.html
- http://www2.ebi.ac.uk/embnet.news/vol5_1/contents.html
- http://www.ie.embnet.org/embnet.news/vol5_1/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://uk.embnet.org/pub/embnet.news/)
- [ftp.be.embnet.org in the directory pub/embnet.news/](ftp://be.embnet.org/pub/embnet.news/)
- [ftp.ebi.ac.uk in the directory pub/embnet.news/](ftp://ebi.ac.uk/pub/embnet.news/)
- [ftp.ie.embnet.org in the directory pub/embnet.news/](ftp://ie.embnet.org/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://uk.embnet.org/pub/embnet.news/)
- [ftp.be.embnet.org in the directory pub/embnet.news/](ftp://be.embnet.org/pub/embnet.news/)
- [ftp.ebi.ac.uk in the directory pub/embnet.news/](ftp://ebi.ac.uk/pub/embnet.news/)
- [ftp.ie.embnet.org in the directory pub/embnet.news/](ftp://ie.embnet.org/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:

Alan Bleasby, SEQNET, Daresbury Laboratory, UK
(bleasby@dl.ac.uk)
FAX +44 (0)1925 603100
Tel +44 (0)1925 603351

Robert Harper, EBI, Hinxton Hall, UK
(harper@ebi.ac.uk)
FAX +44(0)1223 494468
Tel +44(0)1223 494429

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

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

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

Peter Rice, Sanger Centre, Hinxton Hall, UK
(prm@sanger.ac.uk)
FAX +44(0)1223 494919
Tel +44(0)1223 494967

embnet.news

Vol.5, No.1, 1998
March 25, 1998

ISSN 1023-4144