embnet.news

Volume 4 Nr1 (ISSN1023-4144) June 21, 1997

Editorial

Connectivity and the info-tourist

Having worked in Finland I have always been impressed with the networks that exist between the Scandinavian countries and in particular the excellent connections between NORDUnet and the USA. However since moving to the UK I have noticed that at certain times of the day it becomes very tedious even to try to make web connections to the USA. This is because the internet has seen the advent of the "info-tourist". Gone are the days when the research worker from academia reigned supreme.

It would seem that connectivity within countries is fairly good, but as soon as you cross borders then there is a perceived degradation in response times. Indeed EMBnet instituted a PING project[1] a few years ago to monitor the network traffic in Europe, which clearly shows that some of the EMBnet nodes do suffer from poor connectivity.

We can draw a parallel between what has happened to the road networks in the UK and what is happening on the present day Internet. Imagine if you will when the M1 first opened in the UK. You had a 4 lane motorway with hardly any cars driving on it. Look at it today and it is vastly overcrowded with more than the occasional traffic jam. Being stuck in a queue is becoming a way of life. The motorist develops "road-rage" if he has to wait forever in a queue that is going nowhere, and similarly the academic on the infobahn develops "mouse-rage" as he furiously clicks the mouse button in discontent to disconnect himself from a web site that is not responding.

Primary use of Browser

In a recent Web Survey[2] conducted in 1996 it was found that the most common Web activity is simply browsing (77.08%) followed by entertainment (63.79%), education (53.29%), and work (50.9%). Sad to relate academic research was at 36.69% and was rated third from bottom on the table of listed activites. This would indicate that there has been a definite shift in the use of the network. The academic community is no longer in control of what happens on the net and market forces seem to be dictating that the network is primarily being used for fun and entertainment. What can a poor academic do?

Problems using the web

Another area that was investigated[3] was the problems users experience when they try to navigate the WWW. Speed continues to be the number one problem for Web users (76.55%), but in addition there are two other major problems, namely it is difficult to find relevant information on the net (34.09%), and even if you find the relevant information then it is not organised in a way that the end-user finds useful (31.03%), EMBnet has attempted to solve some of these problems. At a EMBnet node you will find solid data, that is current, and well maintained. The common interface for data query and delivery is the SRS system developed by Thure Etzold. So within national borders research workers should have access to an EMBnet node that serves sequence data to them in a timely and easy fashion.

Next Generation Internet/Internet 2 (A network for research and education)

In the USA many of the universities in California are becoming disgruntled by the commercial services that are spreading on the Internet, which are eating up the bandwidth and they are considering rebuilding new high speed local networks to connect universities in their area. Similar plans are being brought forward in Arizona, Chicago, and North Carolina. The notion is that a new network specifically for research and education will be born. In fact there are three different projects in existence. The NSF finances the vBNS which has connected the super computer centres in the USA, and for the Startrek generation there is NGI (Next Generation Internet) which is an initiative organised statewise and is part funded by NSF and finally I2 (Internet 2) which is financed by universities and research institutes themselves. Given the fact that Startrek is so popular and

1. http://www.caos.kun.nl/Ping
2. http://www.cc.gatech.edu/gvu/user_surveys/survey-10-1996/graphs/use/Primary_Use_Of_Browser.html
3. http://www.cc.gatech.edu/gvu/user_surveys/survey-10-1996/graphs/use/Problems_Using_The_Web.html

Contents

that NGI is a snappier more humourous name than I2, I would bet a pound to a penny that it is NGI that will catch the imagination, regardless of its merits or organisation. Life is like that.

Universities in the USA

Discussion regarding Internet 2 have only just begun but it will be based around vBNS (very high speed Backbone Network Service) funded by the NSF (National Science Foundation). There is a position paper on the vBNS, (http://www.fnc.gov/gigapop.html) as well as a clickable map (http://www.vbns.net/snmp/html/clickable_vbns_map.html).

At the present time only the universities in the USA are participating in the I2 project, however the Canadians have formed their own parallel project with network connections being handled through Chicago. At the moment the I2 project does not have any policies regarding international collaboration or connections.

I feel that it is important to keep up with recent developments in the USA for even though the concept of the WWW originated in Europe, it would seem that the academics have dropped the ball, and it has been picked up by the commercial heavies who are steam-rollering their way to another touch-down... do I hear the spectators moan... we should not be beaten on our own pitch.

Here are some references (http://www.hpcc.gov/whats-new/index.html) to position papers which have been presented before congress on the 13th June 1997. They are in the form of a slide presentation so they could come in handy if you need to persuade some offical that there is a desperate need for a new research and education network here in Europe.

Congressional Internet Caucus Briefing on the Next Generation Internet, June 13, 1997

- Next Generation Internet briefing to the Internet Caucus by John C. Toole, Director, National Coordination Office for Computing, Information, and, Communications.
- Presidential Advisory Committee's Initial Report to the President.
- Recommendations on NGI Initiative presented by Raj Reddy, NGI Subcommittee Chair, Advisory Committee on High Performance Computing and Communications, Information Technology, and the Next Generation Internet.
- Internet2 briefing by Mike Roberts, Project Director, Internet2.
- Next Generation Internet (NGI) Initiative briefing by George Strawn, LSN Working Group Co-Chair, NSF.
- Next Generation Internet Technology briefing by Dr. Howard Frank, Director,Information Technology Office, DARPA.

Software Development
The General Menu System GMS

H. de Hilster, A. Thiers and J.H. Noordik
CAOS/CAMM Center; EMBnet NL

Introduction

GMS(General Menu System), is a user interface developed at the CAOS/CAMM Center (Nijmegen University, The Netherlands) with EMBnet support. GMS provides a uniform and consistent organisation and access mechanism for large collections of programs and databases. It enables users to access and execute programs, search databases, manipulate files and upload and download data, through simple mouse clicks.

GMS comes in three flavours. The one to be selected for implementation in a specific environment depends on local computer facilities (the terminal from which the user accesses the network) and the access route to the host system. The order in which the different GMS versions are presented here reflects the computer network development history and GMS development history. Additional information on the availability of GMS and, if required help with the implementation in a specific environment is available on request.

For users with only simple VT100-like terminals, a multi-windowing emulation, controllable from the keyboard is provided in an ASCII menu version of GMS. For users with a local X server (X emulator) an X-menu version is provided.

For technical details, a user manual, and a description of the controls, the interested reader is referred to the http://www.caos.kun.nl/gms/amenu/gms.html" Web pages at the CAOS/CAMM Center.

A view of this interface and further references are accessible from http://www.caos.kun.nl/gms/xmenu/gms.html.
An access mechanism using a Web browser is provided in the WWWmenu eliminating the need to "logon" to the host computer system.

Availability

A detailed description of the latest GMS development, together with User and MDF manuals, a Design and Implementation description and an Installation Manual is available at http://www.caos.kun.nl/gms/webmenu/gms.html.

The WWWmenu is probably the most intuitive GMS version. It presents a Menu of available tools (programs, databases, utilities, HELP ....) in a Web page and, if future applications allow, also the application I/O is presented in the same Web page. In this respect the development of a GCG plugin ("http://www.caos.kun.nl/gms/webmenu/addons/gcgplugin.html") by Koen Cuelenaere and Jack Leunissen (EMBnet NL) is worth mentioning.

The CAOS/CAMM Center runs WWWmenu as the option of choice from its Homepage. GMS has been recently implemented at the ICGEB EMBnet node services in Trieste and other EMBnet nodes are invited to investigate the usefulness for their site. For support mail to: harcoh@caos.kun.nl.

Accessing a Host system.

Many services provided by host system programs (applications) require interactive graphics (e.g. sketching or graphical output).Which of these applications can be used on a client computer(terminal) depends on the capabilities of the local computer and on the way in which a user approaches the host computer system.

Applications which do not use any graphics can be used with almost any local hardware configuration. Applications which use Tektronix graphics, a relative old graphics standard which requires that the terminal (emulator) supports Tek4010, can be used no matter how the host computer system is accessed.

Programs which use X11 graphics require an X server on the local computer and a fast network connection. Unix machines almost always have X11 capabilities, but for PC's and Mac's additional software has to be loaded.

The table below summarizes local computer (terminal) network configurations and the applications which can be used.

For additional information on this topic interested readers can consult the "http://www.caos.kun.nl/access.html" CAOS/CAMM Center Web pages.

GMS' development history
The Client / Server model

Most users of network services used to "logon" to a host computer system through a telnet from simple VT100 like terminals. Therefore in the early development stage of GMS (1991), emphasis was on a character based interface with as much resemblance to a standard graphical client/server type interface as possible. This specification has led to the ASCII version of GMS which was developed to provide users even on those simple VT100 like terminals, with a multi-windowing emulation, controllable from the keyboard.

GMS has been developed as a completely general menu system without any knowledge of the programs it contains. It is NOT an interface tailored to a specific application. Therefore it is an efficient tool to organize and present large collections of programs, databases etc. in a structured way. At the CAOS/CAMM Center it is used to group all the available services (programs, databases and utilities) in discipline oriented categories ; e.g. programs for chemists, programs for bioinformatics etc. or/and database, modeling programs, etc. and within these categories in subgroups of tools for specific purposes.

The menu structure is described in human readable Menu Definition Files. For details see MDF's in "http://www.caos.kun.nl/gms/amenu/mdfmanual.html". These files are easily changed and/or extended. From the character based interface, a Tcl/Tk X-Window version was developed. This "http://www.caos.kun.nl/gms/xmenu/xvers.html" X-Windows version is fully mouse controlled (instead of keyboard-controlled). In its functionality the Xmenu is fully equivalent to the ASCII-menu and it is self explanatory with its on-line HELP facility. Both the ASCII-menu and the Xmenu are maintained from exactly the same Menu Definition Files.

With the recent and rapid development of WWW, particularly since 1995, most users of network services will now have a WWW browser installed on their local computer. With only minor changes in the original client/server design of GMS, this local WWW browser could be used as client. Specifically the recent browser improvements like tables

Local configuration	Applications which can be executed from GMS	
non graphics application	requires VT100 emulator	yes
Tek4010 graphics	requires Tek4010 emulator	yes
WWW application	requires webbrowser	no
X graphics	requires X server	no
no network access	e.g. dial in through a terminal server	
modem network access	e.g. via an Internet service provider	
direct network access	e.g. connection to a local network	

and frames, support for decent page layout, browser file upload for easy file transfer, and java capabilities to support interactive program interaction, made a WWW browser suitable as GMS client. The GMS server would respond with HTML pages instead of interacting with a local client. From these design specifications the WWWmenu was developed.

The difference between the GMS Web menu and e.g. the GMS ASCII menu is mainly in the presentation and the control of the menus, the programs and the helpfiles. Because of the coding of the actual application (program) to be executed from the menu system, in most cases a telnet connection and/or X11 graphics will be required, but this is handled transparently under menu control. With the continuous development of more and more applications as e.g. Java applets, application I/O will also be presented in the Web page eliminating the need to open additional windows on the terminal screen.

Software Development

Pratt : a program for discovering patterns in unaligned protein sequences

Inge Jonassen, Dept. of Informatics, University of Bergen, Norway. Email: inge@ii.uib.no

Introduction

A common problem in protein sequence analysis is to search for common sequence patterns or motifs in groups of functionally related proteins. Such patterns may be the result of common ancestry combined with conservative evolutionary pressure to maintain important residues at active sites and other functionally important parts of the protein. It is not always possible to identify conserved patterns in protein families. When they do occur, however, they can be very simple and useful tools in helping to identify new members of the families and in trying to understand the relationship between sequence, structure and function [2]. The usefulness of patterns is illustrated by the PROSITE database [1], which contains a collection of protein families, and gives for many of them a characteristic pattern. An example of a PROSITE pattern (entry PS00518 in the database) is:

```
ID  ZINC_FINGER_C2H2; PATTERN.
PA  C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H
```

This pattern will match any sequence (string) containing a segment starting with C followed by 2-4 arbitrary symbols followed by a C and 3 arbitrary symbols and a hydrophobic residue (L, I, V, M, F, Y, W or C) and so on.

The program

We have developed a program called Pratt that, given a set of unaligned protein sequences, outputs patterns matching at least a user-defined minimum number of the input sequences (for a more detailed description, see [2] and [3]). Pratt discovers patterns of the type used in the PROSITE database, and is able to discover patterns containing both flexible spacing (e.g., X(2,4)) and ambiguous positions (e.g., [LIVMFY]). The user defines constraints on the patterns that can potentially be discovered, e.g., the amount of flexibility and the degree of ambiguity. For example, the Pratt program can discover the pattern given above for the zinc finger c2h2 family when given as input the set of sequences contained in the family (PROSITE release 13 gives 286 family members). This particular run of Pratt takes less than 30 seconds on a normal desk-top workstation. Note that the program is not given any information about the pattern nor any alignment.

Pratt ranks the discovered patterns according to their information content, which loosely is the amount of information (in bits) obtained about a sequence when one is told that the sequence matches the pattern. The higher the information content of a pattern, the less likely it is to match a random sequence. The scoring scheme has been extended in different ways to take into account the number of sequences matching a pattern, and how diverse (different) the matching sequences are. For each run of Pratt, the user can choose which scoring system is to be used.

An efficient search algorithm is needed for finding the highest scoring conserved patterns. Version 1 of Pratt uses a search algorithm that is (under certain conditions) guaranteed to find the highest scoring conserved patterns [2]. In many cases it is very efficient, but for analysis of sets of relatively similar sequences it may be inefficient, and sometimes it seems to go on indefinitely. A new search algorithm has been implemented in version 2 using branch-and-bound and heuristics to find the highest scoring patterns more efficiently [3]. This gives a more efficient search, but the introduction of heuristics means that the program is no longer guaranteed to find the best conserved patterns. For testing the performance of Pratt, we analysed all the families in release 13 of the PROSITE database. The complete (unaligned) sequences contained in each family were retrieved from the SWISS-PROT database, and analysed by Pratt version 2.1 using default parameters. In total, 1148 families were analysed, and more than 900 families took less than 10 seconds each.

The latest version of Pratt is 2.1 (released February 1997). In this version, the user can specify parameters to be used either using a menu (see below) or using command-line parameters. The same set of parameters (with the same syntax) can be set using command-line parameters. On-line help is also available.

Figure: Pratt's menu.

```
      Pratt version 2.1

      Analysing 166 sequences from file snake

PATTERN CONSERVATION:
   CM: min Nr of Seqs to Match                   166
   C%: min Percentage Seqs to Match            100.0

PATTERN RESTRICTIONS :
   PP: pos in seq [off,complete,start]           off
   PL: max Pattern Length                         50
   PN: max Nr of Pattern Symbols                  50
   PX: max Nr of consecutive x's                   5
   FN: max Nr of flexible spacers                  2
   FL: max Flexibility                             2
   FP: max Flex.Product                           10
   BI: Input Pattern Symbol File                 off
   BN: Nr of Pattern Symbols Initial Search       20

PATTERN SCORING:
   S: Scoring [info,mdl,tree,dist,ppv]          info

SEARCH PARAMETERS:
   G: Pattern Graph from [seq,al,query]          seq
   E: Search Greediness                            3
   R: Pattern Refinement                          on
   RG: Generalise ambiguous symbols              off

OUTPUT:
   OF: Output Filename                 snake.166.pat
   OP: PROSITE Pattern Format                     on
   ON: max number patterns                        50
   OA: max number Alignments                      50
   M: Print Patterns in sequences                 on
   MR: ratio for printing                         10
   MV: print vertically                          off


X: eXecute program
Q: Quit

help: for on-line help

Command:
```

The menu allows the user to set the minimum number of sequences to match a pattern (Cx options), to define constraints on the patterns to be considered (Px, Fx, and Bx options). Using the S option, the user can choose between different schemes for scoring discovered patterns, the G, E, and Rx parameters allows for control over the search to be performed by Pratt. Finally the user can select the format of the output using the Ox, and Mx parameters. More detailed information is given in the documentation available over the WWW (World Wide Web - see below).

PrattWWW

A WWW interface has been developed for Pratt by Kritian Sturzrehm, Jaak Vilo and Inge Jonassen. This uses a form based html page to allow the user to input his/her sequences and to select values for each of Pratt's parameters. Pratt is started at a machine located at the server site (Bergen or Helsinki and at the EBI at the moment). The output file from Pratt is parsed, and relevant data is passed to a new developed Java applet: PatSeq. The applet is designed to show the matches to each of the discovered patterns in each of the sequences, and also shows overlapping matches. The user can select a subset of the patterns for which the matches are to be shown. In the current version, it is not possible to print or save the graphical output from PatSeq.

Conclusions and Further work

The Pratt programs are able to discover patterns of a quite general form allowing for both flexible spacing and ambiguous pattern positions, thus allowing it to automatically find patterns that are difficult to discover using other available programs. The PrattWWW interface makes it possible for users without the necessary computing power available locally to use Pratt, and also provides a nicer interface to the program. The graphical output from the Java applet helps in interpreting the output from Pratt, and could also be used to visualise patterns found by other programs.

The problem of pattern discovery is closely related to local multiple sequence alignment (LMSA). Most programs for LMSA cannot handle gaps, and will therefore not be able to discover patterns with flexible spacing. On the other hand they may be able to find subtle patterns without any strongly conserved positions - that Pratt will not be able to find. We think that Pratt should be used together with other sequence analysis tools (e.g., Clustal W, Gibbs, MacAw, MEME) as the methods have complementary strengths. Future work could include making a strategy for combining ideas from these different approaches in the same sequence analysis tool.

Availability

The source code (ANSI C) for the program is available via anonymous ftp from ftp.ii.uib.no in the directory /pub/bio/Pratt. And at the EBI, ftp.ebi.ac.uk:/pub/software/unix/ The source code is contained in the file Pratt2.1.tar which needs to be untarred (% tar xvf Pratt2.1.tar) and made using the UNIX make program (% make). It has been successfully compiled and run on a variety of UNIX and Linux workstations. Documentation, PrattWWW, etc. is available over the WWW at http://www.ii.uib.no/~inge/Pratt.html. From this URL there is also a link to output from Pratt for each of 1148 families given by the PROSITE families (release 13).

References

[1] A. Bairoch, P. Bucher, K. Hofmann. The PROSITE database, its status in 1995. Nucleic Acids Res. 24 (1996), 189-196.
[2] I. Jonassen, J. F. Collins, D. G. Higgins. Finding flexible patterns in unaligned protein sequences. Protein Science 4 (1995), 1587-1595.
[3] I. Jonassen. Efficient discovery of conserved patterns using a pattern graph. Submitted to CABIOS Febr. 1997.

Database and Software Development :
Fasta3 and Blast services at the EBI

Rodrigo Lopez, EMBL outstation, EBI

Fasta3

The EMBL-EBI has offered for a number of years a database search service to the general user community. The main core of this service has been the fasta algorithm written by W. R. Pearson(*) and D. J. Lipman (1988), “Improved Tools for Biological Sequence Analysis”, PNAS 85:2444-2448, and W. R. Pearson (1990) “Rapid and Sensitive Sequence Comparison with FASTP and FASTA” Methods in Enzymology 183:63-98).

There exist a number of versions of these programs. Since the source code is freely available and in the public domain several variations and improvements have been made to the original code at various sites. The main version, also known as the ‘generic’ version, is maintained by Bill Pearson (William R. Pearson Department of Biochemistry, Box 440, Jordan Hall, U. of Virginia Charlottesville, VA. wrp@virginia.EDU) and remains the main focus of development for packages such as TIGR's grasta and GCG’s FastA, Intelligenetics and recently EGCG (EMBOSS). The latest release of the package is version 30t36. This release contains a parallelised version of fasta which runs on multiprocessor platforms. The main effect of parallelisation of the software is a significant gain in speed (depending on the architecture and the number of processors used).

Blast2

Blast and Blast2 programs (Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman (1990). Basic local alignment search tool. J. Mol. Biol.215:403-10.) have played a very important role in the past years in the sequence database searching race. Blast is known for its speed and efficiency and in many ways it has become somewhat of a standard for searching the databases. The EBI has embarked in making blast and blast2 available via email and interactively.

The EMBL-EBI has run the fasta service for many years now and a major overhaul has taken place. Two SGI Challenge-L with 12x R10000 processors each and two DEC 8400 with 4 x ALPHA processors each are running the service queues. These machines are capable of doing an entire EMBL database search in less than 4 minutes at normal work loads for a fasta search. Since the speed gain is so high the EBI has decided to offer an interactive service thus allowing users to submit sequence searches and get results almost instantaneously.

The services are currently undergoing the final phases of implementation and are thus limited to protein sequence database searches when it comes to blast and blast2. For fasta there is a complete set of EMBL and protein databases to search although complete EMBL searches are currently limited to email submissions. The database sets available are summarised below:

- swissprot
- swissnew
- trembl
- tremblnew
- sptrembl
- pdb
- embl (by division - for interactive searches)
- emblnew

and comprise only the databases that the EMBL-EBI produces.

About the Job Submission Interface

EMBL-EBI has adopted a straightforward approach in this matter. The main component of the interface is an HTML form which allows the user to change some of the parameters typical to fasta and blast. All the program parameters are set automatically to default values and the user is only required to provide their email address, and optional title for the search, the type of search (interactive or email) and the sequence to be compared either by cutting & pasting into a window or by uploading a file from the computer (netscape only). Comprehensive help can be obtained at various locations on the form.

Once a job has been submitted a report will appear on the screen. This report tells the user what to expect depending on how the job has been submitted. If it is interactive the user is asked to stay in that page and after a short pause the result of the search appears on the screen. The output of the searches is already hyper-linked to the SRS5 server at the EMBL-EBI.

One thing the user will notice is that the blast output contains GCG-style database references rather than the NCBI ones. This allows the user to copy/paste/save the output from a search and use it along with other GCG & EGCG applications which are typical at all EMBnet nodes.

The interfaces are located at:
http://www2.ebi.ac.uk/fasta3/ for fasta3 searches
http://www2.ebi.ac.uk/blast2/ for blast2 searches

(*) An interview with Dr. Pearson appears in the previous issue of embnet.news.

TIPS from the computer room

I/O redirection in UNIX

Jose R. Valverde, EMBnet/CNB

One of the most powerful features of UNIX is the ability to instruct a command to send its output or read its input from anywhere you like. You can even tell a command to send its output to another one for further processing hence combining their functionality in versatile ways. This allows you to have short, elegant commands that do a simple job very well and have them cooperate to get complex results.

While there are a few differences among shells, all of them understand a common subset of redirection directives:

Redirecting output

You can send all of a command's output (everything it would otherwise normally send to your terminal) to a file by using the > character after the command followed by the name of the file into which you want to save the output:

```
ls > mylisting.txt
```

This creates a new file named mylisting.txt (if it already exists, it removes the old one before creating it) and sends all the output of ls to it.

The main problem with > is that if the destination file exists, its previous contents are lost. You can instead append the output of the command at the end of any previous contents by using two right angle brackets, >> to specify the destination of the output:

```
ls >> mylisting.txt
```

Input redirection

You may as well instruct a command that normally reads data from your terminal to get its input from a file instead. This is achieved by using the left angle bracket < character:

```
sort < myfile
```

Note that in this case you don't have the choice to append since it doesn't make sense for the command to append to any input from a previous invocation.

Concatenating I/O

While often useful, it is not enough to be able to redirect one command's input or output to a file. Many times one would want to combine the capabilities of one program with those of another one. For instance, one would like to get a full listing of a directory sorted other than the default way, and maybe seeing the listing one page at a time. This can easily be done by first sending the output of one command to a file, and then instructing the next command to read from that file:

```
ls -l > tmpfile1
sort < tmpfile1 > tmpfile2
more tmpfile2
```

But this is cumbersome and requires the creation of many temporary files. Wouldn't it be nice if we could instruct the computer to take the output of a command and use it as the input of the next one? We can achieve this effect with the bar character | to separate two commands. This is called a pipe since it serves as a conduction of data from one command to the next. We can use series of pipes to combine series of programs:

```
ls -l | sort -k9 | less
```

In the first example we get a full directory listing with ls -l and then we pass on the listing to sort so it gets sorted by file name (-k9 means sort by the ninth field, the file name) and finally we browse the listing comfortably using 'less(1)', an enhanced version of the file browser 'more(1)'.

We can concatenate as many commands as we need to achieve our goal. This allows for great versatility. In this context, each of the intermediate commands we use (like, e.g. sort above) is called a filter since it reads some input, performs some operation on it (filters it) and produces some output that is passed on to the next filter.

Software development
W2H: WWW Interface to GCG

Martin Senger, EMBL outstation EBI

Introduction

The W2H is an abbreviation for the WWW interface to the GCG Sequence Analysis Software Package (Genetics Computer Group, Inc., Madison, Wisconsin) or to the derived services (such as EGCG - Extended GCG, Sanger Centre, Hinxton, UK, or HUSAR, Heidelberg Unix Sequence Analysis Resources). With W2H, we tried to cover as much functionality as possible, and we tried to make it as user friendly as we could achieve. It gives the opportunity to access more than hundred programs from any platform where Netscape runs.

The presented interface is being developed as a collaborative project between European Bioinformatics Institute (EMBL-EBI), Hinxton, UK (within the Biostandards project in the Industry Support Programme) and German Cancer Research Center (DKFZ), Heidelberg.

Description

The users knowing the Wisconsin Package Interface will recognize very soon that the W2H interface was designed to be as much compatible with WPI as possible. On the other hand, the W2H interface supports also a classical usage without obligation of working with the working lists. Thus both sequence oriented and application oriented approaches are available.

Also the implementation is quite different from WPI. The W2H presents a real client-server architecture. A user is a client using the Netscape Navigator (WWW browser) on her/his computer, and all requests are transferred via network (using HTTP protocol) to the server computer where the GCG programs are running and making the analysis. We tried our best to minimize the number of necessary round-trips between client and server.

A typical scenario starts by choosing one or more sequences you want to perform some analysis on. Then you select or type the application you want to run. An application program window appears, displaying the selected sequences as input and allowing you to set the required and optional parameters before running the program. From the application window the program is started and the result window appears. Either you refresh the window by clicking a button, or you specify the client-pull method to poll a result buffer automatically.

The interface is quite complex covering besides executing the GCG analysis programs also features like sequence selector, search set builder, pattern chooser, access to the sequence databases, uploading client files to the GCG server or showing and manipulating the graphical outputs. Together it consists of more than 30 HTML frames plus, for each application, a specialized form with all mandatory and optional parameters is automatically generated.

For special environments, like workshops, conferences and company intranets, there is a special mode (Intranet mode) which can be easily set up and used without having the UNIX accounts for all users on the server side.

Software implementation

The W2H is based on the Netscape Navigator version 2.0 or later or other browsers capable of interpreting a JavaScript scripting language embedded in the HTML documents.

The main advantage of an HTML embedded language is that the whole user request is prepared on the client side without necessity to make the network round-trips. The JavaScript language is capable to verify user's inputs, to suggest the default values and to provide sufficient help. Entering a single value in a form can consequently produce the derived values and let them appear in other places not even in the same form, but also in a quite separate window. It makes the user interface much more powerful and user-friendly.

Security

The applications developed to be used through the WWW interface should always carefully consider the possibility of security holes. The WWW tools, specially CGI scripts, are very powerful and used in the wrong way can make the system vulnerable against unwanted attacks.

The W2H design has to take into consideration the security issues even more seriously because it enables access to the server computer completely, although only for registered users. The interface considers how to protect the server machine against the unauthorized users and how to protect the user data against a not-allowed access by other users.

Comparison with other GCG interfaces

There are also other user interfaces to GCG. Here I present a short comparison with the WPI - Wisconsin Package Interface (Genetics Computer Group Inc., Wisconsin - Madison) and with www2gcg (Bioinformatics Unit, Universite Libre de Bruxelles). Of course, the table does not include all features but tries to concentrate on the topics I consider important both from the user's and developer's points of view.

Summary

Moving activities to the client side substantially reduces networking. The W2H does it generally.

Taking advantage of the GCG/WPI configuration files is an important advantage. These .config files contain both parameter definitions and layout description, including dependencies between parameters. They are also supported and quite improved in GCG 9.

Using a platform independent client browser is a general advantage of any WWW interface over other solutions. The W2H depends on using Netscape (because of embedded JavasScript).

Both W2H and www2gcg implements the real client-server architecture which gives the users distributed processing and better relocation of resources.

	WPI 8.1	*www2gcg 2.0*	*W2H 1.2*
Software / Hardware requirements Client platforms	X on UNIX/VMS	X on UNIX, Windows, Mac	X on UNIX, Windows, Mac
Client software requirements	nothing special	any WWW browser	Netscape Navigator 2.0+
Server software requirements	included in GCG distribution	NCSA compatible httpd server	NCSA compatible, Enterprise httpd server
Functionality Inputs validation	special validation routines on the server side	done by GCG software on the server side	mostly by embedded scripting language on the client side
Networking	by every click	when a client form ready	when a client form is ready is and verified
Managing of non-batch asynchronous results	auto refreshing of the client side	client must intervene itself	auto refreshing of the client side
Access to the files on the client-side	no direct access	copy and paste	files upload incorporated
Terminal oriented applications	by special programs	telnet with support	telnet w/o support
Implementation and development Resource for automatically generated application's interface	GCG/WPI native .config files	distributed pregenerated files	GCG/WPI native .config files
Client-server protocol	X11	http	http
Security of user's files	guaranteed by UNIX access rights	not documented	guaranteed by UNIX access rights and http authentication
Keeping the state information	special process on server-side	shared memory block	shared memory block
Language of implementation	C	perl, C	perl, JavaScript

More independent comparison would be welcome. In the table above, very important features, such as performance, complexity, user learning curve, or maintainability, were not considered at all or only partly.

Future directions

Besides extending the basic functions and features (above all implementing batch queue processing and dependency rules between parameters), the W2H will consider in the future further parsing and processing of the applications outputs and linking it to the other information sources (very probably by using SRS).

An another direction, already started, is a design and an implementation of the CORBA-based interface to the the GCG and similar applications.

Availability

The W2H interface has its own homepage http://industry.ebi.ac.uk/w2h/ with links to the documentation, to the latest news and FAQs as well.

The W2H interface is provided as a free software (but useful only for the GCG licensed users) and is available by anonymous ftp at ftp://ftp.ebi.ac.uk/pub/software/unix/w2h and/or ftp://genome.dkfz-heidelberg.de/pub/w2h.

Node Focus

The EMBnet node in China

The China EMBNet node at the Centre of BioInformatics, Peking University, China
Jingchu Luo*

The economical boom in China during the past six years has made it possible to build up a national network infrastructure. In 1994, China Education and Research Network (CERNet), promoted by the State Education Commission of China, announced its birth in the scientific golden-triangle area in North-West Beijing. This is a region where dozens of research institutes, and two of the largest universities in China, Tsinghua University and Peking University, are located. Ten central network nodes are distributed in large cities all over the country. Most of the universities and research institutions are being hooked to the Internet via CERNet. As one of the central nodes of CERNet, Peking University Network Centre plays an important role in network administration and service in northern China.

China's economic growth also offers funding opportunities for Chinese bioscientists. In addition to various projects from basic research for agricultural and medical applications in biological sciences and biotechnology, the Rice Genome Project supported by the National High-tech program has made profound progress. The China Human Genome Project, consisting of some twenty laboratories, was initiated by the National Natural Science foundation in 1993 and is now becoming one of the biggest national projects. In the biocomputing field, scientists from the Institute of Biophysics, the Institute of Biochemistry, Peking University and the China University of Science and Technology, as well as other institutions, have been working on molecular simulation, protein design, structure prediction and bioinformatics.

As suggested by Chris Sander in 1994 when he was co-sponsoring the European Community Bilateral Workshop of Protein Engineering together with his Chinese partner Xiaocheng Gu of Peking University, we have been trying to set up a local database centre. This will serve the ever-growing demand by local users for access to biological databases. Strongly supported by Sandor Pongor after his meeting with Xiaocheng Gu (who was taking part in the annual meeting of ICGEB Council Scientific Advisers in October 1996), the National Laboratory of Protein Engineering and Plant Genetic Engineering at Peking University was accepted as an associate member of the EMBNet. The Centre of Bioinformatics at Peking University was then initiated in March 1997. This centre is composed of staff from the National Laboratory of Protein Engineering and Plant Genetic Engineering and the Network Centre at Peking University.

An SRS5 server for the EMBL, SWISSPROT, PROSITE and ENZYME databases was installed which speeds up the network transmission for local users. Other sequence or structure related databases such as BLOCKS, PRINTS, SBASE, DSSP, HSSP, FSSP are also being installed. A TRANSFAC mirror was provided by Edgar Wingender and Thomas Heinemeyer from GBF, Braunschweig, Germany. A protein domain assignment server created at Michael Sternberg's lab at Imperial Cancer Research Fund in London is also available on the server. Currently, a protein loop classification database is being developed and will be put on the server after its completion.

The URL and contact email address for the node is:

http://www.cbi.pku.edu.cn/
office@cbi.pku.edu.cn

*J LUO is currently visiting the Imperial Cancer Research Fund at London, (luo@icrf.icnet.uk)

Node News

Node news IRELAND

INCBI, the Irish EMBnet node lurched over another funding crisis at the end of March 1997. For the previous three years we were supported by the Irish Science and Technology agency Forbairt and for the last two years also by user subscriptions. As a trial measure we are now being funded by a consortium of the University Librarians of Ireland. One condition of this support has been that we double the user subscription to about 600ecu per research group per year. It is almost true to say that doubling the subscription has halved the subscribers. After lodging in temporary accommodation since February 1996 - which has been broken into 4 times in the last 3 months - we are moving into new premises in the brand new Institute of Genetics this autumn.

Node News SPAIN

The Spanish EMBnet node has switched most of the scientific processing to a new machine, a PowerChallenge with 6 processors, 768 MB of RAM memory and approximately 70GB of hard disk space running Unix. The move has allowed a mirroring of all of EBI's software and databases, as well as partial mirrors of interest areas from other servers. We are also expanding the range of services provided to users including WWW interfaces to GCG, hosting of WWW pages, mailing lists, installation of new software, scientific and development tools, etc...

The new EMBnet/CNB server is also used to supply the needs of the Spanish National Bioinformatics Network, expanding its areas of use beyong plain sequence analysis into molecular dynamics and structure, image reconstruction, mathematical computing, and others. The old server is being retargeted for scientific use in other biocomputing areas and is no longer intended to be used for plain sequence analysis.

As a part of the change, we have also modified membership rules and conditions for use of EMBnet/CNB resources. A web page describing the new conditions is available in spanish from "http://www.es.embnet.org/EMBnet.CNB/Subscripcion/Academica/"

Finally, we are proud to welcome to the EMBnet/CNB team a new member, Gil Martin, a PhD in Molecular Biology with a long standing trajectory in bioinformatics. Gil Martin is a member of Pharmacia's Department of Immunology and Oncology (DIO) which is now hosted at CNB too.

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_1/contents.html***
- ***http://www.be.embnet.org/embnet.news/vol4_1/contents.html***
- ***http://www.no.embnet.org/embnet.news/vol4_1/contents.html***
- ***http://www.ie.embnet.org/embnet.news/vol4_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.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.1, 1997
21 June 1997
ISSN 1023-4144