Chapter 9
Networking for the biologist

R. A. Harper
EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus,
Cambridge, UK.

> Riding along in my automobile my baby beside me at the wheel
> Cruising and playing the radio with no particular place to go
>
> *Chuck Berry*

1 Introduction

Every research worker would like to have the tools on hand to make his job quicker and more efficient, and with the advent of the World Wide Web many of the tasks associated with molecular biology have become freely available online. In the past when a scientist wanted to know something about a particular subject then the first option was to talk to colleagues in the laboratory and ask for their advice. If that was not sufficient then it was off to the library to scan abstracts or the latest journals for the relevant information.

However times are changing and so are working habits. Why ask questions from people in your laboratory when you can ask the same question on the Bionet newsgroups *http://www.bio.net* from research workers all over the world? Why thumb through textbooks for references when you can type in keywords to an Internet search engine such as Lycos or Alta Vista and get a satisfactory answer in no time at all? But often you find that the major search engines index everything on the Web, which makes it difficult to find exactly what you want. So often it is more profitable to use search engines that are totally dedicated to biology.

In Europe you could use BiowURLd *http://search.ebi.ac.uk:8888/compass/* or BioHunt *http://www.expasy.ch/BioHunt*, which deal exclusively with biology-related subjects. Another comprehensive listing exists at the Virtual Library in the BioSciences division. *http://www.vlib.org/Biosciences.html*, and from China there is the NEE-HOW project, *http://biology.neehow.org* which is an invaluable resource for research workers from the Pacific rim.

In the USA one of the original and best lists of Biological resources, put together by Keith Robison can be found at Harvard *http://golgi.harvard.edu/biopages.list* and of course there is the ever popular *Pedro's BioMolecular Research Tools* at *http://www.public.iastate.edu/~pedro/research_tools.html*. If you are looking specifically for software related to bioinformatics, then there is the BioCatalogue at *http://www.ebi.ac.uk/biocat/biocat.html* or if you are looking for an obscure database then there is DBcat *http://www.infobiogen.fr/services/dbcat* from Infobiogen the EMBnet node in France.

There are also a few good newsletters, which deal specifically with what is happening in the world of bioinformatics. EMBnet produces a quarterly newsletter, which gives an update of the latest developments at the different EMBnet nodes throughout Europe. The EMBnet News can be found at the URL *http://www.ebi.ac.uk/embnet.news/embnetmenu.html*.

The EBI has its own industry programme and they produce a newsletter called the Bioinformer: *http://bioinformer.ebi.ac.uk/newsletter/*. One special feature within this newsletter is the BioEvents Calendar, *http://bioinformer.ebi.ac.uk/Events/* that allows people to advertise workshops, conferences, or symposiums. In the USA there are two major newsletters associated with bioinformatics. The NCBI newsletter is at *http://www.ncbi.nlm.nih.gov/Web/Newsltr/index.html* and the National Centre for Genomic Research, the NCGR newsletter, is at *http://www.ncgr.org/ncgr/ncgr_newsletter.html*

The focus of this article is to help research workers avoid the World Wide Wait while using World Wide Web.

2 The changing face of networking

In the early 1990s academic research workers had the networks all to themselves. Today however, the demography of those using the networks and their reasons for using the networks have completely changed. The competition for bandwidth is fierce between the commercial and academic sector.

The Georgia Institute of Technology (*http://www.gvu.gatech.edu/user_surveys/*) has been conducting user surveys on the use of the Internet since 1994 (see *Figure* 1). Over a four-year period there have been many radical changes in attitude towards the use and abuse of the Web. The most recent surveys show that when it comes to using the WWW the two main activities that people engage in are collecting personal information, and using the Web purely for entertainment. The academic no doubt will be distressed that work and education only occupy equal third place. Academics are no longer the only people using the internet and they may feel that their research work suffers because of the 'info-tourists' on the web. Gone are the days when the only people on the network were scientists with Unix boxes. More and more people are coming online from home and the humble PC seems to have cornered the market (see *Figure 2*).

In the past scientists relied on centralized systems, with systems administrators installing and maintaining programmes. Nowadays since the installation of many programmes on PC's has been fully automated, scientists are doing it for themselves. This means that the scientist needs to be aware of the trends that are driving the internet forward. Applications will be written for platforms that

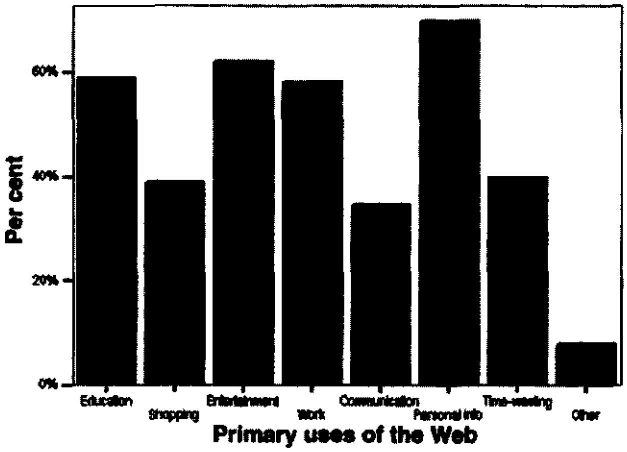


Figure 1 Primary uses of the Web. (Copyright 1994–1998 Georgia Tech Research Corporation. All rights Reserved. Source: GVU's WWW User Survey *www.gvu.gatech.edu/user_surveys*)

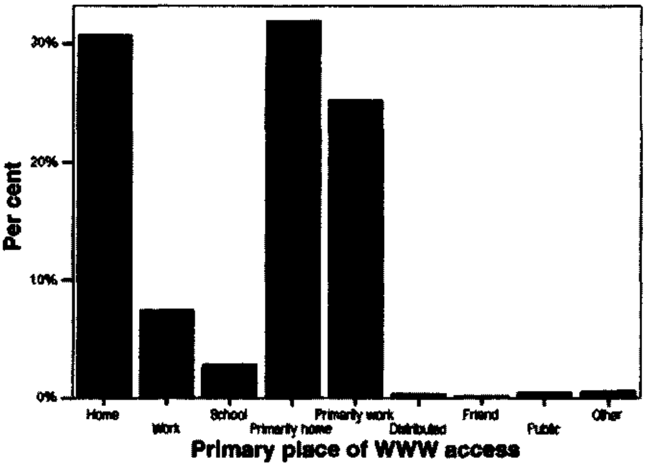


Figure 2 Primary places of WWW access. (Copyright 1994–1998 Georgia Tech Research Corporation. All rights Reserved. Source: GVU's WWW User Survey *www.gvu.gatech.edu/user_surveys*)

are being used the most. If the scientist insists that they can get by with their VT100 terminal and a text based Lynx browser very soon they will be unable to browse sites that are visually rich or rely on Java scripts or corba interfaces. It is clear from the latest survey results that the most used widely used computing platform is Windows 95 (*Figure 3*). No doubt this is partly due to the popularity of Microsoft Internet Explorer which comes bundled with the operating system. The browser wars between Netscape and Microsoft have already led to legal battles in the American courts.

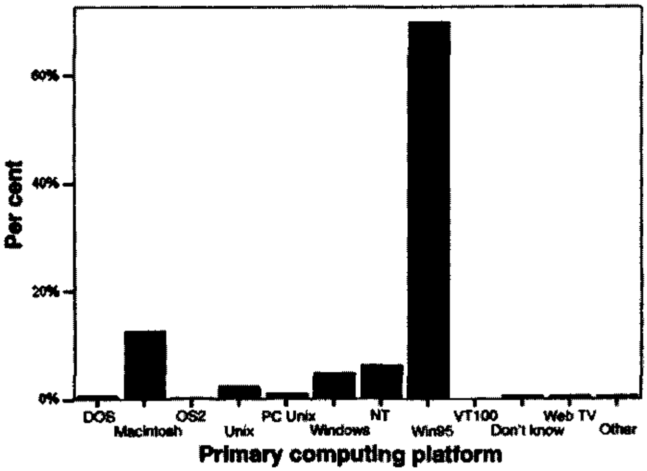


Figure 3 Primary computing platform. (Copyright 1994–1998 Georgia Tech Research Corporation. All rights Reserved. Source: GVU's WWW User Survey *www.gvu.gatech.edu/user_surveys*.)

2.1 Networking in Europe

When it comes to networking not all countries are created equal. The EMBnet organization has developed a service called 'Network Performance monitoring in EMBnet'. This project has monitored the efficiency of networking throughout Europe between the EMBnet nodes. The URL that gives the results from this project is *http://www.cmbi.kun.nl/Ping/*.

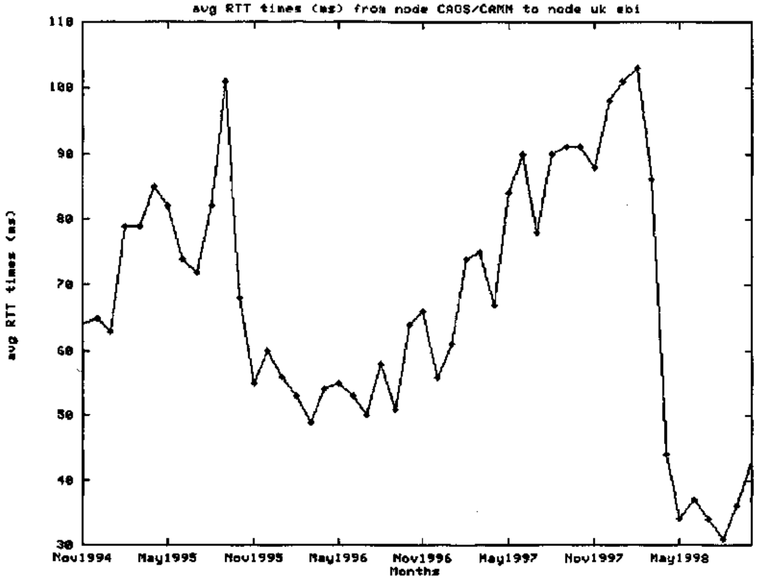


Figure 4 Average round trip times times in ms from EmbNet node CAOS/CAMM in the Netherlands to the EBI in the UK.

In 1995, a *Network Usage and Quality Advisory Group* of the Dutch Network organisation SURFnet, defined '*an upper RTT (Round Trip Time) limit of 125 msec. without packet loss*' as a minimum QOS (Quality of Service) level for interactive online work. The RTT values from the Dutch Embnet node to the EBI can be represented by the following graph.

This shows that the RTT from the Netherlands to the EBI in the UK has consistently been below the recommended time of 125 msec, which means scientists from the Netherlands, should have no difficulties in contacting the EBI web server. It is interesting to note that the results from October 1998 show that of the thirty-two nodes monitored, twenty-two have a RTT of less than 125 msec. This must surely be good news for networking within Europe (*Table 1*).

It is essential that research workers learn to use the services provided for them within their own countries. Penalties are always paid when you network across international borders. It would seem that the more borders you cross, the less efficient the network becomes. However, networking within your own country is more efficient because more often than not a basic infrastructure already exists between the major universities. In the mind of the molecular biologist, however, Mecca is either at the NCBI or EBI and that is the direction they religiously point their browsers to, only to suffer frustration when they cannot get their work done due to poor bandwidth and increased traffic directed towards these sites. For this reason EMBnet tries to co-ordinate their activities so that all the EMBnet nodes provide easy access for database query and retrieval. Many of the EMBnet nodes use a mirror package to update their databases on a daily basis, via remote ftp from the databases stored on the EBI anonymous ftp server at *ftp://ftp.ebi.ac.uk/pub/databases.*

The major databases such as EMBL or SWISS-Prot are then indexed at the EMBnet nodes and can be queried with the SRS package. SRS which was developed at EMBL Heidelberg by Thure Etzold, has been adopted by many of the EMBnet nodes throughout Europe and also abroad. SRS is also unique in that it is able to index very many different databases. A list of all EMBNET sites that use SRS is given in *Table 2.*

2.2 The way we were . . . e-mail servers for sequence retrieval

Networking for the biologist has a very short history and many of the services developed in the eighties are still in use today. Indeed for people with very bad network connections the use of E-mail servers is still the preferred method of obtaining sequences or running homology similarity searches such as FASTA or Blast. The main depositories for sequence data are are found in the UK at the European Bioinformatics Institute (EBI), and at the National Centre for Biotechnology Information (NCBI) in the United States. In addition these two institutes collaborate with the DDJBB in Japan.

Both the EBI and NBCI run e-mail servers that will allow you to retrieve sequences via e-mail. To obtain information on how to run the e-mail server at

Table 1 Table of EMBNET Internet PING results from 1–30 September 1998. These give network response times from the CAOS/CAMM centre in Nijmegen (The Netherlands) to all of the European EMBNET sites. The main figures show the average, minimum, and maximum RTT times.

Node	% Loss	Avg RTT (ms)	Min RTT (ms)	Max RTT (ms)
nl_caoscamm	0	1	1	30
uk_ucl	0	32	20	334
de_embl	98	33	26	86
be_ben	3	34	19	567
uk_hgmp	0	35	22	290
uk_ebi	1	36	22	251
uk_sanger	1	37	23	394
uk_seqnet	2	42	30	289
ch_expasy	5	42	28	328
ch_isrec	3	43	28	258
de_dkfz	5	46	31	336
fi_csc	2	50	43	247
no_bio	0	52	45	156
se_bmc	0	56	44	564
fr_infobiogen	1	56	31	249
dk_biobase	5	59	49	654
at_biocenter	11	64	49	279
es_cnb	3	65	44	415
fr_genethon	1	65	35	851
de_mips	7	67	41	1091
ie_incbi	4	74	39	558
hu_abc	6	87	50	542
it_icgeb	4	132	58	2299
it_cnr	8	176	82	1214
us_ncbi	3	181	106	2804
gr_imbb	15	188	76	808
il_inn	95	275	149	521
pt_pen	3	372	103	3724
pl_ibb	3	588	560	876
au_angis	15	669	433	1828
za_sanbi	9	816	687	2084
cn_peking	34	904	742	3160

EBI you simply send a e-mail message to *netserv@ebi.ac.uk* and include in the main body of the message the word help and full instructions will be sent via e-mail on how to operate the service.

A similar method for sequence retrieval is employed by the NCBI and the e-mail query system utilizes the Entrez retrieval system that they have developed

Table 2 A List of the EMBNET nodes and some other sites around the world which support SRS

WEHI, Melbourne, Australia
Vienna Biocenter EMBnet Node, Vienna, Austria
Belgian EMBnet Node (BEN), Brussels, Belgium
DBBM-IOC, Fiocruz, Rio de Janeiro, Brazil
CBR-NRC, Halifax, Canada The Genome Mine, Base4 Bioinformatics, Canada
CBI EMBnet Node, University of Beijing, China
CSC, Otaniemi, Espoo, Finland
INFOBIOGEN, Villejuif, France Institut Pasteur, Paris, France
LBMRPM INRA/CNRS, Auzeville, Toulouse, France
DKFZ, Heidelberg, Germany
EMBL, Heidelberg, Germany
GBF, Braunschweig, Germany
MIPS–MPG/GSF, Martinsried/Munich, Germany
Bioinformatics Centre, University of Pune, India
INCBI EMBnet Node, Dublin, Ireland
Weizmann Institute BCD, Rehovot, Israel
CNR EMBnet Node, Bari, Italy
CRISCEB, Second University of Naples, Italy
IVR, Kyoto University, Japan
Biotek EMBnet Node, Oslo, Norway
IBB-PAS EMBnet Node, Warsaw, Poland
IGC EMBnet Node, Oeiras, Portugal
SRCG, Novosibirsk, Siberia, Russia
BIC, National University Hospital, Singapore
CNB EMBnet Node, Madrid, Spain
Biomedical Centre (BMC), Uppsala, Sweden
ExPASy, Geneva, Switzerland
CAOS/CAMM Center, Nijmegen, The Netherlands
RIGEB-MRC, Gebze, Kocaeli, Turkey
Adlib, CAB International, Wallingford, UK
EMBL-EBI, Hinxton, Cambridge, UK
HGMP-RC, Hinxton, Cambridge, UK
MBDC Oxford, Oxford University, UK
SEQNET EMBnet Node, Daresbury, UK
Sanger Centre, Hinxton, Cambridge, UK
IUBio, Indiana University, USA

for their website. Many people would argue that getting sequence via e-mail is old-fashioned technology. It is primitive in that it only delivers simple ascii-formatted text. However the e-mail query server at the NCBI is clever enough to be able to return the sequence to you in a variety of different formats including GenBank, FASTA, or Html.

Protocol 1

Using *netserv@ebi.ac.uk* for the retrieval of sequences and software

To request:

- Specific help on the sequence databases such as EMBL or SWISS-PROT
- General help on software
- The sequence with accession number X03392 (nucleotide)
- The sequence called PIP03XX (nucleotide)
- The sequence called WAP_MOUSE (protein)
- The sequence submission form

You would write the following commands directly into the body of an e-mail message:

```
HELP NUC
HELP PROT
HELP SOFTWARE
GET NUC:PIP03XX
GET NUC:X03392
GET PROT:WAP_MOUSE
GET DOC:DATASUB.TXT
```

and then mail the commands to the e-mail address netserv@ebi.ac.uk You would then receive the results back in your mailbox via e-mail.

It is often more convenient to shoot off a query by e-mail and get an answer within a few minutes than it is to struggle with trying to access a website that has bandwidth problems. The address for the NCBI e-mail server is at *query@ncbi.nlm.nih.gov*. To receive full instructions on how the server works just send an e-mail message to *query@ncbi.nlm.nih.gov* and in the main body of the message type the word help. I have often found that people who have used an e-mail server generally have a better understanding of databases and sequence retrieval than those who have only used a WWW interface.

Protocol 2

Using query@ncbi.nlm.nih.gov for sequence retrieval

Examples:

```
DB n
UID U30150
```

Will search the nucleotide database for an entry whose accession number is U30150. Since no DOPT line is present, the record will be displayed the record in the default GenBank format.

```
DB n
UID U30150,U30153
DOPT f
```

Protocol 2 continued

Will search the nucleotide database for entries whose accession numbers are U30150 and U30153, and display them in FASTA format.

```
DB m
UID 88055872
DOPT r
HTML
```

Will search the MEDLINE database for the record with MEDLINE UID 88055872 and display it in MEDLINE Report format. Send the results in HTML format for viewing through a WWW browser.

```
DB p
UID sp|P11598|
DOPT m
```

Will search the protein database, using a FASTA formatted UID, to retrieve the entry whose Swiss-Prot accession number is P11598, and display the MEDLINE links for that protein record as document summaries.

2.3 Similarity searches via e-mail

The two most popular e-mail servers dealing with similarity searches are Blast from the NCBI. and FASTA from EBI. For help regarding these e-mail servers you can send an e-mail message to either *blast@ncbi.nlm.nih.gov* or *fasta@ebi.ac.uk* and complete instructions on how to formulate an e-mail message to be processed by these servers will be returned to you via e-mail. Again it should be stressed that once you understand how to compose an e-mail message to submit a Blast query via E-mail, then you can be more discriminating when you are asked to repeat the procedure via the WWW. As it is most people just opt for the default parameters and never experiment with different options.

Protocol 3

Blast similarity search e-mail server at NCBI

To submit a Blast similarity search at the NCBI a e-mail message should be composed as follows.

```
From: rab.c.nesbit@goven.com Tue Jul 28 21:36:38 1998
Date: 28 Jul 1998 21:29:02-EDT
To: blast@ncbi.nlm.nih.gov
Subject:
PROGRAM blastn
DATALIB month
EXPECT 0.75
```

Protocol 3 continued

```
BEGIN
>XYZ012 mygene XYZ
tgcttggctgaggagccataggacgagagcttcctggtgaagtgtgtttcttgaaatcat
```

The actual search request begins with the mandatory parameter 'PROGRAM' in the first column followed by the value 'blastn' (the name of the program) for searching nucleic acids. The next line contains the mandatory search parameter 'DATALIB' with the value 'month' for the newest nucleic acid sequences. The third line contains an optional EXPECT parameter and the value desired for it. The fourth line contains the mandatory 'BEGIN' directive, followed by the query sequence in FASTA/Pearson format. Each line of information must be less than 80 characters in length. Once the e-mail message has been sent it will be processed automatically at the NCBI and the results returned to your e-mail address once they have been computed.

The BLAST algorithm was developed by the National Center for Biotechnology Information at the National Library of Medicine. The BLAST family of programs employs this algorithm to compare an amino acid query sequence against a protein sequence database or a nucleotide query sequence against a nucleotide sequence database, as well as other combinations of protein and nucleic acid. If you use BLAST as a tool in your published research, the following reference should be cited:

> Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997).Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, Sept. 1, **25**(17), 3389.

It used to be that the NCBI exclusively provided access to BLAST but in recent years you can now run BLAST searches from many different sites around the world, which is a clear indication that this programme has become a very popular method for doing homology searches. The fact that it appears in so many places may be due to the fact that it is available for free from the NCBI anonymous ftp server at *ftp://ftp.ncbi.nlm.nih.gov/blast/*.

Historically the EBI has always provided homology searches through FASTA. The following reference should be cited when you have used FASTA:

> Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence analysis. *Proc. Natl. Acad. Sci. USA*, **85**, 2444.

Web-based FASTA applications can be found at the EBI *http://www.ebi.ac.uk/fasta3* and at DDJB *http://www.ddbj.nig.ac.jp/E-mail/homology.html*

Protocol 4

FASTA similarity search e-mail server at EBI

EXAMPLE OF A SIMPLE SUBMISSION

```
PATH mary.doll@goven.com
TITLE My Sequence
LIB swall
SEQ
MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNAQDFC
TDLAVSSANFIPTVTAISTSPDLQWLVQPALVSSVAPSQTRAPHPFGVPA
END
```

The PATH is the e-mail address of the person to whom the results should be sent. TITLE is anything that you want to appear on the subject line of the returned mail. LIB is the database that you want to search against. The sequence itself should be enclosed between SEQ and END.

2.4 Speed solutions for similarity searches

In recent years there has been an increase in the use of specialized hardware for doing similarity searches. Four companies in particular have pioneered this approach, and the turn around time for running a search against the whole of Swiss-Prot has been reduced to around 10 seconds using the Smith–Waterman algorithm.

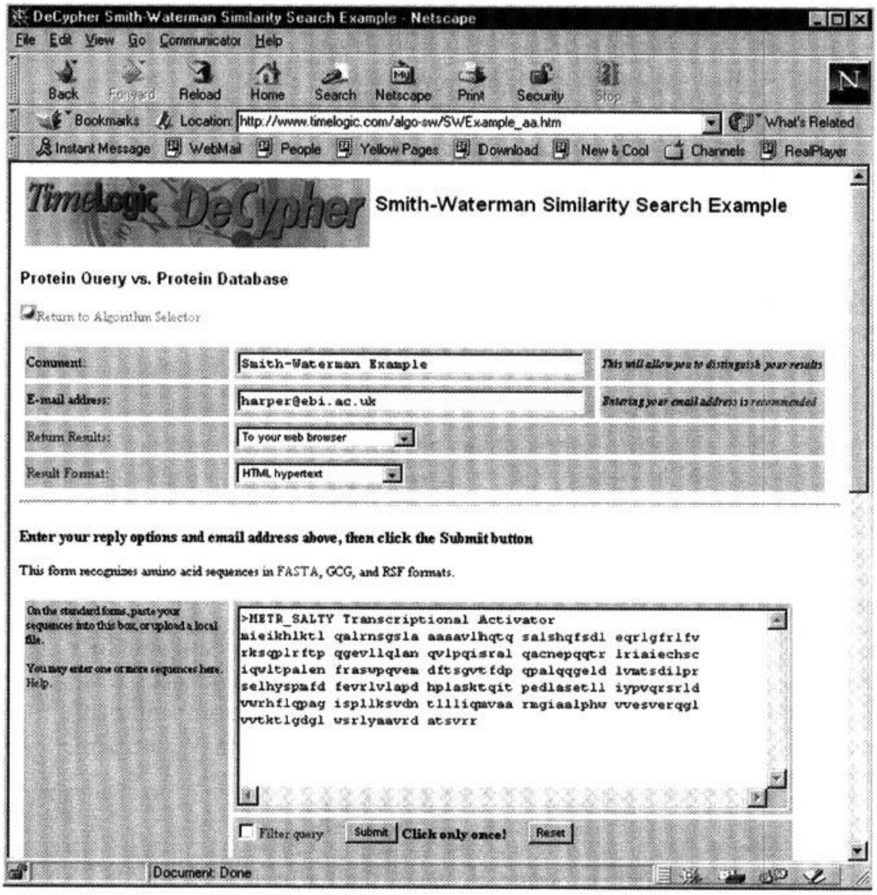
2.4.1 Time Logic

Time Logic (*http://www.timelogic.com*) from the USA has introduced DeCypher Bioinformatics Accelerators and they have implemented a number of algorithms namely, Gapped BLAST 2 (includes entire heuristic search suite: blastn, blastp, blastx, tblastn, tblastx) PSI-BLAST, Affine Smith–Waterman, FrameSearch, ProfileSearch, ProfileScan, and ClustalW with graphical rendition of dendrogram (Java applet). The WWW query interface for a Smith–Waterman similarity search is shown in *Figure 5* and some results from that search are given in *Figure 6*. Timelogic also have a Blast search running on their hardware at the NCGR at the URL:

http://seqsim.ncgr.org/newBlast.html

2.4.2 Compugen

Compugen have succeeded in introducing the Biocellorator to many pharmaceutical companies to aid them in their search for new and novel drugs. The EBI has a biocellorator, which is online and is available for public use. At the EBI there are two different interfaces to this service. The one provided by Compugen called GeneWeb and a simple custom interface developed at the EBI at *http://www.ebi.ac.uk/bic_sw*. The interface to the BIC-SW at the EBI is very compact


DeCypher Smith-Waterman Similarity Search Example - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks Location: http://www.timelogic.com/algo-sw/SWExample_aa.htm
What's Related
Instant Message WebMail People Yellow Pages Download New & Cool Channels RealPlayer
TimeLogic DeCypher
Document: Done

Smith-Waterman Similarity Search Example

Protein Query vs. Protein Database

Return to Algorithm Selector

Comment:	Smith-Waterman Example	*This will allow you to distinguish your results*
E-mail address:	harper@ebi.ac.uk	*Entering your email address is recommended*
Return Results:	To your web browser	
Result Format:	HTML hypertext	

Enter your reply options and email address above, then click the Submit button

This form recognizes amino acid sequences in FASTA, GCG, and RSF formats.

On the standard forms, paste your sequences into this box, or upload a local file.

You may enter one or more sequences here. Help.

```
>METR_SALTY Transcriptional Activator
mieikhlktl qalrnsgsla aaaavlhqtq salshqfsdl eqrlgfrlfv
rksqplrftp qgevllqlan qvlpqisral qacnepqqtr iriaiechsc
iqwltpalen fraswpqvem dftsgvtfdp qpalqqgeld lvmtsdilpr
selhyspmfd fevrlvlapd hplasktqit pedlasetll iypvqrsrld
vwrhflqpag ispllksvdn tllliqmvaa rmgiaalphw vvesverqgl
vvtktlgdgl wsrlyaavrd atsvrr
```

☐ Filter query Submit **Click only once!** Reset

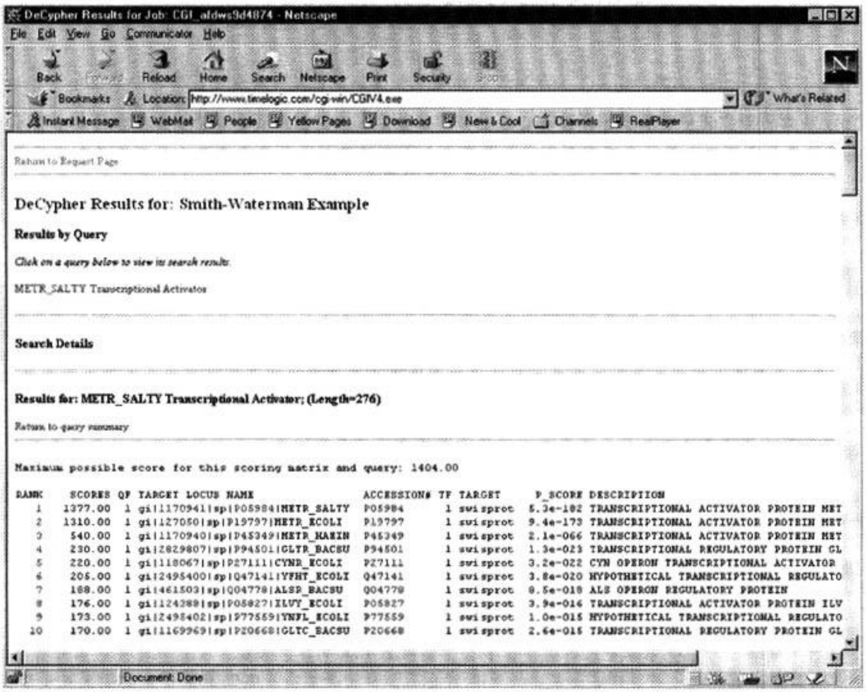
Figure 5 DeCypher interface for Smith–Waterman similarity searches.

and easy to use. Most people just accept the default settings, paste in their query sequences and run the program. The interface page is shown in *Figure 7* and some results are shown in *Figure* 8.

If you are from the Mediterranean area then perhaps it would be more convenient to try the GeneWeb interface from the Weizmann Institute in Israel, which is also open to the public for unregistered users. The URL is (*http://sgbcd.weizmann.ac.il:80/cgi-bin/genweb/main.cgi*)

2.4.3 Paracel FDF

At the Swiss EMBnet node you can find the Paracel Fast Data finder (FDF), which is designed to help bioinformatics departments dramatically increase the rate at which they can find high-scoring potential genomic targets. Paracel claim that GeneMatcher is the first commercially available genetic data analysis system to use custom ASIC technology that can analyse similarities or differences in DNA or protein sequences up to 1000 times faster than traditional computer systems.


DeCypher Results for Job: CGI_afdws3d4874 - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks Location: http://www.timelogic.com/cgi-win/CGIV4.exe
What's Related
Instant Message WebMail People Yellow Pages Download New & Cool Channels RealPlayer
Document: Done

Return to Request Page

DeCypher Results for: Smith-Waterman Example

Results by Query

Click on a query below to view its search results

METR_SALTY Transcriptional Activator

Search Details

Results for: METR_SALTY Transcriptional Activator; (Length=276)

Return to query summary

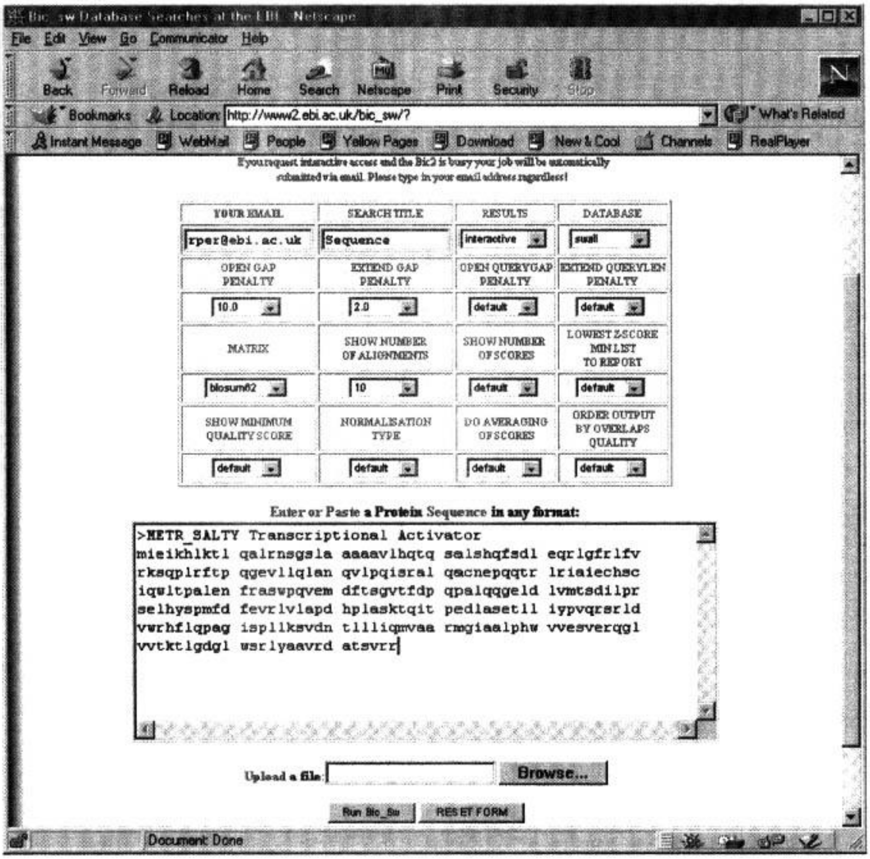
```
Maximum possible score for this scoring matrix and query: 1404.00

RANK    SCORES QF TARGET LOCUS NAME                  ACCESSION# TF TARGET     P_SCORE DESCRIPTION
   1   1377.00  1 gi|1170941|sp|P05984|METR_SALTY   P05984      1 swissprot  5.3e-182 TRANSCRIPTIONAL ACTIVATOR PROTEIN MET
   2   1310.00  1 gi|127050|sp|P19797|METR_ECOLI    P19797      1 swissprot  9.4e-173 TRANSCRIPTIONAL ACTIVATOR PROTEIN MET
   3    540.00  1 gi|1170940|sp|P45349|METR_HAEIN   P45349      1 swissprot  2.1e-066 TRANSCRIPTIONAL ACTIVATOR PROTEIN MET
   4    230.00  1 gi|2829807|sp|P94501|GLTR_BACSU   P94501      1 swissprot  1.3e-023 TRANSCRIPTIONAL REGULATORY PROTEIN GL
   5    220.00  1 gi|118067|sp|P27111|CYNR_ECOLI    P27111      1 swissprot  3.2e-022 CYN OPERON TRANSCRIPTIONAL ACTIVATOR
   6    205.00  1 gi|2495400|sp|Q47141|YFHT_ECOLI   Q47141      1 swissprot  3.8e-020 HYPOTHETICAL TRANSCRIPTIONAL REGULATO
   7    188.00  1 gi|461503|sp|Q04778|ALSR_BACSU    Q04778      1 swissprot  8.5e-018 ALS OPERON REGULATORY PROTEIN
   8    176.00  1 gi|124289|sp|P05827|ILVY_ECOLI    P05827      1 swissprot  3.9e-016 TRANSCRIPTIONAL ACTIVATOR PROTEIN ILV
   9    173.00  1 gi|2495402|sp|P77559|YNFL_ECOLI   P77559      1 swissprot  1.0e-015 HYPOTHETICAL TRANSCRIPTIONAL REGULATO
  10    170.00  1 gi|1169969|sp|P20668|GLTC_BACSU   P20668      1 swissprot  2.6e-015 TRANSCRIPTIONAL REGULATORY PROTEIN GL
```

Figure 6 Results from Decypher Smith–Waterman similarity search.

Competition to discover novel genes is of great interest to pharmaceutical companies because if it is possible to identify just one critical target gene then this can result in an application for a patent on a product. Therefore any method that combines speed with sensitivity is a very valuable tool in the hands of the research worker. The main search interface and some results are shown in *Figures 9* and *10*.

3 Sequence retrieval via the WWW

If you are in a country with a poor Internet connection then working with E-mail servers for the retrieval of sequences is often the best option. However, there are many excellent servers in different parts of the world and they should not be ignored, even if you do live on the other side of the planet. Your geographical location should be the first consideration when accessing a remote site. It is best to access a site that is in close proximity. Two of the most popular services for sequence retrieval are Entrez from the NCBI and SRS from the EBI. However there are other options available and if you are in the Pacific rim area then it might be worthwhile to look at the services offered by DDBJ in Japan *http://www.ddbj.nig.ac.jp/searches-e.html* or the Maestro service from the National Centre for Genomic Research (NCGR) *http://www.ncgr.org/gsdb/maestro/index.html* on the West Coast of the USA.


Bic_sw Database Searches at the EBI - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks
What's Related
Instant Message WebMail People Yellow Pages Download New & Cool Channels RealPlayer
Browse...
Run Bic_Sw
RESET FORM
Document: Done

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

If you request interactive access and the Bic2 is busy your job will be automatically submitted via email. Please type in your email address regardless!

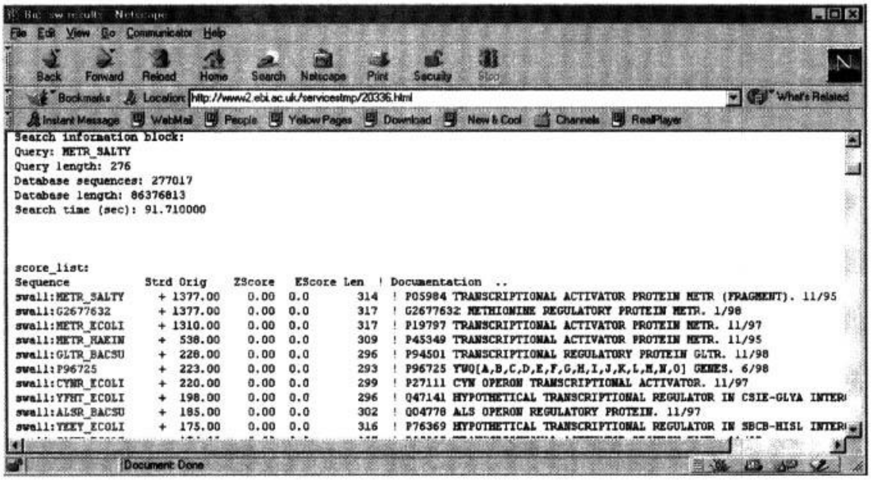
YOUR EMAIL	SEARCH TITLE	RESULTS	DATABASE
rper@ebi.ac.uk	Sequence	interactive	swall
OPEN GAP PENALTY	EXTEND GAP PENALTY	OPEN QUERYGAP PENALTY	EXTEND QUERYLEN PENALTY
10.0	2.0	default	default
MATRIX	SHOW NUMBER OF ALIGNMENTS	SHOW NUMBER OF SCORES	LOWEST Z-SCORE MIN LIST TO REPORT
blosum62	10	default	default
SHOW MINIMUM QUALITY SCORE	NORMALISATION TYPE	DO AVERAGING OF SCORES	ORDER OUTPUT BY OVERLAPS QUALITY
default	default	default	default

Enter or Paste **a Protein** Sequence **in any format:**

```
>METR_SALTY Transcriptional Activator
mieikhlktl qairnsgsla aaaavlhqtq salshqfsdl eqrlgfrlfv
rksqplrftp qgevllqlan qvlpqisral qacnepqqtr lriaiechsc
iqwltpalen fraswpqvem dftsgvtfdp qpalqqgeld lvmtsdilpr
selhyspmfd fevrlvlapd hplasktqit pedlasetll iypvqrsrld
vwrhflqpag ispllksvdn tllliqmvaa rmgiaalphw vvesverqgl
vvtktlgdgl wsrlyaavrd atsvrr
```

Upload a file:

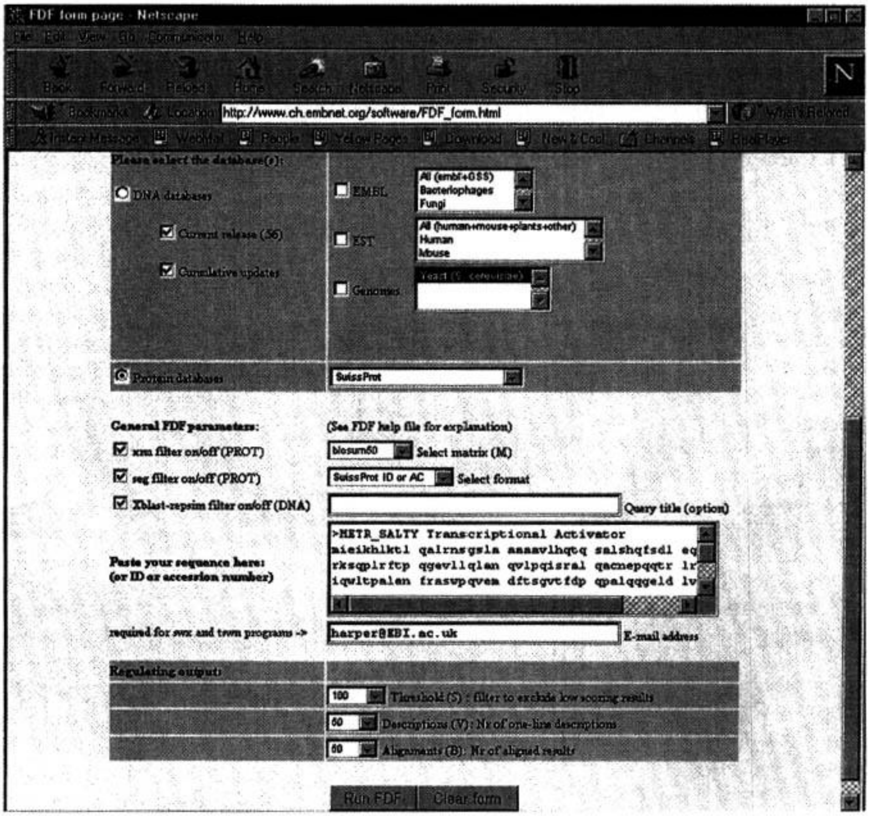
Figure 7 BIC-SW interface for performing Smith–Waterman homology searches.


Bic_sw results - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks
What's Related
Instant Message WebMail People Yellow Pages Download New & Cool Channels RealPlayer
Document: Done

Location: http://www2.ebi.ac.uk/servicestmp/20336.html

```
Search information block:
Query: METR_SALTY
Query length: 276
Database sequences: 277017
Database length: 86376813
Search time (sec): 91.710000



score_list:
Sequence             Strd Orig     ZScore   EScore Len  ! Documentation  ..
swall:METR_SALTY      + 1377.00     0.00   0.0       314  ! P05984 TRANSCRIPTIONAL ACTIVATOR PROTEIN METR (FRAGMENT). 11/95
swall:G2677632        + 1377.00     0.00   0.0       317  ! G2677632 METHIONINE REGULATORY PROTEIN METR. 1/98
swall:METR_ECOLI      + 1310.00     0.00   0.0       317  ! P19797 TRANSCRIPTIONAL ACTIVATOR PROTEIN METR. 11/97
swall:METR_HAEIN      +  538.00     0.00   0.0       309  ! P45349 TRANSCRIPTIONAL ACTIVATOR PROTEIN METR. 11/95
swall:GLTR_BACSU      +  228.00     0.00   0.0       296  ! P94501 TRANSCRIPTIONAL REGULATORY PROTEIN GLTR. 11/98
swall:P96725          +  223.00     0.00   0.0       293  ! P96725 YUQ[A,B,C,D,E,F,G,H,I,J,K,L,M,N,O] GENES. 6/98
swall:CYNR_ECOLI      +  220.00     0.00   0.0       299  ! P27111 CYN OPERON TRANSCRIPTIONAL ACTIVATOR. 11/97
swall:YFHT_ECOLI      +  198.00     0.00   0.0       296  ! Q47141 HYPOTHETICAL TRANSCRIPTIONAL REGULATOR IN CSIE-GLYA INTER
swall:ALSR_BACSU      +  185.00     0.00   0.0       302  ! O04778 ALS OPERON REGULATORY PROTEIN. 11/97
swall:YEEY_ECOLI      +  175.00     0.00   0.0       316  ! P76369 HYPOTHETICAL TRANSCRIPTIONAL REGULATOR IN SBCB-HISL INTER
```

Figure 8 Results from BIC-SW similarity search at the EBI.


FDF form page - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks Location
What's Related
Instant Message WebMail People Yellow Pages Download New & Cool Channels RealPlayer
Run FDF
Clear form

Location: http://www.ch.embnet.org/software/FDF_form.html

Please select the database(s):

○ DNA databases

☑ Current release (56)

☑ Cumulative updates

□ EMBL: All (embl+GSS) / Bacteriophages / Fungi

□ EST: All (human+mouse+plants+other) / Human / Mouse

□ Genomes: Yeast (S. cerevisiae)

● Protein databases: SwissProt

General FDF parameters: (See FDF help file for explanation)

☑ xnu filter on/off (PROT) — blosum60 Select matrix (M)

☑ seg filter on/off (PROT) — SwissProt ID or AC Select format

☑ Xblast-repsim filter on/off (DNA) — Query title (option)

Paste your sequence here: (or ID or accession number)

```
>METR_SALTY Transcriptional Activator
mieikhlktl qalrnsgsla aaaavlhqtq salshqfsdl eq
rksqplrftp qgevllqlan qvlpqisral qacnepqqtr lr
iqwltpalen fraswpqvem dftsgvtfdp qpalqqgeld lv
```

required for swx and twn programs -> harper@EBI.ac.uk E-mail address

Regulating output:

100 Threshold (S) : filter to exclude low scoring results

60 Descriptions (V): Nr of one-line descriptions

60 Alignments (B): Nr of aligned results

Figure 9 Paracel FDF interface at the Swiss EMBnet node.

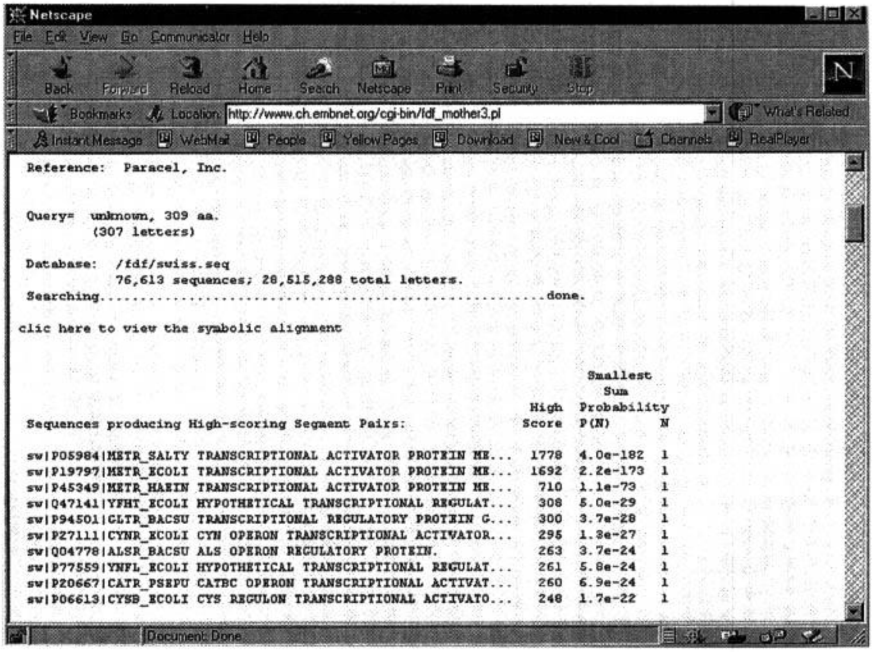
3.1 Entrez from the NCBI

The NCBI is the only place in the world where you will find the Entrez service and it concentrates on a few databases namely, nucleotide sequences, amino acid sequences, 3-D structures, Genomes, Taxonomy, and Literature-PubMed *http://www.ncbi.nlm.nih.gov/Entrez/*. One of its strengths is that it provides access to PubMed and this is a key factor in its popularity and success. Effective August 3, 1998, NLM implemented a system enhancement that dramatically increases the speed of the system. This redesign of the way PubMed stored and retrieved information will improve users search time—a search that previously took approximately 18 seconds to run in PubMed now runs under 2 seconds.

In Entrez you select the database you wish to query, for example the protein database and then you are allowed to string a number of keywords together, like 'Rhizobium Ausubel nodulation' and those entries that meet the criteria will be displayed. An example is shown in *Figure 11*.

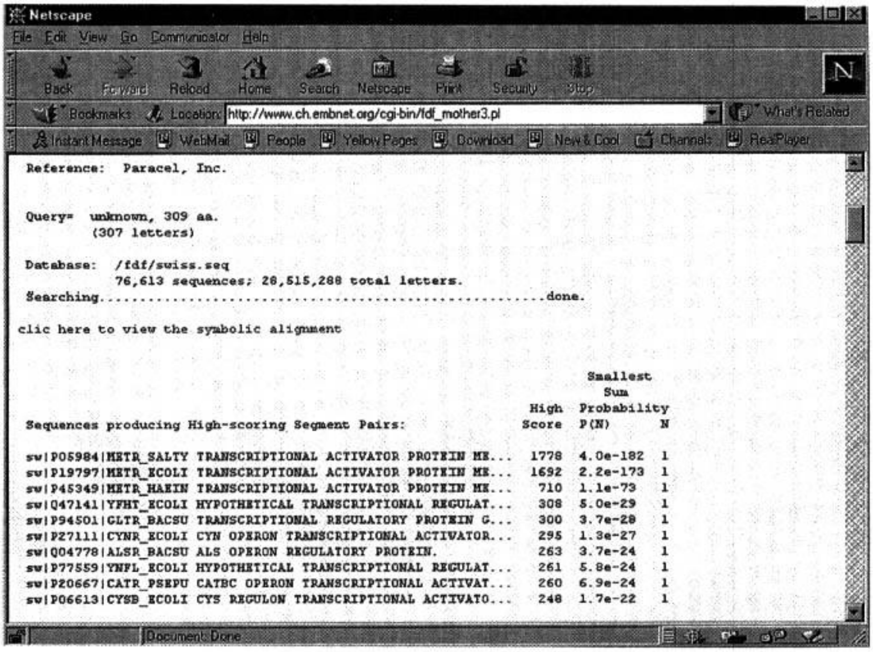
3.2 SRS from the EBI

SRS *http://srs.ebi.ac.uk/* is a very powerful tool for querying databases and it would seem to be the preferred querying system within Europe. You select a database


Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks Location: http://www.ch.embnet.org/cgi-bin/fdf_mother3.pl
What's Related
Instant Message WebMail People Yellow Pages Download New & Cool Channels RealPlayer
Document: Done

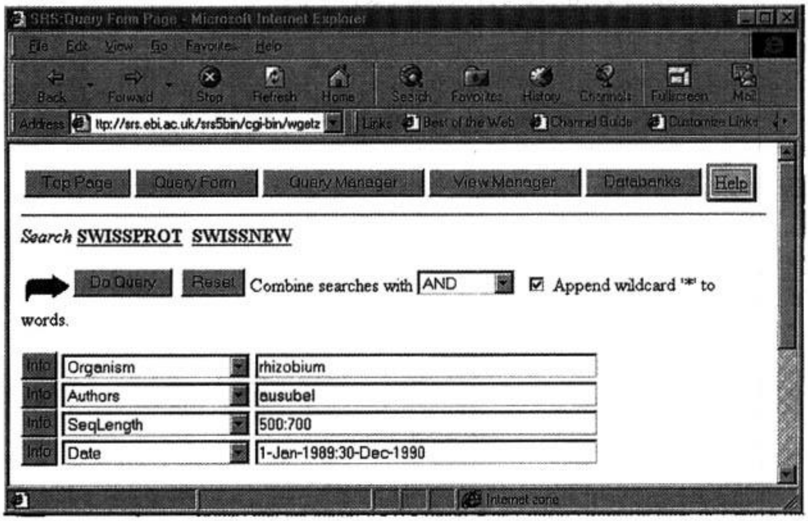
```
Reference:  Paracel, Inc.


Query=  unknown, 309 aa.
       (307 letters)

Database:   /fdf/swiss.seq
           76,613 sequences; 28,515,288 total letters.
Searching..................................................done.

clic here to view the symbolic alignment


                                                                    Smallest
                                                                      Sum
                                                             High  Probability
Sequences producing High-scoring Segment Pairs:              Score  P(N)      N

sw|P05984|METR_SALTY TRANSCRIPTIONAL ACTIVATOR PROTEIN ME...  1778  4.0e-182  1
sw|P19797|METR_ECOLI TRANSCRIPTIONAL ACTIVATOR PROTEIN ME...  1692  2.2e-173  1
sw|P45349|METR_HAEIN TRANSCRIPTIONAL ACTIVATOR PROTEIN ME...   710  1.1e-73   1
sw|Q47141|YFHT_ECOLI HYPOTHETICAL TRANSCRIPTIONAL REGULAT...   308  5.0e-29   1
sw|P94501|GLTR_BACSU TRANSCRIPTIONAL REGULATORY PROTEIN G...   300  3.7e-28   1
sw|P27111|CYNR_ECOLI CYN OPERON TRANSCRIPTIONAL ACTIVATOR...   295  1.3e-27   1
sw|Q04778|ALSR_BACSU ALS OPERON REGULATORY PROTEIN.            263  3.7e-24   1
sw|P77559|YNFL_ECOLI HYPOTHETICAL TRANSCRIPTIONAL REGULAT...   261  5.8e-24   1
sw|P20667|CATR_PSEPU CATBC OPERON TRANSCRIPTIONAL ACTIVAT...   260  6.9e-24   1
sw|P06613|CYSB_ECOLI CYS REGULON TRANSCRIPTIONAL ACTIVATO...   248  1.7e-22   1
```

Figure 10 The first few lines of some search results from an FDF search at the Swiss EMBnet node.


Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks Location: http://www.ch.embnet.org/cgi-bin/fdf_mother3.pl
What's Related
Instant Message WebMail People Yellow Pages Download New & Cool Channels RealPlayer
Document: Done

```
Reference:  Paracel, Inc.


Query=  unknown, 309 aa.
       (307 letters)

Database:   /fdf/swiss.seq
           76,613 sequences; 28,515,288 total letters.
Searching..................................................done.

clic here to view the symbolic alignment


                                                                    Smallest
                                                                      Sum
                                                             High  Probability
Sequences producing High-scoring Segment Pairs:              Score  P(N)      N

sw|P05984|METR_SALTY TRANSCRIPTIONAL ACTIVATOR PROTEIN ME...  1778  4.0e-182  1
sw|P19797|METR_ECOLI TRANSCRIPTIONAL ACTIVATOR PROTEIN ME...  1692  2.2e-173  1
sw|P45349|METR_HAEIN TRANSCRIPTIONAL ACTIVATOR PROTEIN ME...   710  1.1e-73   1
sw|Q47141|YFHT_ECOLI HYPOTHETICAL TRANSCRIPTIONAL REGULAT...   308  5.0e-29   1
sw|P94501|GLTR_BACSU TRANSCRIPTIONAL REGULATORY PROTEIN G...   300  3.7e-28   1
sw|P27111|CYNR_ECOLI CYN OPERON TRANSCRIPTIONAL ACTIVATOR...   295  1.3e-27   1
sw|Q04778|ALSR_BACSU ALS OPERON REGULATORY PROTEIN.            263  3.7e-24   1
sw|P77559|YNFL_ECOLI HYPOTHETICAL TRANSCRIPTIONAL REGULAT...   261  5.8e-24   1
sw|P20667|CATR_PSEPU CATBC OPERON TRANSCRIPTIONAL ACTIVAT...   260  6.9e-24   1
sw|P06613|CYSB_ECOLI CYS REGULON TRANSCRIPTIONAL ACTIVATO...   248  1.7e-22   1
```

Figure 11 An example query using Entrez and the NCBI.


SRS:Query Form Page - Microsoft Internet Explorer
File Edit View Go Favorites Help
Back Forward Stop Refresh Home Search Favorites History Channels Fullscreen Mail
Links Best of the Web Channel Guide Customize Links
Top Page
Query Form
Query Manager
View Manager
Databanks
Help
Do Query
Reset
Internet zone

Address ttp://srs.ebi.ac.uk/srs5bin/cgi-bin/wgetz

Search SWISSPROT SWISSNEW

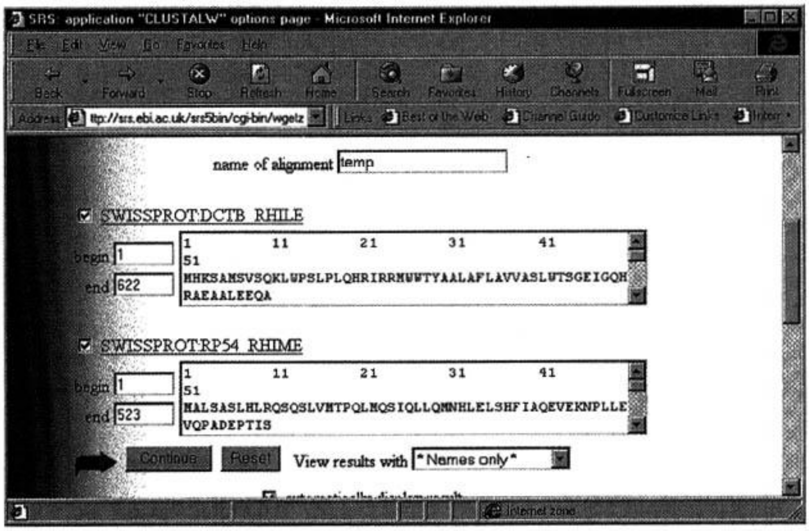
Combine searches with AND ☑ Append wildcard '*' to words.

Organism	rhizobium
Authors	ausubel
SeqLength	500:700
Date	1-Jan-1989:30-Dec-1990

Figure 12 An example query using SRS at the EBI.

and fill in your search criteria as keywords. For example in *Figure 12*, we see a sample query using the fields Organism (Rhizobium), authors (Ausubel), SeqLength (a range 500:700) and the date (a range 1-Jan-1998:30-Dec-1990).

SRS will then display two hits in Swiss-Prot for that particular year with a sequence range between 500 and 700 (*Figure 13*). It should also be noted that SRS also gives the possibility to launch an application such as BLAST or FASTA for any of the sequences that you care to select. You may also select different views of a sequence. For example the FASTA format, which then allows you to launch a multiple sequence alignment using ClustalW, directly from within SRS. This


SRS: application "CLUSTALW" options page - Microsoft Internet Explorer
File Edit View Go Favorites Help
Back Forward Stop Refresh Home Search Favorites History Channels Fullscreen Mail Print
Links Best of the Web Channel Guide Customize Links Internet
Continue
Reset
Internet zone

Address ttp://srs.ebi.ac.uk/srs5bin/cgi-bin/wgetz

name of alignment temp

☑ SWISSPROT:DCTB_RHILE

begin 1
end 622

```
1         11        21        31        41
51
MHKSAMSVSQKLWPSLPLQHRIRRMWWTYAALAFLAVVASLWTSGEIGQH
RAEAALEEQA
```

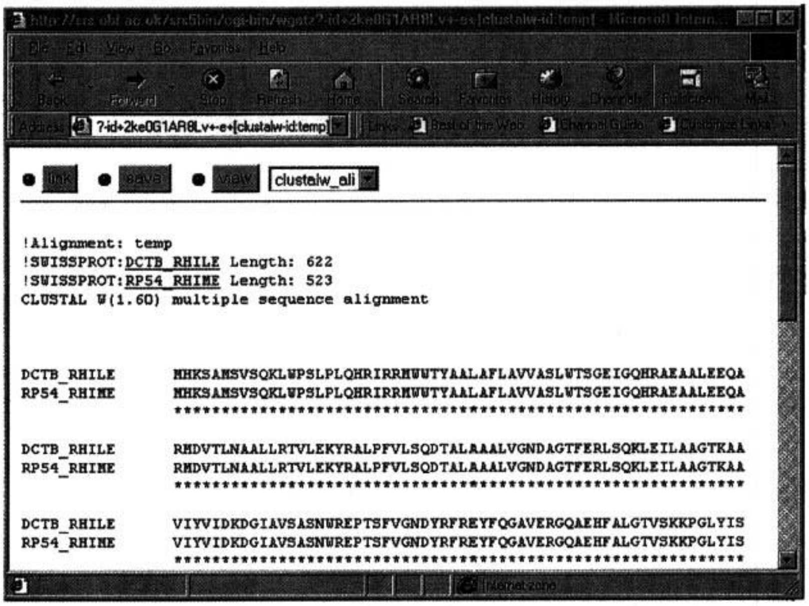
☑ SWISSPROT:RP54_RHIME

begin 1
end 523

```
1         11        21        31        41
51
MALSASLHLRQSQSLVMTPQLMQSIQLLQMNHLELSHFIAQEVEKNPLLE
VQPADEPTIS
```

View results with *Names only*

Figure 13 The results page from the query in Figure 12.



```
!Alignment: temp
!SWISSPROT:DCTB_RHILE Length: 622
!SWISSPROT:RP54_RHIME Length: 523
CLUSTAL W(1.60) multiple sequence alignment


DCTB_RHILE      MHKSAMSVSQKLWPSLPLQHRIRRMWWTYAALAFLAVVASLWTSGEIGQHRAEAALEEQA
RP54_RHIME      MHKSAMSVSQKLWPSLPLQHRIRRMWWTYAALAFLAVVASLWTSGEIGQHRAEAALEEQA
                ************************************************************

DCTB_RHILE      RMDVTLNAALLRTVLEKYRALPFVLSQDTALAAALVGNDAGTFERLSQKLEILAAGTKAA
RP54_RHIME      RMDVTLNAALLRTVLEKYRALPFVLSQDTALAAALVGNDAGTFERLSQKLEILAAGTKAA
                ************************************************************

DCTB_RHILE      VIYVIDKDGIAVSASNWREPTSFVGNDYRFREYFQGAVERGQAEHFALGTVSKKPGLYIS
RP54_RHIME      VIYVIDKDGIAVSASNWREPTSFVGNDYRFREYFQGAVERGQAEHFALGTVSKKPGLYIS
                ************************************************************
```

Figure 14 The results from launching clustalw as an application on the results in Figure 13.

method is a great time saver since there is no need to cut and paste your sequence into a separate ClustalW application (*Figure 14*).

4 Submitting sequences

Not only does the research worker want to query, retrieve and analyse sequences, occasionally they also want to submit their own sequences to the databanks be it GenBank or EMBL. The three major organizations that collect sequence information work in collaboration with each other so that sequences entered into GenBank are transferred daily by FTP to both EBI and DDBJ (and vice versa) in an attempt to keep the major databases synchronized.

At any given time the three institutes are continually swapping data so it is a false idea to believe that any one database is more current than the other. All three institutes have online methods of submitting sequence data through the Web. The NCBI were the first to come online with BANKIT. The EBI then followed with WEBIN and the Japanese at DDBJ have Sakuara.

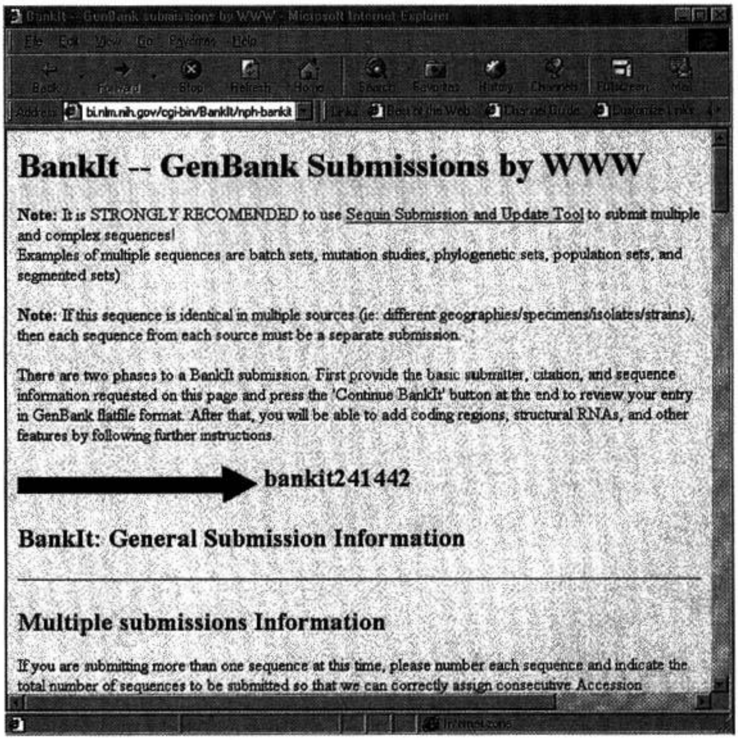
It should also be noted that the NCBI developed a stand-alone programme for MAC's, PC's, and Unix called SEQUIN that allows the end-user to enter their data from a personal computer and to send the submission via e-mail or to simply post the disk to the appropriate institute where it is then uploaded into the database. Sequin is strongly recommended if you have bulk submissions to make.

4.1 Bankit at NCBI

Bankit is convenient for quick submission of sequence data to the NCBI. BankIt allows you to enter sequence information into a form, edit as necessary, and add biological annotation (e.g. coding regions, mRNA features). BankIt transforms your data into GenBank format for you to review and when your record is completed, it can be submitted directly to GenBank. You have the option of adding information by using text boxes to describe in your own words the source of the sequence and its biological features. The entry screen from Bankit is shown in *Figure 15*. The GenBank annotation staff reviews the submitted textual information, incorporates it into the appropriate structured fields, and returns the record by e-mail for your review.

4.2 Sequin from NCBI

Sequin is stand-alone program for the MAC, PC/Windows and UNIX. Sequin is an interactive, graphically oriented program based on screen forms and controlled vocabularies that guide you through the process of entering your sequence and providing biological and bibliographic annotation. Sequin is designed to simplify the sequence submission process and to provide graphical viewing and editing options. This program is optimal for submitting multiple sequences, mutation studies, phylogenetic sets, population sets, and segmented sets. It incorporates



BankIt -- GenBank Submissions by WWW

Note: It is STRONGLY RECOMENDED to use Sequin Submission and Update Tool to submit multiple and complex sequences!
Examples of multiple sequences are batch sets, mutation studies, phylogenetic sets, population sets, and segmented sets)

Note: If this sequence is identical in multiple sources (ie: different geographies/specimens/isolates/strains), then each sequence from each source must be a separate submission.

There are two phases to a BankIt submission. First provide the basic submitter, citation, and sequence information requested on this page and press the 'Continue BankIt' button at the end to review your entry in GenBank flatfile format. After that, you will be able to add coding regions, structural RNAs, and other features by following further instructions.

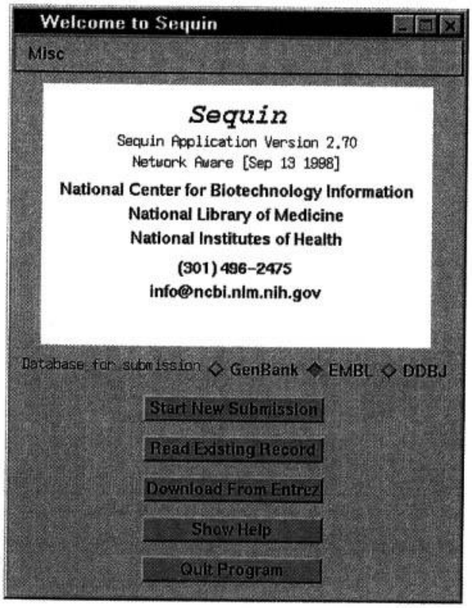
bankit241442

BankIt: General Submission Information

Multiple submissions Information

If you are submitting more than one sequence at this time, please number each sequence and indicate the total number of sequences to be submitted so that we can correctly assign consecutive Accession

Figure 15 The Welcoming page of the Bankit service from the NCBI.



Sequin
Sequin Application Version 2.70
Network Aware [Sep 13 1998]
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
(301) 496-2475
info@ncbi.nlm.nih.gov

Database for submission ◇ GenBank ◆ EMBL ◇ DDBJ

Figure 16 The welcoming screen of the Sequin data submission program from the NCBI.

robust error checking and accommodates very long sequences and complex annotations.

Although Sequin has been implemented by the NCBI, the opening screen allows you to select which database you would like to submit you sequence to be it GenBank, EMBL, or DDBJ. Usually when a sequence is submitted there may be a process whereby the submitter has to be in contact with the annotators of the sequence by telephone to clarify certain details. Therefore it is wise to choose a submission centre in your geographical region if you want to avoid long distance telephone calls. A screen capture of Sequin is shown in *Figure 16*.

Once you have completed the submission depending on which database you have selected at the beginning you will be prompted to send an e-mail to *gb-sub@ncbi.nlm.nih.gov* for NCBI, *datasubs@ebi.ac.uk* for EMBL or *ddbjsub@ddbj.nig.ac.jp* for DDBJ. Sequin runs on Macintosh, PC/Windows, and UNIX computers. The program itself, along with its on-line help documentation, is available by anonymous FTP from the *EBI (UK)* at ftp://ftp.ebi.ac.uk/pub/software/sequin/ or from the *NCBI (USA)* at ftp://ncbi.nlm.nih.gov/sequin/ A useful FAQ to help you if you run into problems during submission can be found at http://www.ebi.ac.uk/~sterk/sqndocs/index.html

4.3 Webin from EBI

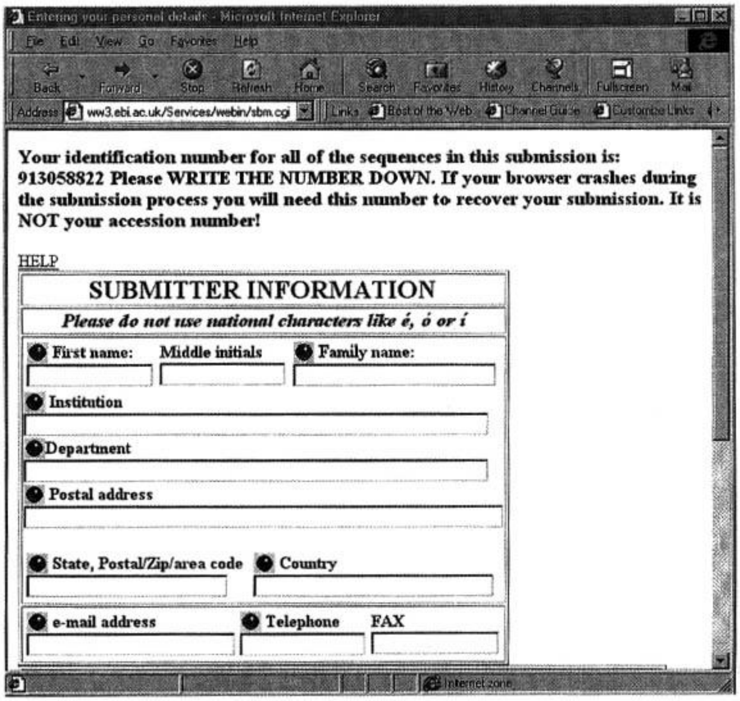
The EBI WWW tool (WebIn) guides the user through a sequence of WWW forms allowing the user to submit sequence data and descriptive information in an interactive and easy way (see *Figure 17*). All the information required to create a database entry will be collected during this process:

(a) Submitter information.
(b) Release date information.
(c) Sequence data, description, and source information.
(d) Reference citation information.
(e) Feature information (e.g. coding regions, regulatory signals etc.).

Data submissions are usually processed within two working days of receipt and the authors are sent notification of their accession number(s). Authors will be asked whether their submitted data can be made available to the public immediately or whether they should be withheld until an author-specified date. Data are never withheld after publication.

Once a database entry has been created from a submission, a copy is sent to the submitter for their reference and for comments or corrections. However, it often happens that the entry is correct when it is created but, with the passage of time, becomes out of date. The authors may make corrections to the sequence itself, or may discover new features of the sequence. Since such findings are often not published, the only way to keep entries correct and up to date is if the authors communicate their new findings to the database. At the EBI this can be done by completing an update form available from the Anonymous FTP, site FTP.EBI.AC.UK in the file: pub/databases/embl/release/update.doc or via the WWW at the URL http://www.ebi.ac.uk/ebi_docs/update.html.

A new service that has been instituted at EBI is scanning for vectors before


Entering your personal details - Microsoft Internet Explorer
File Edit View Go Favorites Help
Back Forward Stop Refresh Home Search Favorites History Channels Fullscreen Mail
Address ww3.ebi.ac.uk/Services/webin/sbm.cgi
Links Best of the Web Channel Guide Customize Links
Internet zone

Your identification number for all of the sequences in this submission is: 913058822 Please WRITE THE NUMBER DOWN. If your browser crashes during the submission process you will need this number to recover your submission. It is NOT your accession number!

HELP

SUBMITTER INFORMATION

Please do not use national characters like é, ó or í

● **First name:** **Middle initials** ● **Family name:**

● **Institution**

● **Department**

● **Postal address**

● **State, Postal/Zip/area code** ● **Country**

● **e-mail address** ● **Telephone** **FAX**

Figure 17 A page from the Webin service of the EBI.

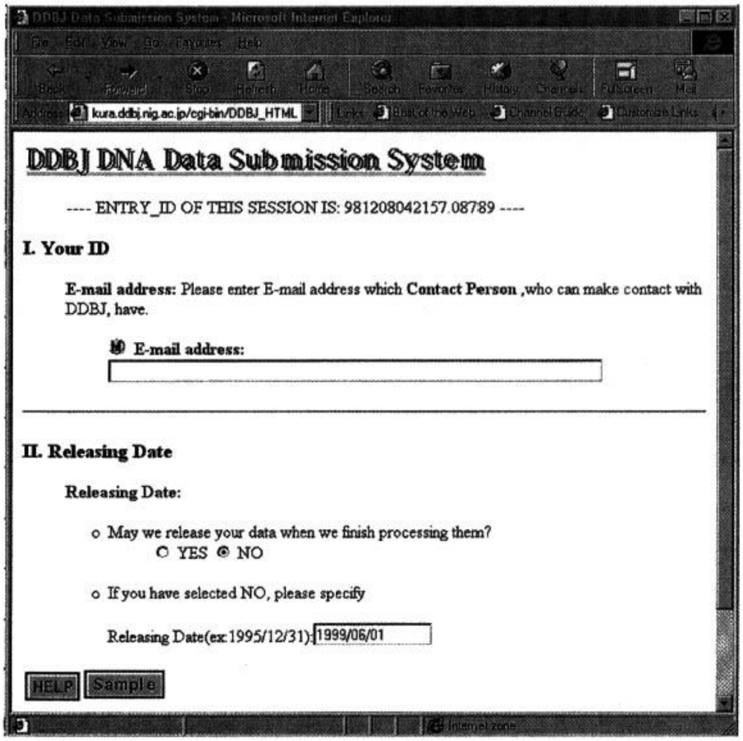
submitting your sequence. You are able to check your sequence data prior to submission for potential vector contamination by running a BLASTN search against EMVEC, a vector database containing information on more than 2000 vectors from the EMBL/GenBank/DDBJ Database SYN(thetic) division. The results will list sequences producing significant alignments and associated information like vector name, score, alignment, etc. The EBI suggests that you remove vector contamination from your sequence data before submitting to the database.

4.4 Sakura from DDJB

SAKURA is a web-based DNA data submission system for DDBJ. The URL for SAKURA is http://sakura.ddbj.nig.ac.jp which can be accessed from the DDBJ Home Page (http://www.ddbj.nig.ac.jp). You can select either the English or Japanese version. However, data input must be done in English only, regardless of language version selected . SAKURA allows you to save your document before completion and submit multiple sequences sequentially (see *Figure 18*).

5 Conclusions

Historically there has been a collaboration between EBI, NCBI, and DDBJ. These three sites are still the only places that have the infrastructure set up to handle


DDBJ Data Submission System - Microsoft Internet Explorer
File Edit View Go Favorites Help
Back Forward Stop Refresh Home Search Favorites History Channels Fullscreen Mail
Address
Links Best of the Web Channel Guide Customize Links
HELP
Sample
Internet zone

kura.ddbj.nig.ac.jp/cgi-bin/DDBJ_HTML

DDBJ DNA Data Submission System

---- ENTRY_ID OF THIS SESSION IS: 981208042157.08789 ----

I. Your ID

E-mail address: Please enter E-mail address which **Contact Person** ,who can make contact with DDBJ, have.

E-mail address:

II. Releasing Date

Releasing Date:

- May we release your data when we finish processing them?
  ○ YES ● NO
- If you have selected NO, please specify

Releasing Date(ex:1995/12/31): 1999/06/01

Figure 18 A page from the SAKURA DNA submission system from DDBJ.

the submission of nucleotide sequences to the databases, be they EMBL or Genbank or DDBJ. For this reason they are also looked upon as the only places where you can do queries and retrieval, or perform homology searches, or multiple sequence alignments. This is no longer true and with the advent of EMBnet, many of the national nodes are able to supply services that are not offered by the major centres. These three major centres have a policy of making all of their databases publicly available, and when distributed network of databases exists in many different parts of the globe then it can only be for the benefit of molecular biologists worldwide.

References

General references to articles about biological services on the internet.

1. Aldhous, P. (1993). Managing the genome data deluge. *Science*, **262**, 502.
2. Altschul, S. *et al.* (NCBI) (1994). Issues in searching molecular sequence databases. *Nature Genet.*, **6**(Feb), 119.
3. Appel, R. D., Sanchez, J.-C., Bairoch, A., Golaz, O., Ravier, F., Pasquali, C., *et al.* (1996). The Swiss-2DPAGE database of two-dimensional polyacrylamide gel electrophoresis, its status in 1995. *Nucleic Acids Res.*, **24**(1), 180.
4. Ashburner, M. and Goodman, N. (1997). Informatics—genome and genetic databases. *Curr. Opin. Genet. Dev.*, **7**, 750.
5. Bairoch, A., Bucher, P., and Hofman, K. (1996). The Prosite Database, its status in 1995. *Nucleic Acids Res.*, **24**(1), 189.
6. Bairoch, A. and Apweiler, R. (1996). The Swiss-Prot Protein sequence data bank and its new supplement Trembl. *Nucleic Acids Res.*, **24**(1), 21.
7. Bairoch, A. (1996). The ENZYME Data Bank in 1995. *Nucleic Acids Res.*, **24**(1), 221.
8. Bairoch, A. (1991). SEQANALREF: a sequence analysis bibliographic reference databank. *Comput. Appl. Biosci.*, **7**(2), 268.
9. Bleasby, A., Griffiths, P., Harper, R., Hines, D., Hoover, K., Kristofferson, D., *et al.* (1992).Electronic communications and the new biology. *Nucleic Acids Res.*, **20**(16), 4127.
10. Coulson, A. (1994). High performance searching of biosequence databases. *Trends Biotechnol.*, **12**, 76.
11. Fuchs, R. (1994). Sequence analysis by electronic mail: a tool for accessing Internet e-mail servers. *Comput. Appl. Biosci.*, **10**(4), 413.
12. Gershon, D. (1997). Bioinformatics in the post-genomic age. Careers and recruitment article. *Nature*, **389**, 417.
13. Gershon, D. (1995). The boom in bioinformatics (employment review). *Nature*, **375**, 262.
14. Harper, R. (EBI). (1995). World Wide Web resources for the biologist. *Trends Genet.*, **11**(6), 223.
15. Holm, L. and Sander, C. (1996). The FSSP database: fold Classification based on structure-structure alignment of proteins. *Nucleic Acids Res.*, **24**(1), 206.
16. Marshall, E. (1996). Hot property: biologists who compute. *Science*, **272**, 1730.
17. O'Donnell, C. (1994). Obtaining software via INTERNET. *Methods Mol. Biol.*, **24**, 345.
18. Peitsch, M. C., Wells, T. N., Stampf, D. R., and Sussman, J. L. (1995). The Swiss-3DImage collection and PDB-Browser on the World-Wide Web. *Trends Biochem. Sci.*, **20**(2), 82.
19. Pietrokovski, S., Henikoff, J. G., and Henikoff, S. (1996). The Blocks database a system for protein classification. *Nucleic Acids Res.*, **24**(1), 197.

20. Roberts, R. J. and Macelis, D. (1996). REBASE—restriction enzymes and methylases. *Nucleic Acids Res.*, **24**(1), 223.
21. Rodriguez-Tomé, P., Stoehr, P., Cameron, G. N., and Flores, T. P. (1996). The European Bioinformatics Institute (EBI). *Nucleic Acids Res.*, **24**(1), 6.
22. Smith, T. F. (1990). The history of the genetic sequence databases. *Genomics*, **6**(4), 701.
23. Stoehr, P. J. and Omond, R. A. (1989). The EMBL Network File Server. *Nucleic Acids Res.*, **17**(16), 6763.
24. Williams, G. W. and Gibbs, G. P. (1990). Automatic updating of the EMBL database via EMBNet. *Comput. Appl. Biosci.*, **6** (2), 122.