New services of the EMBL Data Library

Rainer Fuchs, Peter Stoehr, Peter Rice[1], Roy Omond[1] and Graham Cameron
The EMBL Data Library and [1]EMBL Computer Group, European Molecular Biology Laboratory,
Postfach 10.2209, D-6900 Heidelberg, FRG. E-mail: datalib@embl.bitnet.

ABSTRACT

The existing services of the EMBL Data Library for external users have been improved and extended in several ways. The EMBL File Server has been reorganised, and many new databases and other information relevant to biologists are now accessible via global computer networks. A broad range of software for molecular biology is freely available for different popular computer systems, including the EMBL enhancements to the Wisconsin (GCG) Package. The new Mail-Quicksearch and Mail-FastA services give access to the latest sequence data for database searches by ordinary electronic mail.

INTRODUCTION

The EMBL Data Library (1)—in close collaboration with GenBank (2) and DDBJ (DNA Database of Japan, Mishima, Japan)—collects and distributes nucleotide sequence data worldwide. EMBL also offers a range of computer-based services to the scientific user community, thus providing immediate, reliable and convenient access to these important biological data. This article describes some important improvements of and additions to the existing services.

THE EMBL NETWORK FILE SERVER

In 1987 the EMBL File Server was introduced as a service for external users (3). This fully automatic service, running on the EMBL computer systems, allows the convenient retrieval of database entries and related information using standard electronic mail. Any scientist with access to international computer networks like Bitnet/EARN or Internet can download files, including the most recent EMBL and GenBank database entries which are available as soon as they are created.

Since its introduction the File Server has proven to be very popular with the scientific community. Currently about 1000 requests are answered each month. Comprehensive help and directory documentation is provided which gives new users an easy introduction to the system. To get introductory information send an electronic mail message to

NETSERV@EMBL.BITNET

containing the single line:

HELP

The service has now been extended in two directions: the addition of new databases and the distribution of free software.

New databases

In order to make new databases and related information available the directory structure of the EMBL File Server has been reorganised. Table I summarizes the currently existing directories and the information contained in them. In addition to the EMBL, SwissProt (4) and Brookhaven (5) databases, a number of other relevant data collections are now provided as well, including the Prosite pattern database (6), the new ENZYME database (6), the REBASE restriction enzyme database (7) and the E. coli database (ECD) (8). The complete contents of the quarterly EMBL and SwissProt release tapes is now also available for direct electronic access.

Molecular biological software

At the end of 1989 we started to redistribute free molecular biological software via the EMBL File Server (11). The plan is to establish and maintain a repository of high-quality, but freely available software in this field. At the moment, MS-DOS, Apple Macintosh and VAX/VMS computer systems are supported; UNIX software will be included soon. Software authors are encouraged to directly submit their programs to EMBL for redistribution, thus making them available to a large user community.

This new service has already turned out to be very successful. Within just 6 months of operation about 60 different programs have been submitted to EMBL, and more than 2300 requests for these programs have already been answered.

To allow transmission of binary files over electronic mail networks all programs are encoded and converted to plain ASCII format. Reconversion by the recipient is a simple and straightforward process, with the necessary software being provided on the File Server.

Detailed information about the software archive can be obtained by sending a mail message to

NETSERV@EMBL.BITNET

containing the line

HELP SOFTWARE

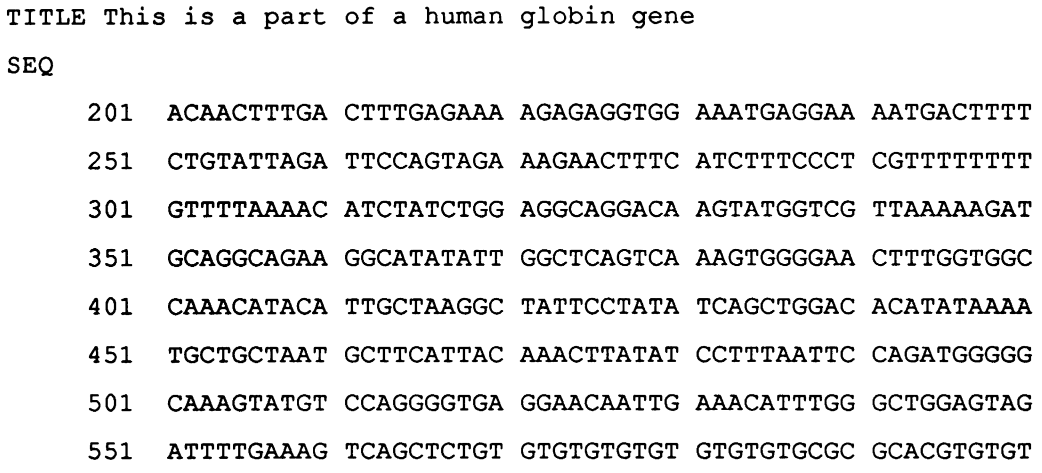
A key position within the software available for VAX/VMS computers is taken by the EMBL enhancements to the Wisconsin (GCG) Package (12), which is one of the most widely used suites of sequence analysis programs. The EMBL enhancements are a collection of programs, written by Peter Rice, which are based on the procedure library of the Wisconsin Package. This collection consists of a variety of programs ranging from the management of large-scale sequencing projects (13) to fast

Table I. The EMBL File Server directory structure. The publicly accessible directories and their contents are shown here. Help files and directory listings are available by sending the File Server commands 'HELP' and 'DIR directory_name' to NETSERV@EMBL.BITNET.

File Server directory	Contents
ECD	E. coli database (8)
ENZYME	Enzyme database (6)
EPD	Eukaryotic promotor database (9)
DOC	General documentation and information for molecular biologists
LIMB	Listing of molecular biological databases (10)
NUC	EMBL, GenBank and DDBJ sequence data plus index files(1,2)
PROSITE	Prosite protein pattern database (6)
PROTEIN	SwissProt protein sequence data plus index files (4)
PROTEINDATA	Brookhaven structural data (5)
REBASE	Restriction enzyme database (7)
REFLIST	Sequence analysis reference list (6)
SOFTWARE	Free molecular biological software (11)

Table II. The EMBL enhancements to the Wisconsin Package. This table lists the programs provided in the three parts of this package.

EMBL enhancements to the Wisconsin Package	
GCGEMBL	Programs for sequence alignment, shotgun sequencing and signal cleavage site prediction.
GCGDBASE	Programs for creating EMBLAll, EMBLNew, GBNew, GBOnly, GBEMBL, SwissPIR and PIROnly databases in GCG format.
GCGQUICK	Programs for very fast database searches.



```
TITLE This is a part of a human globin gene

SEQ

      201  ACAACTTTGA CTTTGAGAAA AGAGAGGTGG AAATGAGGAA AATGACTTTT

      251  CTGTATTAGA TTCCAGTAGA AAGAACTTTC ATCTTTCCCT CGTTTTTTTT

      301  GTTTTAAAAC ATCTATCTGG AGGCAGGACA AGTATGGTCG TTAAAAAGAT

      351  GCAGGCAGAA GGCATATATT GGCTCAGTCA AAGTGGGGAA CTTTGGTGGC

      401  CAAACATACA TTGCTAAGGC TATTCCTATA TCAGCTGGAC ACATATAAAA

      451  TGCTGCTAAT GCTTCATTAC AAACTTATAT CCTTTAATTC CAGATGGGGG

      501  CAAAGTATGT CCAGGGGTGA GGAACAATTG AAACATTTGG GCTGGAGTAG

      551  ATTTTGAAAG TCAGCTCTGT GTGTGTGTGT GTGTGTGCGC GCACGTGTGT
```

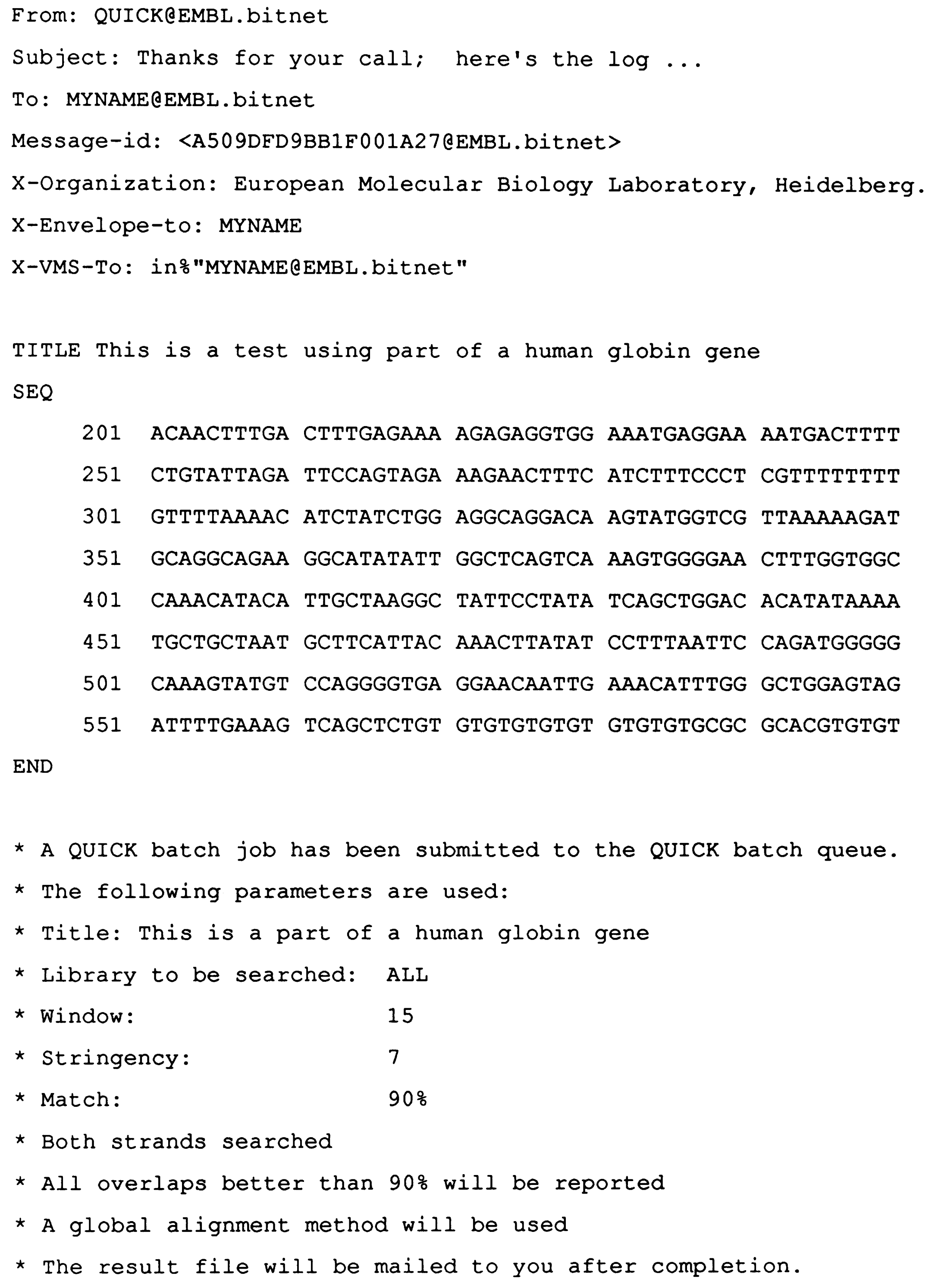
Figure 1: An example Mail-Quicksearch input file. Sending a mail message like this to QUICK@EMBL.BITNET will initiate a sequence comparison against the complete EMBL nucleotide sequence database, using J. Devereux's very fast algorithm (14). Line numbering as shown in this example is optional. When the results are returned to the user Mail-Quicksearch will use the TITLE line as the Subject line of its mail message. TITLE is the only command used in this example. Default values are substituted for all other parameters.

database searching (Table 2) and is suitable for use at any site licensed to use the Wisconsin Package.

DATABASE SEARCHING

The most common application of sequence databases in biological research is probably the comparison of newly determined nucleotide or protein sequences with entire databases to detect similar and related sequences. Two most recent additions to EMBL's services, Mail-Quicksearch and Mail-FastA allow for searching a comprehensive range of databases remotely on the EMBL computer systems. Because new EMBL and GenBank entries are added daily, the most complete and up-to-date data collections are thus available for sequence comparison. The researcher may choose between two different algorithms and select the one best suited to answer his specific question.

Requests are sent to EMBL as ordinary mail files. The receipt of each request is immediately acknowledged, and if any problems occurred while processing it they are reported back to the sender. Well formed requests are entered into a batch queue and database searches are performed in the order of receipt. After completion of the sequence comparison the results are



```
From: QUICK@EMBL.bitnet
Subject: Thanks for your call;  here's the log ...
To: MYNAME@EMBL.bitnet
Message-id: <A509DFD9BB1F001A27@EMBL.bitnet>
X-Organization: European Molecular Biology Laboratory, Heidelberg.
X-Envelope-to: MYNAME
X-VMS-To: in%"MYNAME@EMBL.bitnet"

TITLE This is a test using part of a human globin gene
SEQ
     201  ACAACTTTGA CTTTGAGAAA AGAGAGGTGG AAATGAGGAA AATGACTTTT
     251  CTGTATTAGA TTCCAGTAGA AAGAACTTTC ATCTTTCCCT CGTTTTTTTT
     301  GTTTTAAAAC ATCTATCTGG AGGCAGGACA AGTATGGTCG TTAAAAAGAT
     351  GCAGGCAGAA GGCATATATT GGCTCAGTCA AAGTGGGGAA CTTTGGTGGC
     401  CAAACATACA TTGCTAAGGC TATTCCTATA TCAGCTGGAC ACATATAAAA
     451  TGCTGCTAAT GCTTCATTAC AAACTTATAT CCTTTAATTC CAGATGGGGG
     501  CAAAGTATGT CCAGGGGTGA GGAACAATTG AAACATTTGG GCTGGAGTAG
     551  ATTTTGAAAG TCAGCTCTGT GTGTGTGTGT GTGTGTGCGC GCACGTGTGT
END

* A QUICK batch job has been submitted to the QUICK batch queue.
* The following parameters are used:
* Title: This is a part of a human globin gene
* Library to be searched:  ALL
* Window:                  15
* Stringency:              7
* Match:                   90%
* Both strands searched
* All overlaps better than 90% will be reported
* A global alignment method will be used
* The result file will be mailed to you after completion.
```

Figure 2: Acknowledgement of a Mail-Quicksearch request. If the request was processed without errors the parameter settings as used in the database search are indicated as shown here. Otherwise detailed error descriptions will be returned.

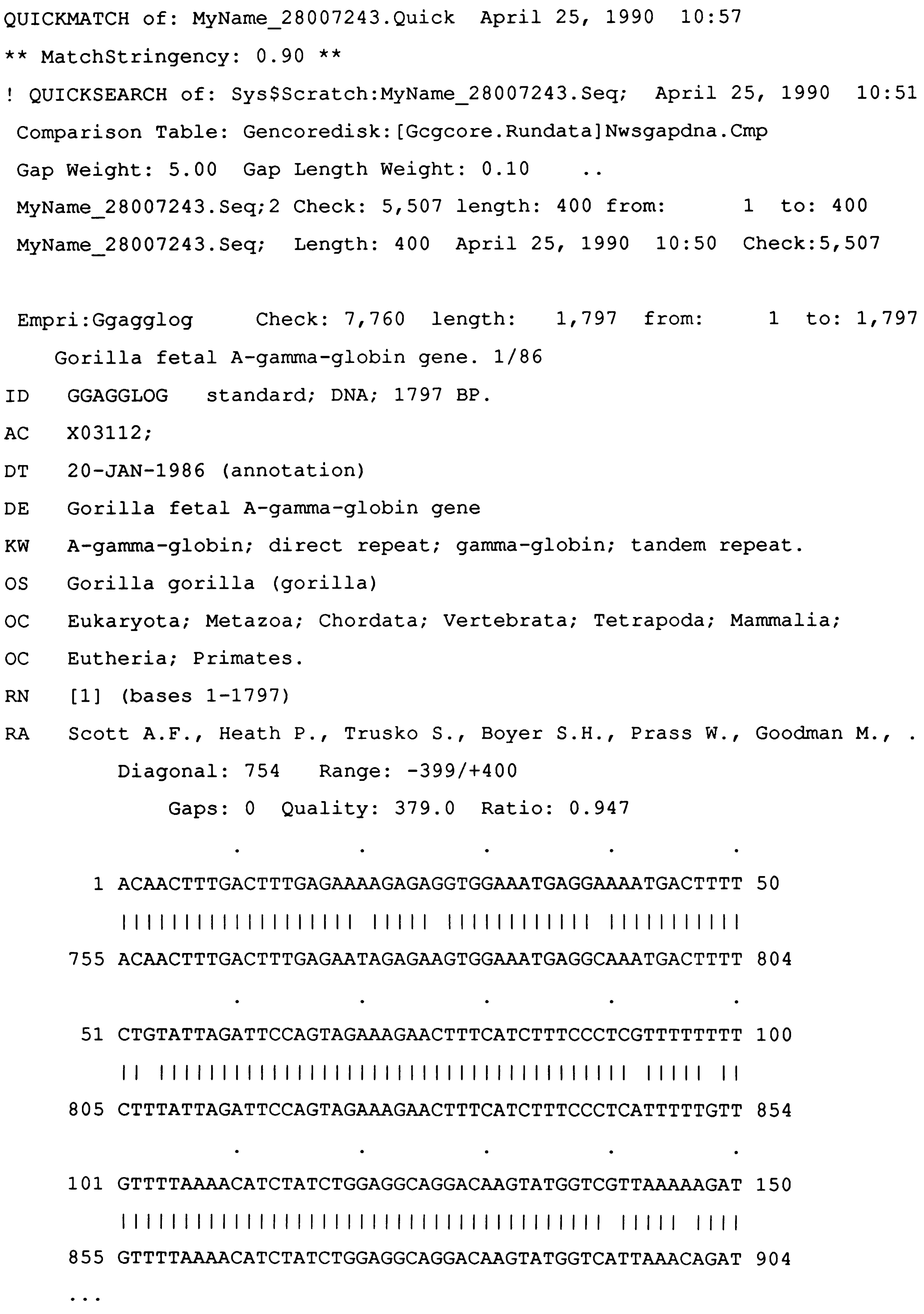
automatically returned. Depending on the network connections and the number of requests waiting in the batch queue results may be obtained in a few minutes.

Both services are of course provided free of charge.

Mail-Quicksearch

This new service is based on the NewQuicksearch and Quickmatch programs as provided in the GCGQUICK part of the EMBL enhancements package. These programs are improvements of the original QuickSearch programs (14) in the GCG Package which make it possible to very quickly search the DNA databases for sequences similar to a query sequence. Currently, protein databases are not available for searching.

Mail-Quicksearch answers the question: is my sequence or a (very) similar one already in the database? It will find any closely related sequences extremely fast, but it will not detect more distant



```
QUICKMATCH of: MyName_28007243.Quick  April 25, 1990  10:57
** MatchStringency: 0.90 **
! QUICKSEARCH of: Sys$Scratch:MyName_28007243.Seq;  April 25, 1990  10:51
 Comparison Table: Gencoredisk:[Gcgcore.Rundata]Nwsgapdna.Cmp
 Gap Weight: 5.00  Gap Length Weight: 0.10    ..
 MyName_28007243.Seq;2 Check: 5,507 length: 400 from:      1  to: 400
 MyName_28007243.Seq;  Length: 400  April 25, 1990  10:50  Check:5,507

 Empri:Ggagglog     Check: 7,760  length:   1,797  from:     1  to: 1,797
    Gorilla fetal A-gamma-globin gene. 1/86
ID   GGAGGLOG   standard; DNA; 1797 BP.
AC   X03112;
DT   20-JAN-1986 (annotation)
DE   Gorilla fetal A-gamma-globin gene
KW   A-gamma-globin; direct repeat; gamma-globin; tandem repeat.
OS   Gorilla gorilla (gorilla)
OC   Eukaryota; Metazoa; Chordata; Vertebrata; Tetrapoda; Mammalia;
OC   Eutheria; Primates.
RN   [1] (bases 1-1797)
RA   Scott A.F., Heath P., Trusko S., Boyer S.H., Prass W., Goodman M., .
         Diagonal: 754   Range: -399/+400
             Gaps: 0  Quality: 379.0  Ratio: 0.947

                 .         .         .         .         .
       1 ACAACTTTGACTTTGAGAAAAGAGAGGTGGAAATGAGGAAAATGACTTTT 50
         ||||||||||||||||||| ||||| |||||||||||| |||||||||||
     755 ACAACTTTGACTTTGAGAATAGAGAAGTGGAAATGAGGCAAATGACTTTT 804

                 .         .         .         .         .
      51 CTGTATTAGATTCCAGTAGAAAGAACTTTCATCTTTCCCTCGTTTTTTTT 100
         || |||||||||||||||||||||||||||||||||||||| ||||| ||
     805 CTTTATTAGATTCCAGTAGAAAGAACTTTCATCTTTCCCTCATTTTTGTT 854

                 .         .         .         .         .
     101 GTTTTAAAACATCTATCTGGAGGCAGGACAAGTATGGTCGTTAAAAAGAT 150
         ||||||||||||||||||||||||||||||||||||||| ||||| ||||
     855 GTTTTAAAACATCTATCTGGAGGCAGGACAAGTATGGTCATTAAACAGAT 904
     ...
```

Figure 3: An example Mail-Quicksearch output file. The first lines include information about the parameters used in the database search. If any similar sequences are found in the database then the first ten lines of each corresponding database entry are shown plus the sequence alignments. In this example the alignment is only shown partially. The entry information can be used to retrieve the database entry from the EMBL File Server.

similarities. A typical application for Mail-Quicksearch is the comparison of newly determined sequences of unknown DNA to the database. In addition to all entries contained in the latest EMBL release, all newly created EMBL and GenBank entries are accessible as well. The hash tables which speed up the searches are rebuilt at EMBL each night.

The Mail-Quicksearch service is used by sending a properly formatted mail file to the network address QUICK@EMBL.BITNET, including sequence data and optionally a few commands. The required syntax is simple and intuitive. No special sequence format is required and in most cases sending merely the sequence is sufficient, since the search parameters will automatically be set to appropriate default values if not specified. Occasionally it may be necessary to directly set these parameters. For this purpose a small set of commands is provided. The help file that is transferred in response to the command HELP gives a detailed description of each of the commands available, including a sample input and output. A typical Mail-Quicksearch command file is shown in Fig. 1. The receipt of each request is immediately acknowledged (Fig. 2) and any syntax errors are reported. The output file of a Mail-Quicksearch run which is returned to the scientist contains all database sequences with more than a certain amount of similarity to the query sequence (normally 90%) and includes the sequence alignments of the database and query sequences (Fig. 3).

Mail-FastA

This service is based on the fast and sensitive database search algorithm implemented in the FastA program by Pearson and Lipman (15), which is probably the most widely used sequence comparison program today.

Mail-FastA answers the question: are there any sequences in the database which show some similarity to my query sequence? In contrast to Mail-Quicksearch it will also detect distantly related sequences, but its cpu-time demands and therefore the response time are much greater. The twenty databases available for searching include EMBL, GenBank, SwissProt, PIR, and subsets of these databases. Again, new EMBL and GenBank entries are added to the data set on a daily basis.

Using Mail-FastA is very similar to using Mail-Quicksearch. A standard mail message containing a few commands and the sequence data should be sent to FASTA@EMBL.BITNET. Like Mail-Quicksearch, no special sequence format is required and default values will be used whenever possible. The only parameter which has to be specified is the database to be searched. The remaining commands may be used for adjusting the sensitivity of a FastA search and the output format. Again, all possible commands are thoroughly explained in the help documentation which is obtained by sending the command HELP to FASTA@EMBL.BITNET.

The output of a Mail-FastA job contains a list of the database entries most similar to the query sequence and optionally the sequence alignments. Following a Mail-FastA or a Mail-Quicksearch run the entry names reported in the result files may subsequently be used to retrieve the corresponding database entries from the EMBL File Server for further analysis.

FURTHER PLANS

Although very effective and convenient, the functionality of services based on electronic mail communication is obviously limited. EMBL is therefore also investigating alternative means for data distribution and data access. TCP/IP networks, to which EMBL is connected, are increasingly popular in Europe, and the possibility of more sophisticated file transfer using them is being explored.

REFERENCES

1. Hamm, G. and Cameron, G. (1986) *Nucl. Acids Res.* ,**14**, 5–10.
2. Burks, C., Fickett, J.W., Goad, W.B., Kanehisha, M., Lewitter, F.I., Rindone, W.P., Swindell, C.D., Tung, C.S. and Bilofsky, H.S. (1985) *CABIOS* ,**1**, 225–233.
3. Stoehr, P. and Omond, R. (1989) *Nucl. Acids Res.*, **17**, 6763–6764.
4. Bairoch, A. (1990) University of Geneva, Geneva, and EMBL Data Library, EMBL, Heidelberg.
5. Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) *J. Mol. Biol.*, **112**, 535–542.
6. Bairoch,A. (1990) University of Geneva, Geneva.
7. Roberts,R.J. (1985) *Nucl. Acids Res.*, **13**, r165-r200.
8. Kroeger,M. (1989) *Nucl. Acids Res.*, **17**, r283-r309.
9. Bucher,P. and Trifonov,E.N. (1986) *Nucl. Acids Res.*, **14**, 10009–10026.
10. Lawton, J.R., Martinez,F.A. and Burks,C. (1989) *Nucl. Acids Res.*, **17**, 5885–5899.
11. Fuchs,R. (1990) *CABIOS*, **6**, 120–121.
12. Devereux,J., Haeberli,P. and Smithies,O. (1984) *Nucl. Acids Res.* , **12**, 387–395.
13. Edwards,A.; Voss,H., Rice,P., Civitello,A., Stegemann,J., Schwager,C., Zimmermann,J., Erfle,H., Caskey,C.T. and Ansorge,W. (1990) *Genomics*, **6**, 593–608.
14. Devereux,J. (1988) Ph. D. thesis.
15. Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.