CABIOS

Vol. 12 no. 2 1996
Pages 129–133

FastAlert–an automatic search system to alert about new entries in biological sequence databanks

F.Eggenberger[1], N.Redaschi and R.Doelz

Abstract

This paper describes a new tool enabling awareness of new sequence databank entries of interest. The FastAlert system relieves the researcher from the burden of repeating FASTA searches in order to keep up with the rapidly growing amount of information found in biological sequence databanks. The query sequence can be submitted from any computer connected to the Internet. Upon registration, the databank, including the updates, is scanned at periodic intervals with the sequence provided. The results, so-called FastAlert reports, are delivered via electronic mail. The reports contain the FASTA best-scores list and the similarity statistics for each entry listed.

Introduction

Investigation of a newly determined sequence usually starts by comparing it with all sequences available in the appropriate database, such as the EMBL nucleotide sequence library (Higgins *et al.*, 1992; Rice *et al.*, 1993) or the SWISS-PROT (Bairoch and Boeckmann, 1993) protein sequence databank. However, molecular sequence databanks are frequently updated to include newly submitted sequences and entries which have experienced changes in their contents. Therefore the findings of a once accomplished databank scan will be valid only temporarily. Systematic methods are required to cope with the steadily increasing volume of databank updates. The total volume of updates of the EMBL and GenBank nucleotide sequence databases typically range between 20 and 100 Mbytes. It is thus rather impractical to get the latest entries of interest by manually searching the updates via conventional network tools, such as Usenet News, which may be unavailable to non-academic sites. Automatic services have recently been made available to the Internet community which allow people to search the latest entries on a remote computer conveniently. ExPASy's Swiss-Shop service (Peitsch, 1995) has especially been designed to meet the requirements of the users of the protein sequence databank SWISS-PROT. Swiss-Shop allows any user to compare the new sequences entered in SWISS-PROT with user-defined criteria, such as words in the sequence description or sequence pattern. The Belgian EMBnet node's (BEN) CSA and CSSA services (Alard, 1995) provide similar capabilities to search the EMBL and GenPept databanks. Both Swiss-Shop and BEN's awareness tools are accessible via WWW and will search only the databank updates.

In this paper we describe a new service, the FastAlert system, which allows to search fully updated protein or DNA sequence databanks using the FASTA program (Lipman and Pearson, 1985; Pearson and Lipman, 1988). We have implemented this system on a client/server system on the Internet (see below). The sequence submission can be done from any computer connected to the Internet by using specific requester software. Unlike Swiss-Shop and BEN's awareness tools, FastAlert does not use the WWW interface but is built on top of the HASSLE v5 network protocol (Doelz, 1994; Doelz *et al.*, 1994; Doelz, 1995; Redaschi *et al.*, 1995) which handles the resource discovery in a fully transparent way. Thus, the user's query sequence is registered automatically at the nearest server which is able to handle the request. Upon registration, the sequence is searched at periodic intervals against the appropriate set of databases and the results, so-called FastAlert reports, are sent to the user via electronic mail.

System and methods

The FastAlert system follows the client/server paradigm on the basis of IP network as implemented using the HASSLE protocol. The system comprises the FastAlert requester on the user side and the FastAlert server and mailer modules on the provider side (Figure 1). The FastAlert requester is written in ANSI C using the NCBI Vibrant windowing system (Ostell, 1993) and the ICLUI class library (Leong *et al.*, 1995) and runs on MS-Windows 3.1, OS/2 Warp, Motif, and MacOS. A text-based version is available for UNIX, VMS and OS/2 platforms (Figure 2). The server and mailer modules are written in ANSI C and the 'csh' scripting language. Programs launched on the provider side include READSEQ (D.G. Gilbert, Indiana University) for sequence

BioComputing, Basel University, Biozentrum, Klingelbergstrasse 70, CH-4056 Basel, Switzerland. Email: embnet@comp.bioz.unibas.ch

[1]To whom correspondance should be addressed at present address: Union Bank of Switzerland, SYNE/SYKA-ENF, PO Box, CH-8021 Zürich, Switzerland. Email: florian.eggenberger@ubs.ch

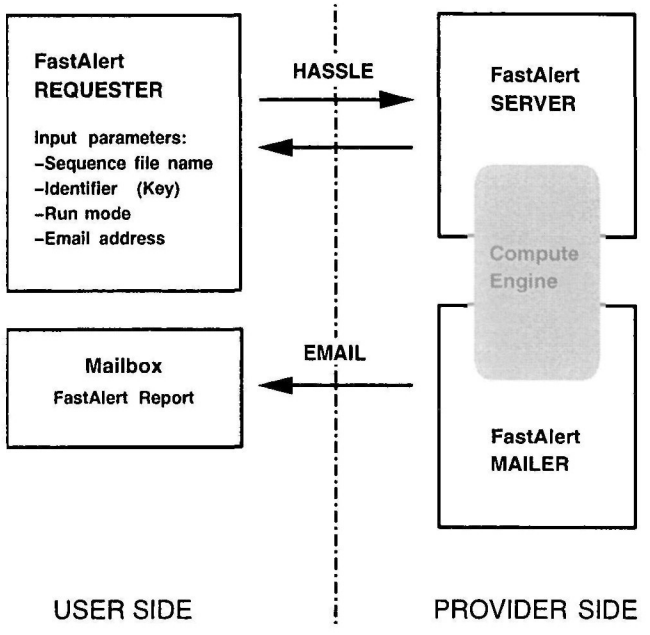
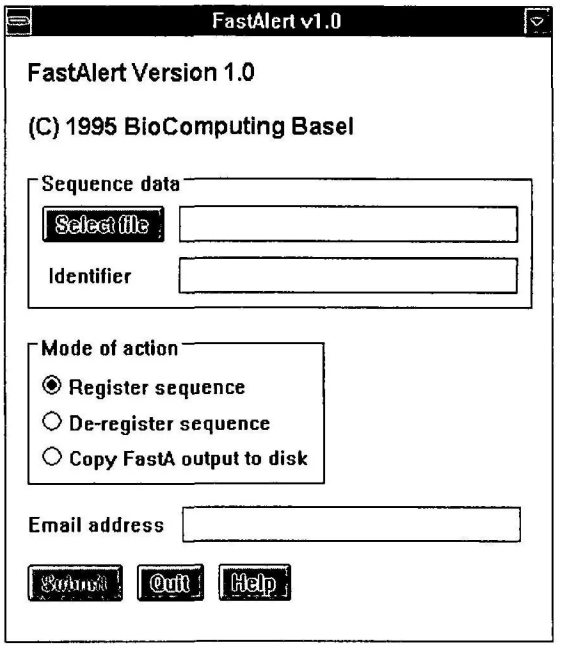


Fig. 1. Flowchart of data processing. Network communication is established by using either electronic mail or the HASSLE (hierarchical access system for sequence libraries in Europe) protocol

format conversion, REFORMAT (Genetics Computer Group, 1994) for sequence reformatting, FASTA (Genetics Computer Group, 1994) for similarity searches, and PRDF (W.R. Pearson and D.Lipman, University of Virginia) for estimating the probability of FASTA's similarity scores.

Requester module

FastAlert currently supports three operational modes: (i) registration of a sequence, (ii) de-registration of a sequence, and (iii) transmission of the original FASTA output to the user's local disk. The FastAlert requester provides a graphical interface to enter the input-parameters needed to handle the query. The user is asked to supply detailed parameters as shown in Figure 1. If the format of all parameters is correct, the requester launches a HASSLE communication to the FastAlert server. Upon successful negotiation (which includes authentication and security checks) the query is transmitted to the provider. As defined in the HASSLE communication protocol, all data are transparently compressed and encrypted. On the provider side the integrity of the transmitted data is checked and acknowledged. While the FastAlert server performs several data and system checks (using the programs listed above), the communication terminates and the requester starts listening for the reply from the server. This asynchronous communication model, which is a common feature of HASSLE applications, has the advantage that the line has not to be kept alive while the




```
$ fastalert

FastAlert does weekly FastA scans of DNA or protein
databanks with a query sequence. The results, so-called
FastAlert reports, are sent to the email address provided.
The reports contain the FastA best-scores list and the
probability estimates for the optimized FastA similarity
scores

 FastAlert with what query sequence ? clon17.seq

 Please provide a unique sequence identifier· fe13sept95

 Mode of action:

 1) Register sequence
 2) De-register sequence
 3) Copy FastA output to disk

 Please choose one (* 1 *). 1

 Please provide your email address:
eggenber@comp.bioz.unibas.ch

 Submit request ? (* y *) y

--- HASSLE version D5-0-1 (95/07/25) (C) BioComputing
Basel 1995
NOTE: BABY.URZ.UNIBAS.CH is used as provider
NOTE: Listening for result on port 15017
NOTE: Wrote h4054.job
NOTE: Wrote clon17.sta

FastAlert v1 0: Job completed successfully, sequence
'clon17.seq' registered. The results will be delivered by
email to 'eggenber@comp.bioz unibas.ch'

$
```

Fig. 2. Operation of the FastAlert requester on the user side A graphical user interface for MS-Windows and a text-driven version for the VMS operating system are shown.

job on the provider side is running. After all checks are completed the server launches a HASSLE communication to the requester which is still listening on the previously negotiated port. This is to transmit the success status of the user's request which is subsequently analysed and displayed. Finally, in order to return the search result by electronic mail later, as described below, the requester terminates while the FastAlert server will still process the query.

Server module

The data transmitted by the FastAlert requester are stored until de-registration in a request-specific directory, which is named reflecting the user-supplied parameters. Upon a registration request, the server first checks the format of the transmitted data and runs a system self-test. The user-supplied sequence file is converted to GCG format by the programs READSEQ and REFORMAT and then scanned against itself using GCG's FASTA. Upon successful completion, a welcome message is sent to the user. The success status of the routines executed so far is then transmitted to the FastAlert requester as described above. If all tests are passed successfully, the actual registration procedure is launched. First the sequence is searched against the appropriate set of sequence databanks using GCG's FASTA. The output of the FASTA run is then submitted to the routine SCORECMP which executes the sequence shuffling program PRDF for each databank entry present in FASTA's best-scores list. Finally, a so-called FastAlert report is produced and sent by electronic mail to the user. The report contains each entry of the FASTA best-scores list followed by the probability estimate for the optimized similarity score as calculated by PRDF.

Mailer module

The FastAlert mailer is run at regular intervals to alert users about changes in FASTA's output with respect to the previous run. The routine first charges the user's HASSLE account using the HASSLE v5 API and then launches GCG's FASTA to search the query sequence against the appropriate set of databanks. Upon successful completion of the FASTA search, the FastAlert mailer executes the routine SCORECMP to compare the current and the previous FASTA run, to calculate the similarity statistics and to produce the FastAlert report, as described above. Unlike the first report produced by the server module, the reports produced by FastAlert mailer contain only those entries of the FASTA best-scores list which did not appear in the previous run. Finally, the report is sent by electronic mail to the user. The authentication mechanism provided by HASSLE prevents that anonymous users can exploit the service provider. If an account has been overdrawn, the sequence is de-registered automatically and the user is notified by electronic mail.

Discussion

The volume of the major nucleotide and protein sequence databases needed for a fully functional search and retrieval service currently amounts to approximately 5 Gbyte of disk space and is growing continuously. Given this huge amount of data, many research sites can no longer make the databanks locally available without investing significantly in additional equipment and manpower. Searching the current DNA or protein databanks with the usually available PC equipment may not only take ages but often requires access to multiple CD-ROM drives. With the ongoing spread of network connectivity, remote searching over the Internet provides a possible solution to this problem. Databank access over the network has been implemented using Email approaches or dedicated systems, such as the Hierarchical Access System for Sequence Libraries in Europe (HASSLE). Email search servers are widely distributed but require that the user formulates the query in a server-specific format. By contrast, the HASSLE system does not require any additional learning efforts from the user. Most local search tools can be adapted to support remote searching via HASSLE. The HASSLE system is fully transparent. Thus, when switching from a local to a remote searching approach, the user can formulate the query the same way as before. Moreover, the remote site offering the requested service does not have to be known, HASSLE automatically locates the nearest server which is able to handle the request.

To stay aware of the latest sequences of interest, today's molecular biologists have to redo a databank search at periodic intervals. Automatic sequence awareness services which relieve the user from the burden of doing searches manually have not been available until recently. Principally, two approaches are possible: either only the updates or the fully updated databank sets are scanned. Both methods may have their advantages and drawbacks. Considering the resources needed to process several hundred thousand sequences, searching only the updates might be the more appropriate method, at least from the provider's point of view. Using this approach, however, the user still has to search the full databank release somewhere else, otherwise the results obtained from the awareness service are rather worthless. Searching the complete, updated databanks, on the other hand, provides access to the entire databank. The user, therefore, does not need to run any other database search and may even be

tempted to utilize the awareness service for unique databank searches.

The methods to estimate the statistical significance of the similarity of two sequences are either based on theoretical models or permutations of the observed data. FastAlert makes use of the latter approach. The probability estimates of the similarity scores produced by FASTA are calculated using the PRDF program, which is an improved version of RDF (Pearson and Lipman, 1988). Basically, PRDF compares two sequences by calculating initial and optimal similarity scores and then repeatedly shuffles the second sequence and calculates the similarity scores again. Extreme value distributions are fit to the distributions of the scores (Altschul *et al.*, 1994). This allows to estimate the probability that each of the unshuffled sequence scores would be obtained by chance.A well-known example for a method that is based on a theoretical model is the BLAST algorithm (Altschul *et al.*, 1990). BLAST uses a direct approximation of the similarity statistics assuming that the probability of finding each amino-acid at each position in a protein is simply proportional to the protein's composition in the database as a whole. This assumption might not always be valid for small databank sets, such as a few hundred (or even less) sequence updates usually searched by Swiss-Shop or BEN's CSSA awareness tools.

Due to its performance and sensitivity, the BLAST program is currently one of the most popular sequence comparison tools. In contrast to FASTA 1.x (as distributed with the Wisconsin package), the BLAST output provides probability estimates for the databank entries found. The most recently released version 2.x of the FASTA program now provides similarity statistics as well but it takes still significantly more time to scan a sequence databank with FASTA than with BLAST. This is because FASTA allows for gaps while BLAST does not. To search a DNA databank with a DNA query sequence, BLAST ignores segment pairs which do not contain at least 12 matches. As a consequence, FASTA should theoretically be much more sensitive for DNA databases searches than BLAST.

FastAlert in its current implementation uses GCG's FASTA program and the sequence shuffling program PRDF v. 1.8, as described above. The design of the FastAlert system is, however, open to allow the incorporation of any other combination of search engine and similarity statistics program. Hence, it is possible that a future version of FastAlert will integrate with BLAST or FASTA 2.x instead.

The FastAlert service is currently available on two DEC/AXPs at BioComputing Basel. These servers are primarily intended to meet the requirements of Swiss Universities and EMBnet members (Doelz, 1993). For all other users, restrictions apply which reflect the limited amount of resources available to run the service in Basel. Non-commercial users have access on a temporary basis. In spite of these restrictions, the system has been received very well by the international community. Two thirds of a total of 120 sequences which have been registered in the last two months, were submitted by research institutes outside Switzerland (Figure 3).

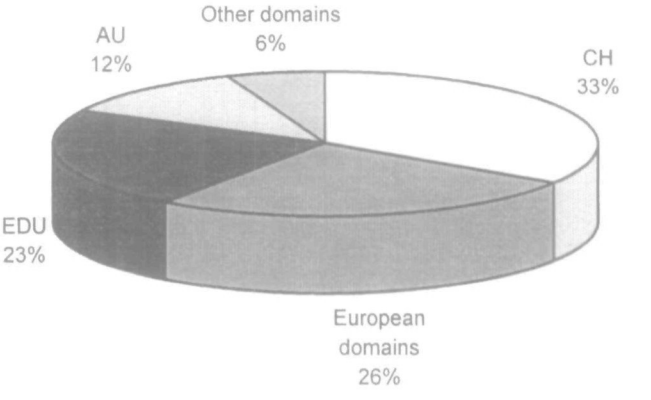


Fig. 3. FastAlert usage at BioComputing, Basel, July—September 1995. Percentages of service requests from hosts of various Internet domains.

Availability

The FastAlert requester suited to run on the most popular operating systems in use today can be downloaded by anonymous FTP from *nic.switch.ch* in the directory *mirror/embnet-ch/bioftp-sw/fastalert*. Text-based versions and the FastAlert provider suite can be obtained on request.

Acknowledgements

Thanks to Bill Pearson and Paul Baumgartner for comments on an earlier version of this system. IBM Switzerland provided help with the program development on OS/2. Financial supported was provided by the University of Basel, the Swiss National Science Foundation and the Bundesamt für Bildung and Wissenschaft.

References

Alard,P. (1995) CSA and CSSA, BEN's new sequence arrival warning tools. *Embnet.news* 2, 1–2. Available via anonymous FTP from: ftp.no.embnet.org in the directory pub/embnet.news/.

Altschul,S.F., Gish,W. Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic Local Alignment Search Tool. *J. Mol. Biol.*, **215**, 403–410.

Altschul,S.F., Boguski,M.S. Gish,W. and Wooton,J.C. (1994) Issues in searching molecular sequence databases. *Nature Genetics*, **6**, 119–129.

Bairoch,A. and Boeckmann,B. (1993) The SWISS-PROT protein sequence data bank, recent developments, *Nucleic Acids Res*., **21**, 3093–3096.

Doelz,R. (1993) The EMBnet Project. *Computer Networks and ISDN Systems*, **25**, 464–468.

Doelz,R. (1994) Hierarchical access system for sequence libraries in Europe (HASSLE): a tool to access sequence databases remotely. *Comput. Applic. Biosci.*, **10**, 31–34.

Doelz,R. (1995) Biology and networks: new tools to manage distributed resources. *Computer Networks and ISDN Systems*, **26** (Suppl. 4), 157–162.

Doelz,R., Eggenberger, F. and Wadley, C. (1994) Biocomputing on a server network. *Embnet.news* 1, 6–8. Available via anonymous FTP from: ftp.no.embnet.org in the directory pub/embnet.news/.

Genetics Computer Group, Inc. (1994) Program Manual for the Wisconsin Package, Version 8.

Higgins,D.G., Fuchs,R., Stoehr,P.J. and Cameron,G.N. (1992) The EMBL Data Library, *Nucleic Acids Res.*, **20**, 2071–2074.

Leong,K., Law,W., Love,R., Tsuji,H. and Oson,B. (1995) *OS/2 Class Library: Power GUI Programming with C Set++*. Van Nostrand Reinhold, New York.

Lipman,D.J. and Pearson,W.R. (1985) Rapid and sensitive protein similarity searches. *Science*, **227**, 1435–1441.

Ostell,J. (1993) NCBI Software Development ToolKit. Available by anonymous FTP from the National Center for Biotechnology Information at ncbi.nlm.nih.gov.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Peitsch,M. (1995) The Swiss-Shop BioComputing Server. Accessible via WWW at the URL http://expasy.hcuge.ch/swisshop/SwissShopReq.html

Redaschi,N., Doelz,R. and Eggenberger,F. (1995) *Advanced Computer Network Communication: Hierarchical Access System for Sequence Libraries in Europe (HASSLE v5)*. Dr. U. Dolz Verlag, Basel.

Rice,C.M., Fuchs,R., Higgins,D.G., Stoehr,P.J. and Cameron,G.N. (1993) The EMBL data library, *Nucleic Acids Res.*, **21**, 2967–1971.

Received on October 30, 1995; revised and accepted on February 14, 1996