CABIOS APPLICATIONS NOTE

Vol. 13 no. 6 1997
Pages 619-620

The Sequence Alerting Server—a new WEB server

Hedvig Hegyi, Jen-Mai Lai and Peer Bork[1]

EMBL, Meyerhofstraße 1, 69012 Heidelberg and Max-Delbruck-Centre for Molecular Medicine, Berlin-Buch, Germany

Received on May 9, 1997; revised and accepted on June 6, 1997

Abstract

Summary: *A Sequence Alterting Server with a WWW interface is described which informs users with query sequences in database searches about new entries in protein databases related to their query.*
Availability: *The server address is http://www.bork.embl-heidelberg.de/alerting/*
Contact: *bork@embl-heidelberg.de*

Database searches have become powerful tools in biological research. However, (i) the accumulated information is becoming more redundant as the same sequence appears simultaneously in several databases; (ii) owing to the continuous updating of databases, a search will be valid only temporarily.

In order to circumvent these difficulties, we created a Sequence Alerting Server (address: http://www.bork.embl-heidelberg.de/Alerting/) which regularly compares users' subscribed sequences to the new entries in all the public protein databases by using the BLAST programs (Altschul *et al.*, 1994). The server is based on a non-redundant protein database, nrdb, which contains every publicly available protein sequence once, eliminating the multiple appearance of the same sequence in the different protein databases. Users can subscribe their query sequences via a WWW interface. The subscribed sequences are searched daily against the new subset of the nrdb (see below); thus, users are kept informed about new entries in protein databases related to their queries.

The non-redundant protein database, nrdb, is built up every day at EMBL, using a script provided by the NCBI (W.Gish) from the following databases: SWISSPROT, PDB, TREMBL, PIR, GENPEPT and WORMPEPT. As new entries arrive daily in the databases, nrdb is continuously growing. The nrdb database is stored in FASTA format (each entry consisting of one annotation line and the sequence), the annotation line listing all the entries whose sequences are identical to the nrdb entry in question. They are called 'synonyms'. For example, the SWISSPROTNEW entry MGLA_SALTY has two synonyms, ST40492_1 and 1345116, in TREMBL and GENPEPT, respectively: >/:swissprotnew|P23924|MGLA_SALTY GALACTOSIDE TRANSPORT ATP BINDING PROTEIN MGLA.//: trembl|U40492|ST40492_1 gene: 'mglA'; product: 'MglA'; Salmonella typhimurium galactoside transport ATP-binding protein //:genpept|U40492|1345116 MglA/ [Salmonella typhimurium] MGSTISPPSGEYLLEMRGINKSFPGVK-ALDNVNLNVRPHSIHALMGENGAGKST...

To derive a daily difference non-redundant database, ddnrdb, containing only the differing database entries between two subsequent days' nrdbs, they were compared in the following way: the previous day's entry list is compared to the actual nrdb database, and only those entries are included in ddnrdb which neither in their names, nor in their list of synonymous names, contain any entry of the previous day's entry list. The selection procedure ensures that if the same sequence appears later as a distinctive entry with a different name in another protein database, it will still appear only once in ddnrdb.

The server performs a BLASTP or a BLASTX search for the subscribed protein and DNA sequences, respectively, on a daily basis. Users are allowed to customize several parameters when they subscribe their query sequences (see Figure 1).

To test the alerting server, we subscribed a SWISSPROT entry, CC4H_HUMAN, in June 1996. The protein is said to be similar to a yeast cell division control protein (Feuchter *et al.*, 1992), but we were unable to reproduce this similarity. However, since last June, several hits were recorded by the Alert: (i) human and other mammalian 'beige' proteins (trembl:HSU67615_1 and trembl:MMU52461_1) which are known to play a role in Chediak–Higashi syndrome (Nagle *et al.*, 1996); (ii) a lysosomal trafficking regulator protein (trembl:MMU70015_1; Barbosa *et al.*, 1996); (iii) several *Caenorhabditis elegans* proteins of unknown function (trembl:CEF35G12_14, trembl:CET01H10_8, swissprot:YSM2_CAEEL, swissprot:YSM3_CAEEL); (iv) several proteins containing the WD-domain (trembl:BGU49437_1, trembl:HVRACK1_1) known to play a role in signal trans-

[1]*To whom correspondence should be addressed*

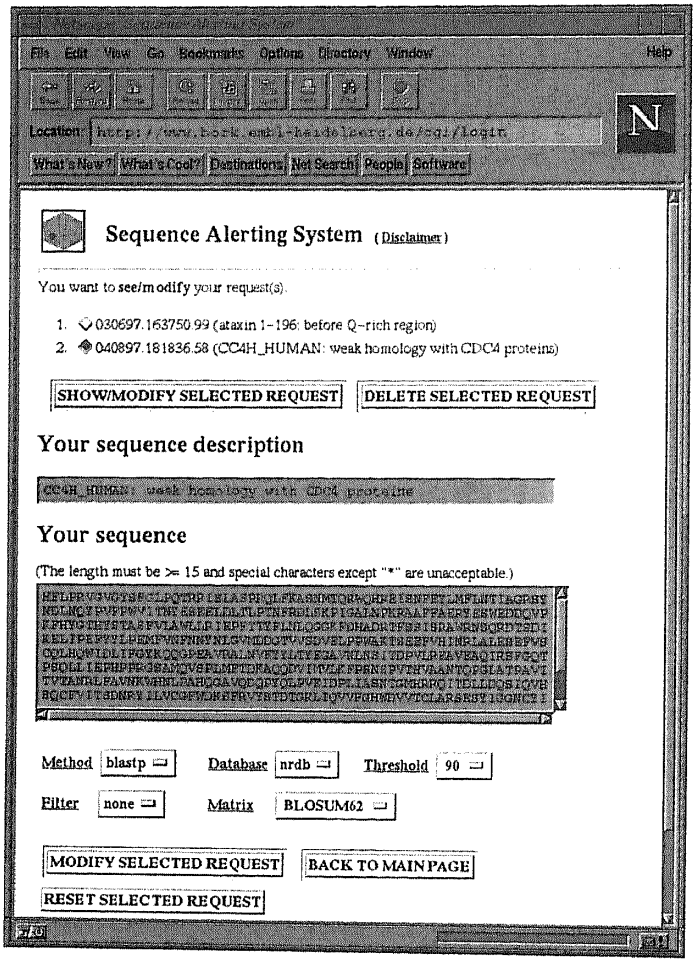

Fig. 1. The World Wide Web interface to the Sequence Alerting Server. Users can see/modify/delete their subscribed sequences and customize several parameters according to their needs, such as: (i) method; (ii) database; (iii) score threshold; (iv) filter; (v) comparison matrix (default values are provided). For optimizing scoring and other parameters, see Altschul *et al.* (1994).

duction (Neer *et al.*, 1994); (v) a human FAN protein (trembl:HSFAN_1), which is known to couple the p55 TNF-receptor to neutral sphingomyelinase (Adam-Klages *et al.*, 1996). Taken together, these findings show that the protein may play a role in important biochemical pathways and regulation of vesicular transport in the cell, rather than in cell division, as was originally suggested.

The server provides a reliable tool for those researchers who do not want to search protein databases regularly themselves in order to remain updated about new database entries related to their sequences of interest. The server is user friendly, easy to use, and unlike FastAlert (Eggenberger *et al.*, 1996) and DBWatcher (http://www-igbmc.u-strasbg.fr/BioInfo/LocalDoc/DBWatcher/), it does not require any specific software to be installed locally, only a WWW browser. It performs the searches in the most complete protein database, nrdb, unlike BEN's awareness tools (URL address: http://ben.vub.ac.be/ben/CSSA.html) and Swiss-Shop (http://expasy.hcuge.ch/swisshop/SwissShopReq.html), which use the translated EMBL and SWISSPROT, respectively.

The difference database generating procedure makes sure that every protein sequence appears only once in the system, even if it appears in several protein databases at different times. It performs the searches in the most complete publicly available protein database, nrdb. The system is robust, running on a Silicon Graphics Power Challenge with 16 75 MHz R8000 processors and 2 GBytes of main memory; one BLAST search takes only a few seconds.

References

Adam-Klages,S., Adam,D., Wiegmann,K., Struve,S., Kolanus,W., Schneider-Mergener,J. and Kronke,M. (1996) FAN, a novel WD-repeat protein, couples the p55 TNF-receptor to neutral sphingomyelinase. *Cell*, **86**, 937–947.

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

Barbosa,M.D.F.S. *et al.* (1996) Identification of the homologous beige and Chediak-Higashi syndrome genes. *Nature*, **382**, 262–265.

Eggenberger,F., Redaschi,N. and Doelz,R. (1996) FastAlert—an automatic search system to alert about new entries in biological sequence databanks. *Comput. Applic. Biosci.*, **12**, 129–133.

Feuchter,A.E., Freeman,J.D. and Mager,D.L. (1992) Strategy for detecting cellular transcripts promoted by human endogenous long terminal repeats: identification of a novel gene (CDC4L) with homology to yeast CDC4. *Genomics*, **13**, 1237–1246.

Gish,W. ftp://ncbi.nlm.nih.gov/pub/nrdb/README

Nagle,D.L. *et al.* (1996) Identification and mutation analysis of the complete gene for Chediak-Higashi syndrome. *Nature Genet.*, **14**, 307–311.