3792–3794 *Nucleic Acids Research, 2003, Vol. 31, No. 13*
DOI: 10.1093/nar/gkg535

BLAST2SRS, a web server for flexible retrieval of related protein sequences in the SWISS-PROT and SPTrEMBL databases

Konstantinos Bimpikis, Aidan Budd, Rune Linding and Toby J. Gibson*

European Molecular Biology Laboratory, Postfach 10.2209, 69012 Heidelberg, Germany

Received February 13, 2003; Revised and Accepted March 25, 2003

ABSTRACT

SRS (Sequence Retrieval System) is a widely used keyword search engine for querying biological databases. BLAST2 is the most widely used tool to query databases by sequence similarity search. These tools allow users to retrieve sequences by shared keyword or by shared similarity, with many public web servers available. However, with the increasingly large datasets available it is now quite common that a user is interested in some subset of homologous sequences but has no efficient way to restrict retrieval to that set. By allowing the user to control SRS from the BLAST output, BLAST2SRS (http://blast2srs.embl.de/) aims to meet this need. This server therefore combines the two ways to search sequence databases: similarity and keyword.

INTRODUCTION

BLAST2 is the most widely used tool to query databases by sequence similarity search (1). The major BLAST2 webservers, as at NCBI, EBI or ExPASy, allow the user to retrieve sequences returned in the output. This is essential for enabling further analyses such as multiple alignment, calculating a tree, investigating residue conservation and other forms of comparative sequence analysis.

Often a user will be interested in a defined subset of sequences, for example just mammalian or just purple bacteria. Servers such as at NCBI and ExPASy allow prior taxonomic restriction of the database, saving the computational search time.

However, we find that there is a more common situation in which the user will not be interested in all the protein sequences found by BLAST, which, for example in the case of protein kinases, will be thousands, but that the set of interesting sequences is dependent on what is returned by the search and, as the analysis of the hits proceeds, may need continual revision. As an example, if one is interested in the evolution of vertebrate multigene families, an invertebrate outgroup may be needed. But which? Chordates like amphioxus or *Ciona* would be desirable but the sequence may not have been determined yet. The *Drosophila* proteome is available but the homologous protein may be absent—true for many vertebrate extracellular matrix proteins in this soft-centred organism for example. Shall we next try *Caenorhabditis* or *Aplysia*? Our choice of sequence to use as outgroup depends on what sequences are found to be available.

While the user can click through an output list to find out which species are present, the cryptic entry IDs used by SPTrEMBL (and other databases except SWISS-PROT) are a hindrance. Furthermore no major BLAST server provider currently includes sufficient useful information like species or gene name in the output to allow rapid perusal, even though this information can be easily parsed into the BLAST-formatted binary databases.

Experienced SRS users can run BLAST from within an SRS server (e.g. http://srs.ebi.ac.uk/) and use query linking to select desired data subsets. The approach is powerful, though difficult for naive users. More recently BLAST at NCBI has allowed the user to apply ENTREZ keyword queries to select sequence entries from the output. As noted above, however, both servers currently provide uninformative BLAST outputs so the user may still have difficulty determining what is present.

To meet the need for flexible on-line retrieval, both for ourselves and others, we want a server that allows the user to apply taxonomic and other keywords to delimit sequence collection from BLAST output. One way to do this is by harnessing the SRS retrieval tool (2,3) in tandem to BLAST, principally by setting up the BLAST web output with SRS controls. A comprehensive sequence similarity search is initially performed, from which the results, a list of database hits, are saved for subsequent rounds of selection/filtering and viewing by different criteria. Additionally, by carrying useful database annotation into the BLAST output, the user can be better informed as to what is available for retrieval.

Here we summarise the functionality of the BLAST2SRS server. It is designed for flexible retrieval of subsets of related sequence entries in the SWISS-PROT and SPTrEMBL protein sequence databases (4). The focus is on this alone and features, such as graphical alignment summaries, found on more sophisticated user interfaces are not implemented here. Nor is this server aimed at maximising sensitive detection of remote homologues, where PSI-BLAST (1), Profile (5) or HMM (6) searches would be more appropriate.

*To whom correspondence should be addressed. Tel: +49 6221387398; Fax: +49 6221387517; Email: toby.gibson@embl.de

Table 1. The set of controls available in the BLAST2SRS header

Name of button	Purpose of button or text box
Species selection	Check boxes for every species found in the output.
Exclude fragments	Exclude any known partial sequences (as reported in the database DE record). On by default.
Cutoff value	Type in the desired E-value cutoff.
Update sequence selection	Reselect the sequences in the output list according to the current header settings.
Invert selection	Invert the current selection.
Get sequences in FASTA format	Collect the selected sequences in a new window in the widely used FASTA format. Names are made by concatenating the entry ID, a five character species identifier and the gene name.
Type in SRS query	Type in a keyword or several keywords separated by logical operators.
Send query to SRS	Send a query to SRS combining the current sequence selection and the SRS text query.

BLAST2SRS SERVER DETAILS

To meet our aims with the server we cannot support arbitrary flat file database formats. Therefore only the SWISSALL database group (4) is used, allowing for consistent parsing and preprocessing of the database. This includes the SWISS-PROT database, the best annotated of all the protein sequence databases but with a relatively small set of entries (4). SWISSALL is updated weekly and now includes >1 000 000 protein sequences. It should be fairly comprehensive as regards annotated genes in sequence data processed by the major DNA databases; for up-to-the-minute data, the user should access individual genome servers. Our formatted binary BLAST database includes the species, gene name, description and fragment fields in addition to the ID and Accession numbers.

The server is run on an IBM Intel processor-based FreeBSD server using Apache 2. Parser scripts are written in Python using available library routines for flat file parsing (7). NCBI BLAST2 (1) version 2.2.5 is currently used. The BLAST XML output is parsed into an HTML format with a species list in the header, E-value cutoff and other controls. Links are provided to the EMBL SRS server (currently SRS version 5; http://srs.embl-heidelberg.de:8000/srs5/). BLAST2SRS is at http:/blast2srs.embl.de/.

In a future revision, the parser will be rewritten to replace the XML with the standard BLAST output. Using the XML output has resulted in an inefficient parser that impacts server performance.

USING BLAST2SRS

When a user requires an arbitrary set of sequences (arbitrary from the perspective of the server provider), they will have several main reasons in mind for keeping or rejecting entries. These include the significance of the match (E-value), the species, the protein/gene name (when there are paralogous groups in a multiprotein family), whether the protein is a fragment (useless for phylogeny and many other analyses). These are foreseeable requirements and specific control is provided to the user by BLAST2SRS. A list of included species is given in the header: very useful in those situations where taxonomy is important. There are less foreseeable requirements such as desiring a higher level taxonomic node (e.g. Chordata) and/or wishing to exclude transmembrane receptors from the output of a tyrosine kinase query: SRS query language can support these demands so a query input box is available to the user.

The user invokes BLAST2SRS in the usual way by copy-pasting a protein sequence into the input box, choosing the database and filtering option and submitting the search. Limited control over other BLAST parameters is also available. The search and output parsing will take longer than typical for one of the major service providers, a minute or more depending on query, DB and output sizes. The output page can be bookmarked to guarantee retrieval and to return to at a later date. Table 1 summarises the retrieval controls that are to be found in the output header.

Figure 1 shows an example of output from BLAST2SRS. Imagine we are interested in the evolution of TFIIB transcription factors and we want to know if any occur in prokaryotic organisms. Using the human TFIIB sequence (SWISS-PROT Accession Q00403) we search SWISSALL with BLAST2SRS. There are >70 significant matches from ~40 species. [In this case there is no clear border between the true and false matches with eukaryotic paralogous sequences (BRF/TFIIIB) scoring a little below the default E-value cutoff at e^{-3}.] Archaea such as *Sulfolobus* and *Pyrococcus* are represented with good E-values ($\ll e^{-20}$). Archaeal experts can simply tick the archaeal species listed, update the selection and then collect the sequences in FASTA format, suitable, for example, for ClustalX multiple alignment (8). Less knowledgeable researchers can select 'All Species' and type 'archaea' into the SRS text box and use SRS to retrieve all the archaeal sequences, choosing formats such as FASTA or SWISS-PROT. At a quick glance there are no significant eubacterial matches in the list but, to be sure, a query with 'bacteria' in the SRS text box will confirm that there are none. We have now established that TFIIB homologues exist in archaea but not in eubacteria and retrieved the significant archaeal matches for further processing. This example has a smallish list of matches so is easy to work through to become familiar with BLAST2SRS: the utility of the server increases with the size of the BLAST output since it is impractical to work through hundreds of entries by individual retrieval and scrutiny.

Why use BLAST2SRS?

BLAST2SRS can be helpful in any situation where both homology and keyword are needed to define a set of

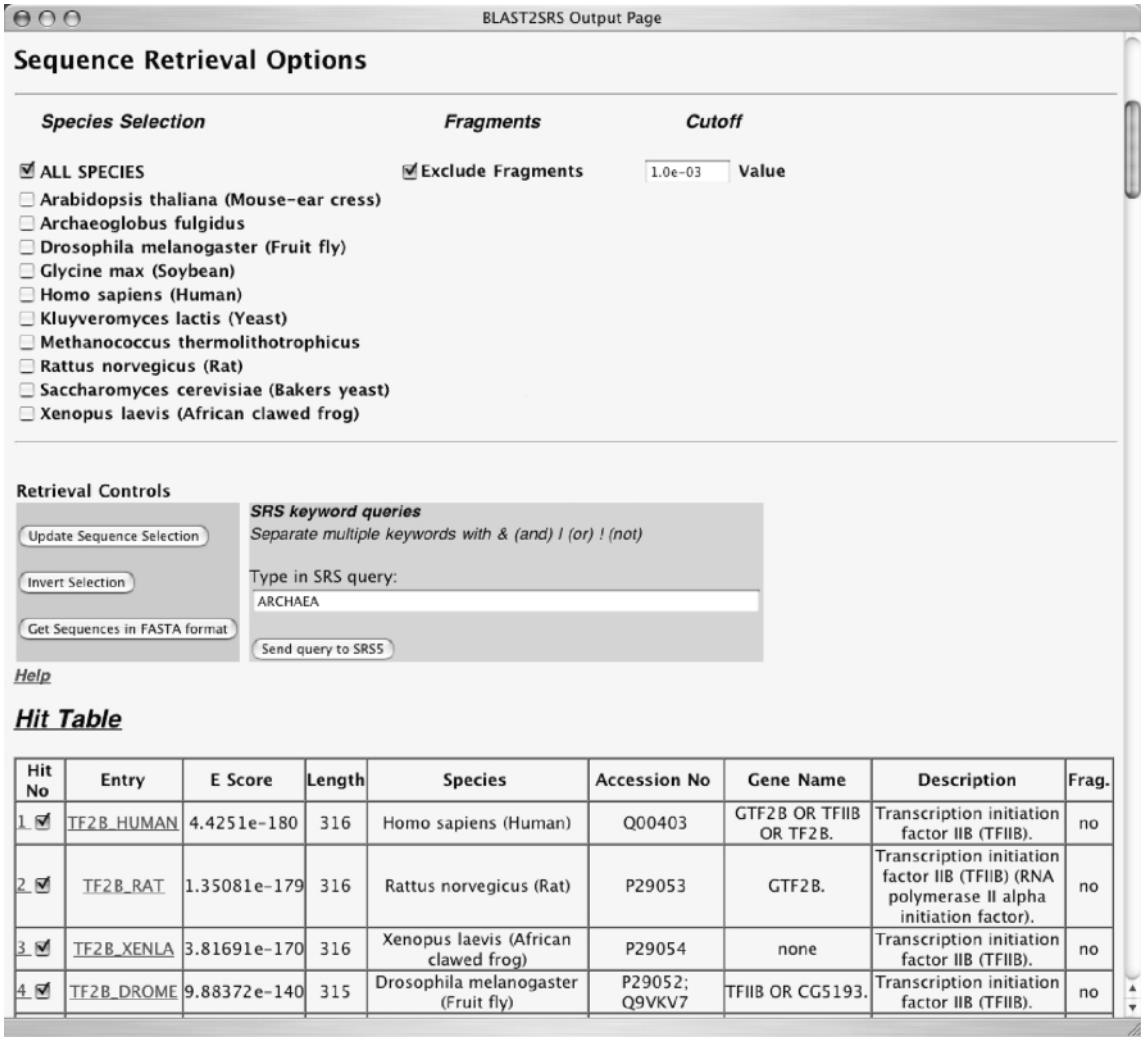

Figure 1. Detail of BLAST2SRS TFIIB output header showing the species list and control buttons. 'ARCHAEA' has been typed into the SRS query box: clicking on [Send query to SRS] will retrieve the archaeal sequences, excluding any fragments. The number of hits displayed was limited to 10 to make the figure.

sequences. This list illustrates some typical situations where this is true:

- all human entries below E-value 10^{-20}
- all human and *Drosophila*, excluding any sequence fragments
- vertebrate entries only
- animals, excluding chordates
- insect entries only
- green plant entries only
- eukaryotic entries only
- protein kinases, excluding tyrosine kinases

ACKNOWLEDGEMENTS

We thank Chenna Ramu and Christine Gemünd for help, discussions and Python modules.

REFERENCES

1. 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.*, **25**, 3389–3402.
2. Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.
3. Zdobnov,E.M., Lopez,R., Apweiler,R. and Etzold,T. (2002) The EBI SRS server-new features. *Bioinformatics*, **18**, 1149–1150.
4. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
5. Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
6. Eddy,S.R., Mitchison,G. and Durbin,R. (1995) Maximum discrimination hidden Markov models of sequence consensus. *J. Comput. Biol.*, **2**, 9–23.
7. Ramu,C., Gemund,C. and Gibson,T.J. (2000) Object-oriented parsing of biological databases with Python. *Bioinformatics*, **16**, 628–638.
8. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **25**, 4876–4882.