Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W143–W146***
doi:10.1093/nar/gkl157

Cascade PSI-BLAST web server: a remote homology search tool for relating protein domains

R. Bhadra[1], S. Sandhya[1,2], K. R. Abhinandan[1], S. Chakrabarti[2], R. Sowdhamini[2] and N. Srinivasan[1,*]

[1]Molecular Biophysics Unit, Indian Institute of Science, 560 012, Bangalore, India and [2]National Center for Biological Sciences, Tata Institute of Fundamental Research, GKVK Campus, Bellary Road, Bangalore 560 065, India

Received February 3, 2006; Revised February 22, 2006; Accepted March 20, 2006

ABSTRACT

Owing to high evolutionary divergence, it is not always possible to identify distantly related protein domains by sequence search techniques. Intermediate sequences possess sequence features of more than one protein and facilitate detection of remotely related proteins. We have demonstrated recently the employment of Cascade PSI-BLAST where we perform PSI-BLAST for many 'generations', initiating searches from new homologues as well. Such a rigorous propagation through generations of PSI-BLAST employs effectively the role of intermediates in detecting distant similarities between proteins. This approach has been tested on a large number of folds and its performance in detecting superfamily level relationships is ∼35% better than simple PSI-BLAST searches. We present a web server for this search method that permits users to perform Cascade PSI-BLAST searches against the Pfam, SCOP and SwissProt databases. The URL for this server is http://crick.mbu.iisc.ernet.in/~CASCADE/CascadeBlast.html.

INTRODUCTION

Remote homology sequence search tools are powerful methods for detecting distant and non-obvious similarities among proteins. Such methods are often challenged with detecting relationships in proteins that are usually restricted to the realm of structural comparisons. The detection of such similarities that are often exemplified by poor sequence similarities remains the Holy Grail of most sequence search methods. Databases such as SCOP (1) and CATH (2) perform a hierarchical clustering of protein structures and help define evolutionary relationships. The SCOP database clusters proteins that show high similarities in sequence, structure and function into families such that an evolutionary relationship between family members is expected. It also defines superfamily as a collection of protein families that show similarities in structure and function but poor sequence identity.

Several interesting approaches have employed in different ways, profiles (3,4), templates (5), HMMs (6) and intermediate sequences (7–9) to detect such relationships. We have developed recently a remote homology search approach (10) that establishes such relationships by cascading an entirely sequence analysis based method. The principle behind this approach is a rigorous implementation of an intermediate sequence based search procedure through an extensive application of PSI-BLAST (3,10). In such procedures, intermediate sequences serve as 'links' and bridge the sequence gap between proteins. Earlier applications of such search methods (8,9) overcome the intensity of the search by restricting the propagation of searches through a few representative hits. Programs that perform such searches are available in the public domain for download. This is the first implementation of an online server which performs cascade propagation of PSI-BLAST searches through all hits identified for a query. Since the searches are rigorous and time-intensive depending on the query and database selected, results are intimated to the users through e-mail response at the end of every generation.

CASCADE PSI-BLAST APPROACH

Briefly, in the cascade PSI-BLAST search method (10), we propagate PSI-BLAST searches through hits identified at the end of each search until no new hits are detectable through further searches. The rigorous propagation through each hit can help overcome the query-dependence and asymmetry of

*To whom correspondence should be addressed. Tel: +91 80 2293 2837; Fax: +91 80 2360 0535; Email: ns@mbu.iisc.ernet.in
Present addresses:
K. R. Abhinandan, Department of Biochemistry and Molecular Biology, University College, London, Gower Street, London WC1E 6BT, England
S. Chakrabarti, Computational Biology Branch, National Center for Biotechnology, Information (NCBI), National Library of Medicine (NLM), NIH, 8600, Rockville Pike, Bethesda, MD 20894, USA

traditional use of PSI-BLAST (11). The search is initiated against a database with a single query and we term this 'a first generation' search. All hits identified at the end of this search serve as queries in what we term as 'second generation' to possibly detect hits not identified in the earlier search. New hits, if detected are then used to propagate further generations of the search. Such a cascade propagation of PSI-BLAST searches continues until a convergence is reached and no new hits are detectable. An intense propagation of PSI-BLAST through such cascaded searches for a single query results in, on an average, 500–1000 individual PSI-BLAST searches through all detected hits over many generations.

CASCADE PSI-BLAST WEB SERVER

The web-server implementation of our Cascade PSI-BLAST search protocol (Figure 1) accepts a protein sequence as an input. It is best if the query corresponds potentially to a protein domain. The databases available for querying the sequence are the SwissProt-Release **48.9** (12), SCOP-Release 1.69 (1) and PFAM-Release 19.0 (13) databases. Since this is a very time-intensive process and each hit is given a chance to scan protein sequence space, we have chosen databases that contain experimentally verified sequences and larger databases, that may compromise on speed and time taken to perform the searches, such as the NR are not included. The homologues are identified in cascaded searches within the database selected by the user. Default *E*- and *H*-values of 0.001 and 0.0001 are employed. However, the user can change the values if desired. Also, to avoid false positives being detected in the searches, a length alignment filter of 75% is employed in the searches. This filter may also be relaxed or made more stringent by the user. In addition, the user can choose low complexity filter options for the searches. The web server is interactive and allows selection of those hits through which further generations of the search are propagated. At the end of each generation, an Email, with a link to the hits, is sent to the user. The result pages are descriptive with annotations for each hit and indicate the domain boundaries of the hits identified in the PSI-BLAST searches. The detailed PSI-BLAST output files that contain alignments are made available to the user

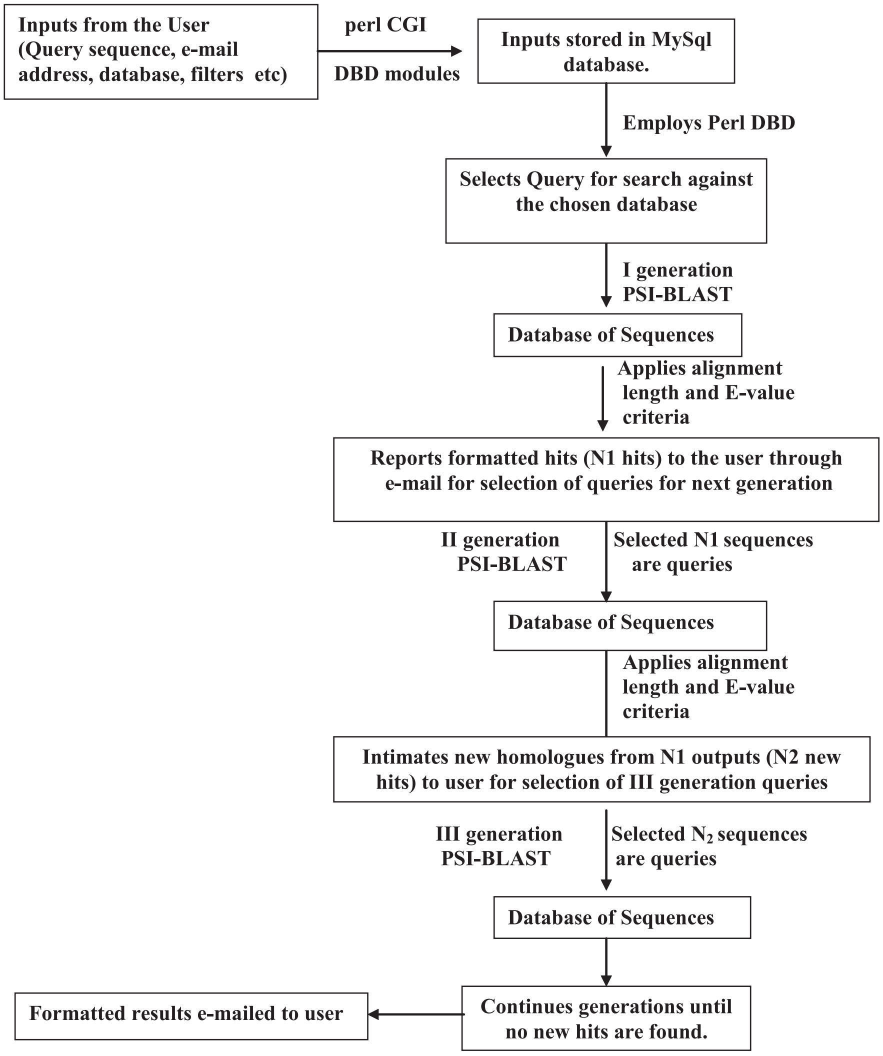
Inputs from the User (Query sequence, e-mail address, database, filters etc)
perl CGI
DBD modules
Inputs stored in MySql database.
Employs Perl DBD
Selects Query for search against the chosen database
I generation PSI-BLAST
Database of Sequences
Applies alignment length and E-value criteria
Reports formatted hits (N1 hits) to the user through e-mail for selection of queries for next generation
II generation PSI-BLAST
Selected N1 sequences are queries
Database of Sequences
Applies alignment length and E-value criteria
Intimates new homologues from N1 outputs (N2 new hits) to user for selection of III generation queries
III generation PSI-BLAST
Selected N₂ sequences are queries
Database of Sequences
Continues generations until no new hits are found.
Formatted results e-mailed to user

Figure 1. Flowchart of the Cascade PSI-BLAST web server.

as a gzipped tar file at the end of every generation. All hits identified at the end of each generation of the search are made available for download in FASTA format for subsequent analysis. In order to overcome limitations on computing time involving multiple server requests, the next generation of the search can be seeded through a maximum of 30 sequences at a time. If the user wishes to seed the searches in a generation with >30 sequences, such hits may be selected as queries subsequent to the completion of the previous runs. A maximum of up to four generations of the search is permitted although our earlier assessments indicate that up to three generations of the search can effectively detect most of the remote hits for a query. As can be seen in Figure 1, hits identified at each generation of the search are stored in a MYSQL database and these are queried through a PERL-CGI interface. Programs that employ Perl DBD modules are used to process the searches and results. Links to results are notified to the user through Email since searches can be computationally intensive and time consuming. Results are presented in two formats as clickable links in the Email sent to the user at the end of every generation. Link 1 lists the annotated sequence details of the hits with *E*-values at which detection is made. These sequences correspond to the alignment region reported in the search and not to the full length of the protein and may be used to generate sequence libraries or multiple alignments of all remote homologues identified by the search since the sequences may be downloaded. Link 2 presents the results in tabular format. This representation allows a quick assessment of those hits in each generation that have served as the actual ‘intermediate links’ in propagating the searches and elucidating the evolutionary relationships. In searches in Pfam (13), the protein family name of the hits associated with the query is provided. Protein annotations are also made available as links in searches involving the SwissProt database. The scop code of the hits is provided in searches in the SCOP database (1). This is a very useful feature since the examination of these defined codes allows the user to quickly ascertain whether the hits appearing in each generation of the search belong to the same family, superfamily or fold of the query. Such an analysis allows evaluation of the evolutionary significance of the relationships detected. We had demonstrated such relationships in the superfamilies of the TIM and Globin folds (10). In our earlier analysis, we have shown an improvement in coverage of family members of ∼15% and detect ∼35% more superfamily level relationships than typical PSI-BLAST searches (10).

CONCLUSIONS

Proteins may evolve to the extent that they have little similarity in sequence and yet preserve their inherent relationships through similarities in structure and function. When sequences mutate so extensively, it is difficult to ascertain whether they may be true relatives through pure sequence analysis. Structural comparisons are often successful in detecting such remote similarities as seen in SCOP and CATH databases (1,2). Park *et al*. and others (7,10) have shown that intermediate sequences can serve as missing links and bridge protein sequence space. With the escalation in genome sequence initiatives, there is an excellent and exciting opportunity to employ these sequences in a rigorous manner to scan protein space further and detect more meaningful hits.

Large sequence dispersions within protein families and superfamilies are often the cause for difficulties in detecting evolutionary relationships. While search methods such as PSI-BLAST are powerful in detecting most family relationships, we have shown an improved coverage at the level of both family and superfamily since more than one generation of the search is effective in detecting related members of a protein query (10). Large sequence dispersions may also result in sub-clustering of family members. Thus, ‘profile traps’, caused due to over-representation of some of the sub-clusters often become a bottleneck in detecting other members. Since every hit can propagate the search in a non-uniform direction, we attempt to overcome the inherent asymmetry of other approaches (10). Our search method has exciting implications in fold recognition and in detecting distant similarities through an entirely sequence analysis based approach.

ACKNOWLEDGEMENTS

R.S. thanks NCBS (TIFR) for financial and infrastructural support. N.S. thanks the Department of Biotechnology, New Delhi for support through the computational genomics project. This work is supported by the award of Senior Research Fellowships in Biomedical Sciences by the Wellcome Trust, London to R.S. and N.S. R.B. is supported by Department of Biotechnology, New Delhi. S.S. and K.R.A. were also supported by the Wellcome Trust, London. S.S. is a Senior Research Fellow of the Council of Scientific and Industrial Research, India. The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*., **247**, 536–540.
2. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
3. 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.
4. Sandhya,S., Kishore,S., Sowdhamini,R. and Srinivasan,N. (2003) Effective detection of remote homologues by searching in sequence dataset of a protein domain fold. *FEBS Lett*., **552**, 225–230.
5. Yi,T.M. and Lander,E.S. (1994) Recognition of related proteins by iterative template refinement (ITR). *Protein Sci*., **3**, 1315–1328.
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. Park,J., Teichmann,S.A., Hubbard,T. and Chothia,C. (1997) Intermediate sequences increase the detection of homology between sequences. *J. Mol. Biol*., **273**, 349–354.
8. Salamov,A.A., Suwa,M., Orengo,C.A. and Swindells,M.B. (1999) Combining sensitive database searches with multiple intermediates to detect distant homologues. *Protein Eng*., **12**, 95–100.

9. Li,W., Pio,F., Pawlowski,K. and Godzik,A. (2000) Saturated BLAST: an automated multiple intermediate sequence search used to detect distant homology. *Bioinformatics.*, **16**, 1105–1110.
10. Sandhya,S., Chakrabarti,S., Abhinandan,K.R., Sowdhamini,R. and Srinivasan,N. (2005) Assessment of a rigorous transitive profile based search method to detect remotely similar proteins. *J. Biomol. Struct. Dyn.*, **23**, 283–298.
11. Aravind,L. and Koonin,E.V. (1999) Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *J. Mol. Biol.*, **287**, 1023–1040.
12. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
13. Sonnhammer,E.L., Eddy,S.R., Birney,E., Bateman,A. and Durbin,R. (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.*, **26**, 320–322.