Nucleic Acids Research, 2006, Vol. 34, Database issue ***D243–D246***
doi:10.1093/nar/gkj043

MulPSSM: a database of multiple position-specific scoring matrices of protein domain families

V. S. Gowri[1], O. Krishnadev[1], C. S. Swamy[1,2] and N. Srinivasan[1,*]

[1]Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, India and
[2]National Centre for Biological Sciences, GKVK Campus, Bangalore 560065, India

Received August 15, 2005; Revised and Accepted October 3, 2005

ABSTRACT

Representation of multiple sequence alignments of protein families in terms of position-specific scoring matrices (PSSMs) is commonly used in the detection of remote homologues. A PSSM is generated with respect to one of the sequences involved in the multiple sequence alignment as a reference. We have shown recently that the use of multiple PSSMs corresponding to an alignment, with several sequences in the family used as reference, improves the sensitivity of the remote homology detection dramatically. MulPSSM contains PSSMs for a large number of sequence and structural families of protein domains with multiple PSSMs for every family. The approach involves use of a clustering algorithm to identify most distinct sequences corresponding to a family. With each one of the distinct sequences as reference, multiple PSSMs have been generated. The current release of MulPSSM contains ~33 000 and ~38 000 PSSMs corresponding to 7868 sequence and 2625 structural families. A RPS_BLAST interface allows sequence search against PSSMs of sequence or structural families or both. An analysis interface allows display and convenient navigation of alignments and domain hits. MulPSSM can be accessed at http://crick.mbu.iisc.ernet.in/~mulpssm.

INTRODUCTION

Multiple sequence alignments of protein families are commonly represented as hidden Markov models (HMMs) (1–4) or position-specific scoring matrices (PSSMs) (5). Tremendous power of such protein profiles in enabling the detection of remote homologues is well known. For example, PSI_BLAST (5) uses PSSM generated, at the end of every round of search, as an input to the next round. Such a use of dynamic PSSMs is known to be extremely effective in the detection of distant relatives of query in the sequence database. RPS_BLAST and IMPALA use a database of static PSSMs corresponding to homologous proteins and enables rapid match of the query sequence with various PSSMs in the database (6,7). Programs employing PSSM matching algorithms are generally faster than the HMM matching programs and are commonly used in large-scale analyses (8,9).

Recently we have focused on an important and sensitive feature of PSSMs (10). It is the use of a reference sequence in the construction of a PSSM starting from a multiple sequence alignment. Reference sequence should be any one of the sequences involved in the multiple sequence alignment. Basically, PSSM integrates two kinds of information at every alignment position of a multiple sequence alignment. (i) Extent of occurrence of each of the 20 residue types. (ii) Potential of replacement of the residue in the reference sequence with each one of the 20 residue types. The reference sequence corresponds to the query in the case of PSI_BLAST. For RPS_BLAST searches, reference sequence for generating a PSSM is chosen arbitrarily from the multiple sequence alignment. Hence, the PSSMs generated for a multiple sequence alignment with different homologues as reference sequences will be different. It has been shown convincingly that the sensitivity of the PSSM is strongly dependent on the choice of the reference sequence. During RPS_BLAST searches use of multiple PSSMs corresponding to different reference sequences for a given alignment results in remarkably improved specificity, sensitivity and error rate compared with the use of single PSSM corresponding to an alignment and even HMM (10).

In this paper, we report the ready availability of multiple PSSMs for every multiple sequence alignment corresponding to large datasets of sequence and structural families. A web interface allows convenient use of RPS_BLAST to search these datasets of multiple PSSMs and analysis of results.

*To whom correspondence should be addressed. Tel: +91 80 2293 2837; Fax: +91 80 2360 0535; Email: ns@mbu.iisc.ernet.in

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

DATASETS, CLUSTERING OF SEQUENCES AND GENERATION OF PROFILES

MulPSSM database consists of PSSMs of protein domain sequence and structural families. The set of protein domain sequence families corresponds to Protein Families Database of alignments and HMMs (Pfam version 17.0), which consists of 7868 families (11,12). The seed alignments available in PfamA version have been used in the present work. The dataset of structural domain families has been obtained from the latest update of the database on phylogeny and alignment of homologous protein structures (PALI release 2.4) corresponding to 2625 families (13,14). The list of protein domain families of known structures in PALI release 2.4 has been obtained from 1.67 release of structural classification of proteins (SCOP) database (15,16). The multiple sequence alignments in PALI families have been derived on the basis of three-dimensional structural superposition of homologues using structural alignment of multiple proteins (STAMP, version 4.2) program (17). Homologues of yet unknown structure corresponding to the PALI families have been identified by mining the Universal Protein Resource (UNIPROT) database (18,19) using PSI_BLAST (E and h value cut-offs are 10^{-5} and 10^{-6} respectively; cut-off for both query length coverage in the alignment and sequence identity are fixed at 30%).

Generation of multiple PSSMs with two very closely related sequences as references is unlikely to be worthwhile and it will unnecessarily add to the computer time while searching the PSSM database. Hence choice of reference sequences from a multiple sequence alignment should be optimal so that it maximizes the sensitivity during search in the PSSM database and minimizes the computer time. Hence we clustered sequences in every family of protein domains and identified a set of most disperse sequences within each family. Blastclust was used for the generation of clusters of non-redundant sequences from the set of homologues of the family sequences (5). In order to identify optimal level of clustering following exercise was performed: the clustering was performed at cut-offs of 30–100% sequence identity for 250 randomly picked SCOP families. The sequences in PALI database were then queried against the multi PSSM database at each cut-off value. The number of true positives (Superfamily- or Fold-related sequences) and false positives (class-related or across class connections) were found for each cut-off value to assess the sensitivity and specificity of the profiles generated at each cut-off. It was found that at a sequence identity cut-off of 70%, the sensitivity of the method was equal to the sensitivity at 100% sequence identity cut-off whereas the number of profiles had decreased from ~25 000 at 100% cut-off to 14 000 at 70% cut-off. At a sequence identity cut-off of 50%, the sensitivity is 75% of the sensitivity at 100% cut-off. Apart from this point, considering the time taken for search that is dependent on the number of PSSMs in the database a cut-off of 50% was chosen as optimal. In our previous work, we have shown that in the multiple PSSM approach, the false positive rate is 2% at an E-value cut-off of 10^{-5} (10). The false positive rate remains the same after clustering the sequences at a cut-off of 50% sequence identity. For the structural families, the multiple sequence alignments were biased towards the structure-based alignment of the family. In principle, generation of profiles can be carried out using sequence-based alignments alone and it would make no difference to the quality of profiles if the sequences are not very divergent. In case of very divergent sequences, the structural alignments are more accurate than sequence alignments and thus in such cases, using the structural information will lead to more robust profiles.

GENERATION OF PSI_BLAST PROFILES

After identifying the reference sequences present in the multiple sequence alignment using the clustering algorithm, the following procedure is used for the generation of PSSMs. The multiple sequence alignment and the reference sequence are given as inputs to a PSI_BLAST run to iterate against a database of sequences present in the input multiple sequence alignment. In such iterative searches, any hit identified is already present in the multiple sequence alignment fed to the search program. Hence, the profile generated corresponds to the multiple sequence alignment fed to the PSI_BLAST program. The PSSMs output from such PSI_BLAST runs form MulPSSM database.

The current version of MulPSSM database consists of ~33 000 and 38 000 PSSMs corresponding to 7868 sequence and 2625 structural families, respectively. We believe, based on our earlier analysis (10), that the use of multiple profiles for each family can lead to better annotation of genome sequences than is provided by HMM based or single profile based searches. For example, the protein gil23508377 from *Plasmodium falciparum* is annotated as a hypothetical protein in the NCBI genome database. Using PSI_BLAST search or Pfam HMM based searches no known domain could be detected in the sequence with significant E-values. Using the multiple PSSM approach, we detected the relationship between this protein and the family of a Peptidase_M10. The E-value is 10^{-4} with a sequence identity of 22% over the alignment length of about 80 residues. Such observations generate reasonable hypothesis for exploration using experimental studies.

WEB INTERFACE

In the MulPSSM website (http://crick.mbu.iisc.ernet.in/~mulpssm), users can select datasets of PSSMs corresponding to either sequence families or structural families or both. Although we recommend an E-value cut-off of 10^{-4}, the choice of E-value is left to the users. The results are represented in pictorial form, with ASCII characters, to show the presence of various domains that may be present in the query sequence. Links are provided to the alignment corresponding to the PSSM with the best E-value, list of all the profiles within and across families that are hit to the query, links to the details of the families (Figure 1).

Lists of all the sequence and structural families are provided in the MulPSSM site with alignments and indication of reference sequences for each family. Reference sequences used in every family are shown highlighted and the user can download all the PSSMs corresponding to a given family for use in a local machine in a RPS_BLAST search. The complete set of all the PSSMs may be obtained from the authors upon request. This set is expected to be invaluable for rapid and sensitive genome-wide domain assignments.

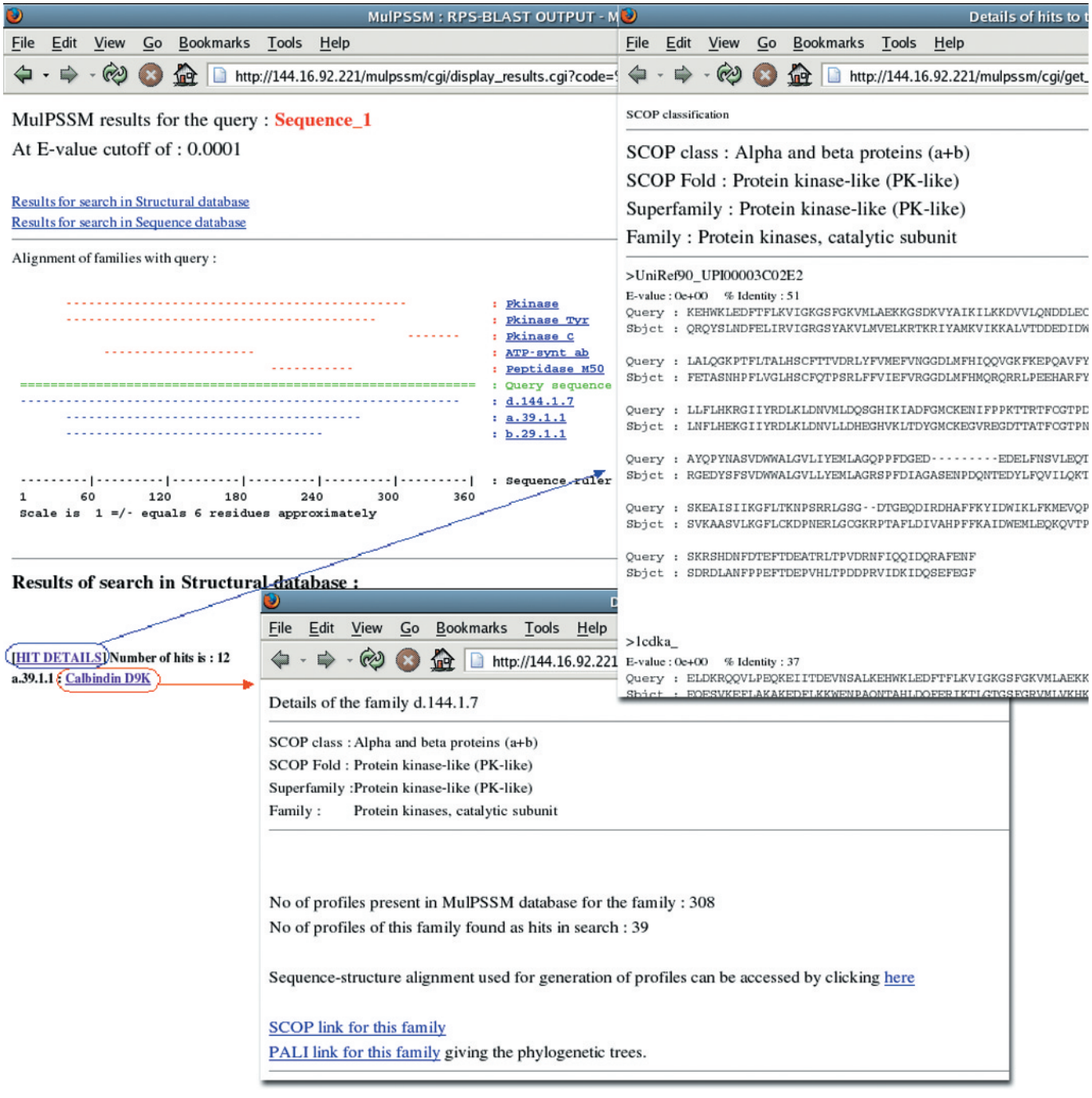

Figure 1. Typical outputs of a search made in MulPSSM site. The main window gives the alignment of the families found as hits in a semi-graphical layout. The main window has a link to the details of the multiple hits for a family (indicated by blue encircle) which opens in a new window (indicated by a blue arrow). Similarly, the details of the family can be obtained by clicking on the name of the family (indicated by a red encircle). The detailed view of the alignments and other features of a family hit (e.g. the fraction of PSSMs of a family found as hits) can help in assessing the accuracy of a hit.

ACKNOWLEDGEMENTS

O.K. and C.S.S. are supported by Council of Scientific and Industrial Research, New Delhi and Wellcome Trust, London, respectively. This research is supported by the award of an International Senior Fellowship in biomedical sciences to N.S. by the Wellcome Trust, UK, NMITLI project from CSIR, India and computational genomics project funded by the Department of Biotechnology, India. The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Baldi,P., Chauvin,Y., Hunkapiller,T. and McClure,M.A. (1994) Hidden Markov models of biological primary sequence information. *Proc. Natl Acad. Sci. USA*, **91**, 1059–1063.

2. Krogh,A., Brown,M., Mian,I.S., Sjolander,K. and Haussler,D. (1994) Hidden Markov models in computational biology. Applications to protein modelling. *J. Mol. Biol.*, **235**, 1501–1531.
3. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
4. Karplus,K., Barrett,C. and Hughey,R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.
5. 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.
6. Schaffer,A.A., Wolf,Y.I., Ponting,C.P., Koonin,E.V., Aravind,L. and Altschul,S.F. (1999) IMPALA: matching a protein sequence against a collection of PSI_BLAST-constructed position-specific score matrices. *Bioinformatics*, **12**, 1000–1011.
7. Marchler-Bauer,A., Panchenko,A.R., Shoemaker,B.A., Thiessen,P.A., Geer,L.Y. and Bryant,S.H. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.*, **30**, 281–283.
8. Muller,A., MacCallum,R.M. and Sternberg,M.J. (1999) Benchmarking PSI_BLAST in genome annotation. *J. Mol. Biol.*, **293**, 1257–1271.
9. Lindahl,E. and Elofsson,A. (2000) Identification of related proteins on family, superfamily and fold level. *J. Mol. Biol.*, **295**, 613–625.
10. Anand,B., Gowri,V.S. and Srinivasan,N. (2005) Use of multiple profiles corresponding to a sequence alignment enables effective detection of remote homologues. *Bioinformatics*, **21**, 2821–2826.
11. Sonnhammer,E.L., Eddy,S.R. and Durbin,R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**, 405–420.
12. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L.L. *et al.* (2004) The Pfam Protein Families Database. *Nucleic Acids Res.*, **32**, D138–D141.
13. Balaji,S., Sujatha,S., Kumar,S.S. and Srinivasan,N. (2001) PALI—a database of Phylogeny and ALIgnment of homologous protein structures. *Nucleic Acids Res.*, **29**, 61–65.
14. Gowri,V.S., Pandit,S.B., Karthik,P.S., Srinivasan,N. and Balaji,S. (2003) Integration of related sequences with protein three-dimensional structural families in an updated version of PALI database. *Nucleic Acids Res.*, **31**, 486–488.
15. 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.
16. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.
17. Russell,R.B. and Barton,G.J. (1992) Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, **14**, 309–323.
18. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2004) UniProt: the Universal Protein Knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
19. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.