The MetaFam Server: a comprehensive protein family resource

Kevin A. T. Silverstein, Elizabeth Shoop, James E. Johnson, Alan Kilian, John L. Freeman, Timothy M. Kunau, Ihab A. Awad, Margaret Mayer and Ernest F. Retzel*

Computational Biology Centers, Academic Health Center, University of Minnesota, Mayo Mail Code 43,
420 Delaware Street, SE Minneapolis, MN 55455-0312, USA

Received August 28, 2000; Revised and Accepted September 13, 2000

ABSTRACT

MetaFam is a comprehensive relational database of protein family information. This web-accessible resource integrates data from several primary sequence and secondary protein family databases. By pooling together the information from these disparate sources, MetaFam is able to provide the most complete protein family sets available. Users are able to explore the interrelationships among these primary and secondary databases using a powerful graphical visualization tool, MetaFamView. Additionally, users can identify corresponding sequence entries among the sequence databases, obtain a quick summary of corresponding families (and their sequence members) among the family databases, and even attempt to classify their own unassigned sequences. Hypertext links to the appropriate source databases are provided at every level of navigation. Global family database statistics and information are also provided. Public access to the data is available at http://metafam.ahc.umn.edu/.

INTRODUCTION

The classification of unknown protein sequences is a crucial problem in this era of large-scale genome sequencing projects. Thus it is not surprising that, even in browsing this database issue, one encounters numerous protein family classifications. Each of these databases defines its family members using different methodologies. PROSITE (1) characterizes families using regular expression patterns, and profiles (2). Blocks (3) and PRINTS (4) each use groups of ungapped position-specific scoring matrices (PSSMs) to model conserved regions that typify a family; Pfam (5) uses Hidden Markov Models (HMMs) (6,7), a gapped statistical generalization of both PSSMs and profiles. PIR (8) and SBASE (9) employ expert curators analyzing sequence similarity results and other data. Finally, DOMO (10,11), ProDom (12) and PROTOMAP (13) all utilize fully automated algorithms built upon sequence similarity protocols. This variety of methods yields a diverse set of family definitions. Surprisingly, there is both remarkable consistency and complementarity among them.

To find information on a specific protein or family of proteins, it has been necessary to consult each of these complementary databases sequentially, gleaning the desired information from each. It is very difficult to adequately compare each database's family sets using this approach. To aid the researcher, we have created a unified resource, MetaFam, to serve as a springboard to this rich set of complementary public data. MetaFam collects together each family database's notion of glucose-1-6-phosphatase, for example, into a single family superset. A convenient visualization tool allows the user to simultaneously view characteristics of each family within the superset, their overlap, their members and the full domain architecture of each member (as defined by each of the family databases). The methods used to create MetaFam (14,15) are automated. Thus we are able to provide updates at frequent intervals.

From the MetaFam server, users may also attempt to classify their own sequences. We have integrated the results of the native searching routines of many of these databases in a tool called PANAL (16). In this paper, we describe these and other tools accessed through the MetaFam server.

METAFAM CONTENT

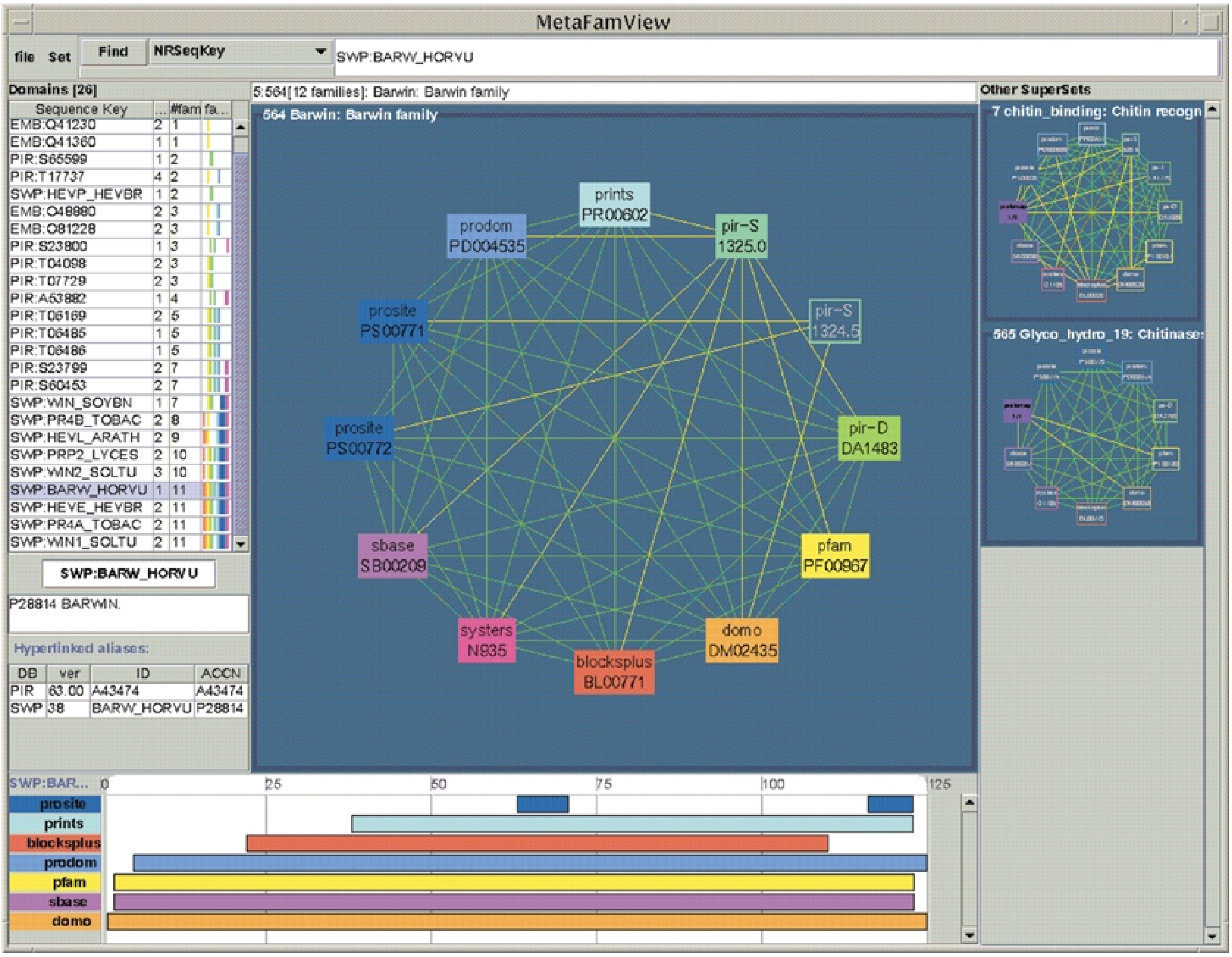
Non-redundant protein set

In order to properly relate protein membership among the various family databases, we have translated all protein sequence identifiers into a non-redundant set of protein sequence keys. Specifically, identical sequences found in SWISS-PROT and TrEMBL (17), GenPept (18), PIR and NRL3D (19) are mapped to these non-redundant keys.

MetaFam family supersets

MetaFam attempts to group together all corresponding families among the family databases. This allows the user to focus on a particular family, and to compare the results of the different classification methods. The central entity of MetaFam is called a family superset. A family superset representing the enolase family, for example, contains a set of enolase families from each database, and the union of the sequence-domain members pooled from each enolase family. Pairwise family correspondences

*To whom correspondence should be addressed. Tel: +1 612 625 5957; Fax: +1 612 626 6069; Email: metafam@ahc.umn.edu. Correspondence may also be addressed to Kevin A. T. Silverstein at the same address and phone number.



MetaFamView

file Set Find NRSeqKey SWP:BARW_HORVU

Domains [26]

Sequence Key	...	#fam	fa...
EMB:Q41230	2	1	
EMB:Q41360	1	1	
PIR:S65599	1	2	
PIR:T17737	4	2	
SWP:HEVP_HEVBR	1	2	
EMB:O48880	2	3	
EMB:O81228	2	3	
PIR:S23800	1	3	
PIR:T04098	2	3	
PIR:T07729	2	3	
PIR:A53882	1	4	
PIR:T06169	2	5	
PIR:T06485	1	5	
PIR:T06486	1	5	
PIR:S23799	2	7	
PIR:S60453	2	7	
SWP:WIN_SOYBN	1	7	
SWP:PR4B_TOBAC	2	8	
SWP:HEVL_ARATH	2	9	
SWP:PRP2_LYCES	2	10	
SWP:WIN2_SOLTU	3	10	
SWP:BARW_HORVU	1	11	
SWP:HEVE_HEVBR	2	11	
SWP:PR4A_TOBAC	2	11	
SWP:WIN1_SOLTU	2	11	

SWP:BARW_HORVU

P28814 BARWIN.

Hyperlinked aliases:

DB	ver	ID	ACCN
PIR	63.00	A43474	A43474
SWP	38	BARW_HORVU	P28814

5:564[12 families]: Barwin: Barwin family

Figure 1. A snapshot of MetaFamView, the graphical visualization tool providing access to the database. The features of MetaFamView are briefly summarized in the text and demonstrated in the Supplementary Material.

within the superset have been identified using the percentage of intersecting members. We have avoided using textual descriptions to make the match because of the potential for naming conflicts.

PANAL: an integrated resource for Protein sequence ANALysis

Often a user wishes to classify their own uncharacterized sequence. Several of the protein family databases provide software that searches for motifs and domains via their native method (e.g., profiles, HMMs). These programs appear on a variety of sites that output results in diverse formats. We have installed several of these programs on site, and provide a single interface and graphical output summary that integrates the results.

METAFAM ACCESS

The main page of the MetaFam server (http://metafam.ahc.umn.edu/) provides several levels of access to the contents of our database. Global statistics and summaries of the constituent databases are provided. Java server pages retrieve links to sequences or families of interest. A Java applet, MetaFamView, allows users to navigate the interrelationships among families and sequences stored in the database. Finally, a link to PANAL is provided to aid classification of new sequences. In the following subsections we further discuss access to the Java server pages and MetaFamView.

Java Server pages

Often a researcher has a sequence identifier or accession number from one of the primary sequence databases (e.g., GenPept), and wishes to know what that sequence is called at other sources (e.g., NRL3D, PIR, Swiss-Prot or TrEMBL). Users can simply type in the sequence identifier or accession on our alias page and obtain links to the corresponding sequences. Sometimes a researcher is interested in a particular family, and wishes to know the corresponding families at the different family databases, or the composite list of members. At our superset lookup page, users can type in a keyword, or the identifier/accession of a particular family or sequence to find the appropriate MetaFam superset. A summary of the superset, including links to all sequence and family members is returned.

MetaFamView

MetaFamView combines features of our quick-access Java server pages into an interactive graphical tool to peruse the database. An online help button summarizes features of the tool. Users may find supersets using keywords, sequence identifiers (or accession numbers) and family identifiers/accessions. The first superset appears in a central panel, with the remaining ones along the right side (Fig. 1). Families in the superset are represented as boxes on a wheel, and sequence-domain members appear as sorted lists along the left panel. The color of the lines connecting two families in a superset indicates the degree of agreement (i.e. number of intersecting members) between them. When a sequence is selected from the list, a full map of all domain boundary assignments on that sequence is shown. All of the views are coordinated in a manner that allows the user to compare the family and domain boundary assignments among the databases. Also, at every level, hypertext links to the original sequence and family sources are provided. A demonstration of MetaFamView is provided in the Supplementary Material, and at http://metafam.ahc.umn.edu/demo.

DEVELOPMENT AREAS

MetaFam has been steadily evolving over the past year, and will continue to do so. We will soon link the results of PANAL directly to MetaFam. Users will also be able to export sequences from a superset in FASTA format to be saved in files, or sent directly to Clustal-W (20) and a Java-based sequence alignment viewer. We are also working on improvements to MetaFam's basic clustering algorithm, as well as the speed of superset retrieval via keyword queries. Inclusion of other sequence-based family databases (21,22) and ultimately structure-based databases (23–25) are also being considered.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported by an SCA with the USDA-ARS, grants from NSF (9872565, 9872633), DOE (DE-FC07 97ID13550), USDA (58-3625-8-117) and the University of Minnesota Academic Health Center. Some hardware used in this work was supplied by an AEG from Sun Microsystems, Inc.

REFERENCES

1. Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.
2. Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
3. Henikoff,J.G., Greene,E.A., Pietrokovski,S. and Henikoff,S. (2000) Increased coverage of protein families with the Blocks database servers. *Nucleic Acids Res.*, **28**, 228–230.
4. Attwood,T.K., Croning,M.D.R., Flower,D.R., Lewis,A.P., Mabey,J.E., Scordis,P., Selley,J.N. and Wright,W. (2000) PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res.*, **28**, 225–227.
5. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.
6. Krogh,A., Brown,M., Mian,I.S., Sjölander,K. and Haussler,D. (1994) Hidden Markov Models in computational biology: Applications to protein modeling. *J. Mol. Biol.*, **235**, 1501–1531.
7. Eddy,S.R. (1996) Hidden Markov Models. *Curr. Opin. Struct. Biol.*, **6**, 361–365.
8. Srinivasarao,G.Y., Yeh,L.-S.L., Marzec,C.R., Orcutt,B.C., Barker,W.C. and Pfeiffer,F. (1999) Database of protein sequence alignments: PIR-ALN. *Nucleic Acids Res.*, **27**, 284–285.
9. Murvai,J., Vlahovicek,K., Barta,E., Cataletto,B. and Pongor,S. (2000) The SBASE protein domain library, release 7.0: a collection of annotated protein sequence segments. *Nucleic Acids Res.*, **28**, 260–262. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 58–60.
10. Gracy,J. and Argos,P. (1998) Automated protein sequence database classification. I. Integration of compositional similarity search, local similarity search and multiple sequence alignment. *Bioinformatics*, **14**, 164–173.
11. Gracy,J. and Argos,P. (1998) Automated protein sequence database classification. II. Delineation of domain boundaries from sequence similarities. *Bioinformatics*, **14**, 174–187.
12. Corpet,F., Servant,F., Gouzy,J. and Kahn,D. (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.*, **28**, 267–269.
13. Yona,G., Linial,N. and Linial,M. (2000) ProtoMap: automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Res.*, **28**, 49–55.
14. Silverstein,K.A.T., Shoop,E., Johnson,J.E. and Retzel,E.F. (2000) MetaFam: a unified classification of protein families. I. Overview and statistics. *Bioinformatics*, in press.
15. Shoop,E., Silverstein,K.A.T., Johnson,J.E. and Retzel,E.F. (2000) MetaFam: a unified classification of protein families. II. Schema and query capabilities. *Bioinformatics*, in press.
16. Silverstein,K.A.T., Kilian,A., Freeman,J.L., Johnson,J.E., Awad,I.A. and Retzel,E.F. (2000) PANAL: an integrated resource for Protein sequence ANALysis. *Bioinformatics*, in press.
17. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
18. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2000) GenBank. *Nucleic Acids Res.*, **28**, 15–18.
19. Barker,W.C., Garavelli,J.S., Huang,H., McGarvey,P.B., Orcutt,B.C., Srinivasarao,G.Y., Xiao,C., Yeh,L.-S.L., Ledley,R.S., Janda,J.F., Pfeiffer,F., Mewes,H.-W., Tsugita,A. and Wu,C. (2000) The Protein Information Resource (PIR). *Nucleic Acids Res.*, **28**, 41–44. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 29–32.
20. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
21. Schultz,J., Copley,R.R., Doerks,T., Ponting,C.P. and Bork,P. (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.*, **28**, 231–234.
22. Huang,H., Xiao,C. and Wu,C.H. (2000) ProClass protein family database. *Nucleic Acids Res.*, **28**, 273–276.
23. Orengo,C.A., Pearl,F.M.G., Bray,J.E., Todd,A.E., Martin,A.C., Conte,L.L. and Thornton,J.M. (1999) The CATH database provides insights into protein structure/function relationships. *Nucleic Acids Res.*, **27**, 275–279. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 223–227.
24. Hubbard,T.J.P., Ailey,B., Brenner,S.E., Murzin,A.G. and Chothia,C. (1999) SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, **27**, 254–256.
25. Holm,L. and Sander,C. (1999) Protein folds and families: sequence and structure alignments. *Nucleic Acids Res.*, **27**, 244–247.