Nucleic Acids Research, 2005, Vol. 33, Database issue ***D197–D200***
doi:10.1093/nar/gki067

FunShift: a database of function shift analysis on protein subfamilies

Saraswathi Abhiman and Erik L. L. Sonnhammer*

Center for Genomics and Bioinformatics, Karolinska Institutet, S-17177 Stockholm, Sweden

Received August 13, 2004; Revised and Accepted October 5, 2004

ABSTRACT

Members of a protein family normally have a general biochemical function in common, but frequently one or more subgroups have evolved a slightly different function, such as different substrate specificity. It is important to detect such function shifts for a more accurate functional annotation. The FunShift database described here is a compilation of function shift analysis performed between subfamilies in protein families. It consists of two main components: (i) subfamilies derived from protein domain families and (ii) pairwise subfamily comparisons analyzed for function shift. The present release, FunShift 12, was derived from Pfam 12 and consists of 151 934 subfamilies derived from 7300 families. We carried out function shift analysis by two complementary methods on families with up to 500 members. From a total of 179 210 subfamily pairs, 62 384 were predicted to be functionally shifted in 2881 families. Each subfamily pair is provided with a markup of probable functional specificity-determining sites. Tools for searching and exploring the data are provided to make this database a valuable resource for protein function annotation. Knowledge of these functionally important sites will be useful for experimental biologists performing functional mutation studies. FunShift is available at http://FunShift.cgb.ki.se.

INTRODUCTION

One of the fundamental goals of the genomic era is to extract information about the function of proteins from sequence data on a large scale. To this end, many databases have been developed that group homologous protein sequences into families, for example, Pfam (1), SMART (2), TIGRFAMs (3), PROSITE (4), BLOCKS (5), PRINTS (6) and InterPro (7). InterPro, Pfam and SMART are the most widely used among these databases.

The membership of a protein to a particular family generally indicates the broad function it may perform. If more detailed functional aspects are sought, it is often necessary to analyze the subfamily membership within that family (8).

A subfamily can be viewed as a set of proteins with related functions and domain organizations resulting from a particular line of evolution within a family. With the rapid growth of the sequence databases, the number of sequences belonging to a particular protein family is increasing sharply. As a consequence, it is becoming necessary to analyze the relationships between the numerous members of a protein family by categorizing them into subfamilies. Even though efforts have been made in this direction, they have only been applied to a handful of families (8–10). PANTHER is an exception, but is not freely available to the scientific community (11).

Many protein families have evolved to accommodate a wide range of functions, with each subfamily performing a specific function even though the general function may be the same for all the subfamilies. Hence it is necessary to identify subfamilies in protein families and analyze them for function shifts to enable better functional annotation of protein sequences.

Conservation patterns in protein multiple sequence alignments can be used to analyze the evolutionary constraints operating on different subfamilies. We use here two kinds of sites to predict function shift between subfamilies. These are conservation shifting sites (CSS), which are conserved in two subfamilies but using different amino acid residues, and rate shifting sites (RSS), which have different evolutionary rates in two subfamilies.

Here, we present a new database called FunShift that provides subfamily classifications and function shift analysis of the subfamilies derived from full alignments of the Pfam database.

GENERATION AND STATISTICS OF THE DATABASE

Subfamily generation

The division of a protein family into subfamilies is often performed by inspecting the phylogenetic tree of the family and deciding the subfamily membership of proteins. However,

*To whom correspondence should be addressed. Tel: +46 8 524 863 95; Fax +46 8 337 983; Email: Erik.Sonnhammer@cgb.ki.se

there are no clear criteria for dividing the tree into subfamilies, and it would also be time consuming for large-scale analysis. Sjolander (10,12) developed a method called BETE, which uses total relative entropy (TRE), the average relative entropy of all the columns in an alignment between two subfamilies. In this method, a neighbor-joining tree is constructed using TRE as distance measure. The subfamilies are defined using an encoding cost function that strives to minimize the number of subfamilies at the same time as it maximizes the sequence homogeneity within each subfamily. This method is completely automatic and hence can be used for large-scale analysis.

Subfamilies for the Pfam families were generated using the BETE method. The size and sequence diversity of the subfamilies thus generated is similar to the PANTHER database (11), where expert curators divided the subfamilies after inspecting the phylogenetic tree of each family manually. Function shift between subfamilies was predicted by identifying two kinds of sites, namely CSS and RSS.

Conservation shifting sites

Positions conserved in all members of a family are considered to be important for maintaining the structural scaffold or the core function. However, some positions may be conserved in different subfamilies but using different amino acids. Such positions are likely to be responsible for subfamily-specific functions. It is probable that these subfamilies have slight changes in function, such as different substrate specificities. Positions that exhibit such subfamily-specific conservation patterns are termed as CSS and can thus be used as indicators of function shift. CSS between the subfamilies were identified using the method developed by us (S. Abhiman and E. L. L. Sonnhammer, submitted for publication), which is similar to the method of Sjolander (10). Essentially, the amino acid distribution at each position in an alignment is computed and used to calculate the relative entropy between two subfamily alignments. The cumulative relative entropy is then converted into a *Z*-score, which is a normalized measure of conservation dissimilarity between two subfamilies.

Rate shifting sites

Sites in a protein evolve at different rates, with some functionally constrained sites evolving slowly and some others evolving faster. Some sites also evolve at different rates in different subfamilies of a family. Sites with such shifts in evolutionary rates between two subfamilies are referred to as RSS. Detecting a large number of such positions between two subfamilies suggests that the function has diverged between them. RSS between subfamilies in a family were determined using the LRT method (13). Each position in the alignment is analyzed individually and the program generates *U*-values that specify the likelihood that there is a rate change for each alignment position between the subfamilies under consideration.

Prediction of functionally divergent subfamily comparisons

In each family, the subfamily pairs were compared all-against-all for CSS and RSS. Subfamilies that had at

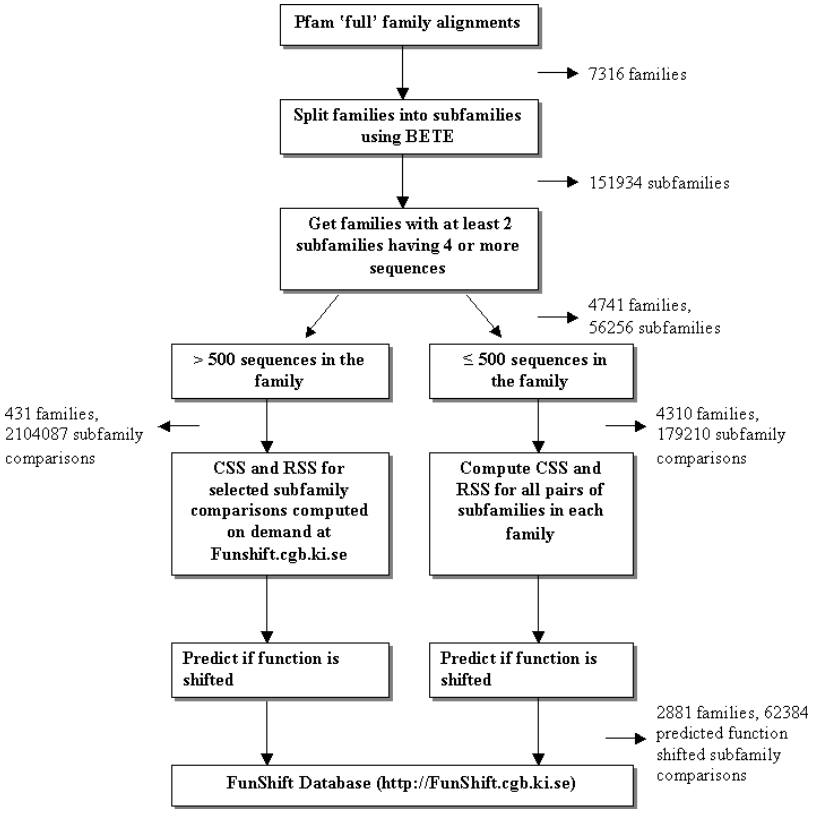

Figure 1. Schematic representation describing the process of generating the FunShift database.

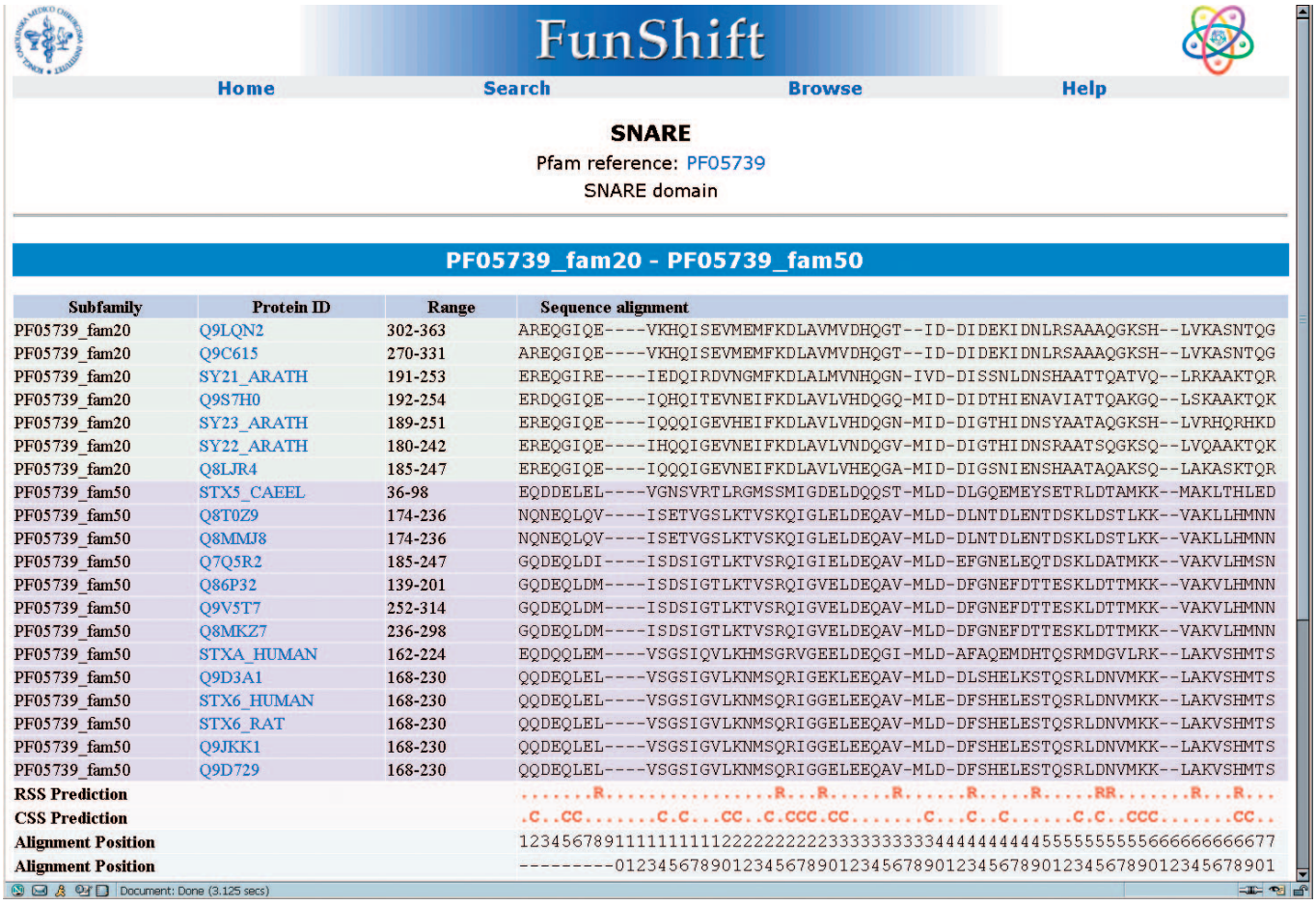

Figure 2. Example of a subfamily comparison from the FunShift database. The Screenshot shows the markup of RSS ('R' symbol) and CSS ('C' symbol) for a subfamily pair from the SNARE domain family (Pfam: PF05739).

least four sequences were only considered for this analysis. A function shift between a subfamily pair was predicted by using the percentage of CSS and RSS as variables in classification functions. These classification functions were derived from a previous analysis of functionally divergent subfamilies derived from enzyme families (S. Abhiman and E. L. L. Sonnhammer, submitted for publication).

RESULTS

The primary data were derived from the Pfam database (Version 12.0) of protein domain families and alignments. A total of 7300 'full' alignments from Pfam, with a maximum of 10 000 sequences were divided into subfamilies. This resulted in 151 934 subfamilies, of which 58 696 subfamilies had four or more sequences. Since it is computationally intensive to consider all subfamily pairs (2 283 297), we only precomputed RSS and CSS for families up to 500 sequences (4310 families; 179 210 subfamily pairs). Large families can be computed on demand on the website. The calculations on $\leqslant$500 sequence families predicted that 62 384 subfamily pairs (35%) in 2881 families are functionally shifted. The general scheme for the generation of the database is shown in Figure 1.

FEATURES OF THE DATABASE

Subfamily alignments and phylogenetic trees

Each Pfam family has a link to the subfamily alignments and the corresponding phylogenetic tree defining the subfamilies, generated by BETE. The subfamily alignments are provided in the standard FASTA format as well as in the Stockholm format, used by Pfam.

Comparison of subfamily pairs for function shift

Each subfamily pair within a family was compared to identify RSS and CSS. The positions were marked up as RSS or CSS when the *U*-values and *Z*-scores exceeded the cutoffs 4.0 and 0.5, respectively (see above) (Figure 2). The criteria for defining these cutoffs have been described in detail elsewhere (S. Abhiman and E. L. L. Sonnhammer, submitted for publication). The subfamily alignments along with predictions of function shift and RSS/CSS markup are available for browsing and download at the FunShift web server.

ACCESS TO THE DATABASE

FunShift is available via the World Wide Web (http://FunShift.cgb.ki.se). The data are stored in easy-to-access

flat files and can be downloaded. The web interface has a user-friendly navigation system to explore the information and provides basic text search tools for searching by keywords, family name and protein name. Methods for displaying selected families, subfamilies, comparisons and function shift analysis were built in Perl, and implemented in a Unix environment.

DISCUSSION

The FunShift database of protein subfamilies annotated with predicted CSS and RSS, and functionally distinct subfamilies are intended as a resource for the functional genomics and evolution research communities. This dataset may be used for a number of studies such as investigating the distribution of CSS and RSS residues on the three-dimensional structure of the proteins, identifying function subtypes and testing of functional divergence principles. Many of these studies have only been carried out on single protein families and will be of more general value when using the FunShift database. Furthermore, the CSS and RSS can be used as primary candidates for site-directed mutagenesis in function elucidation of proteins from laboratory experiments. The database will be periodically updated and will follow the Pfam version numbers. Additional methods for predicting function shift between subfamilies of a protein family are being investigated and will be incorporated into the database in future.

ACKNOWLEDGEMENTS

We thank Bjarne Knudsen for providing the Rate shift program, Kimmen Sjolander for providing the BETE program and for helpful discussions. We thank David A. Liberles for suggestions about our research, Markus Wistrand and other members of Sonnhammer's group for discussions. This work was supported by the Pfizer Corporation and the Swedish Knowledge Foundation.

REFERENCES

1. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al*. (2004) The Pfam protein families database. *Nucleic Acids Res*., **32**, D138–D141.
2. Letunic,I., Copley,R.R., Schmidt,S., Ciccarelli,F.D., Doerks,T., Schultz,J., Ponting,C.P. and Bork,P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res*., **32**, D142–D144.
3. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res*., **31**, 371–373.
4. Hulo,N., Sigrist,C.J., Le Saux,V., Langendijk-Genevaux,P.S., Bordoli,L., Gattiker,A., De Castro,E., Bucher,P. and Bairoch,A. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res*., **32**, D134–D137.
5. 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.
6. Attwood,T.K. (2002) The PRINTS database: a resource for identification of protein families. *Brief Bioinformatics*, **3**, 252–263.
7. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al*. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*., **31**, 315–318.
8. Hannenhalli,S.S. and Russell,R.B. (2000) Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol*., **303**, 61–76.
9. Gaucher,E.A., Miyamoto,M.M. and Benner,S.A. (2001) Function-structure analysis of proteins using covarion-based evolutionary approaches: elongation factors. *Proc. Natl Acad. Sci. USA*, **98**, 548–552.
10. Sjolander,K. (1998) Phylogenetic inference in protein superfamilies: analysis of SH2 domains. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **6**, 165–174.
11. Thomas,P.D., Campbell,M.J., Kejariwal,A., Mi,H., Karlak,B., Daverman,R., Diemer,K., Muruganujan,A. and Narechania,A. (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res*., **13**, 2129–2141.
12. Sjolander,K. (1997) Bayesian Evolutionary Tree Estimation. In *Proceedings of the Eleventh International Conference on Mathematical and Computer Modelling and Scientific Computing, Computational Biology Session: Conference Computing in the Genome Era 1997*, Georgetown University Conference Center, Washington DC, March 31–April 3.
13. Knudsen,B. and Miyamoto,M.M. (2001) A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proc. Natl Acad. Sci. USA*, **98**, 14512–14517.