Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

The SUPERFAMILY database in structural genomics

Julian Gough

MRC Laboratory of Molecular Biology, Hills Road, cambridge CB2 2QH, England

Correspondence e-mail: gough@supfam.org

Received 31 January 2002
Accepted 21 August 2002

The SUPERFAMILY hidden Markov model library representing all proteins of known structure predicts the domain architecture of protein sequences and classifies them at the SCOP superfamily level. This analysis has been carried out on all completely sequenced genomes. The ways in which the database can be useful to crystallographers is discussed, in particular with a view to high-throughput structure determination. The application of the SUPERFAMILY database to different target-selection strategies is suggested: novel folds, novel domain combinations and targeted attacks on genomes. Use of the database for more general inquiry in the context of structural studies is also explained. The database provides evolutionary relationships between target proteins and other proteins of known structure through the SCOP database, genome assignments and multiple sequence alignments.

1. Introduction

As a result of the increase in the rate of experimental determination of DNA sequences and the consequent success of the genome-sequencing projects, there are now (publicly available) over 60 completely sequenced genomes spanning all kingdoms of life. More recently, there have arisen structural genomics projects (Smith, 2000), which are expected to begin reaching their production phases during 2002–2003. These projects will accelerate the production of new protein structures and hence significantly increase the available structural data. Because of the difficulty and cost of solving three-dimensional structures, it is not possible to attempt to solve the structure of every protein. A more targeted approach than that used by the sequencing projects is needed to maximize the return on the projects' efforts. Using experimental and computational tools, targets can be selected which are expected to be in some way novel. Some of the ways in which the SUPERFAMILY (Gough & Chothia, 2002) database can contribute to targeting strategies are discussed here.

Most targeting strategies aim to achieve as complete as possible coverage of something, for example a proteome or a functional pathway. There are a limited number of common structural folds (Chothia, 1992) and some targeting strategies of structural genomics projects will significantly increase the proportion of this limited number for which we have a structural representative. The improved completeness of fold coverage will change the current view of protein-structure space.

2. SUPERFAMILY

The SUPERFAMILY database (Gough & Chothia, 2002; Gough *et al.*, 2001) is a library of hidden Markov models

(HMMs; Eddy, 1996; Hughey & Krogh, 1996; Krogh *et al.*, 1994) of domains of known structure created using the SAM (Karplus *et al.*, 1998) software package. Services and data are available at http://supfam.org.

2.1. What it does

The purpose of SUPERFAMILY is to detect and classify in protein sequences evolutionary domains for which there is a known structural representative. Given a protein about which nothing is known other than the amino-acid sequence, the object is to assign known structural domains or more specifically domains at the SCOP (Murzin *et al.*, 1995) superfamily level.

2.1.1. SCOP. The SCOP database classifies all proteins in the PDB (Berman *et al.*, 2000) into domains which are hierarchically organized at levels of similarity. Small proteins consist of a single domain, whereas medium-sized proteins may consist of one or more domains. Large proteins consist of multiple domains. A domain is defined as the minimum evolutionary unit, so a protein will only have parts classified into separate domains if those parts are observed independently in nature either on their own or in combination with other domains.

Structural, functional and sequence information is used to group together in the hierarchy at the superfamily level domains for which there is evidence for a common evolutionary ancestor. Domains belonging to the same superfamily have very similar structure and hence usually the same or a related function.

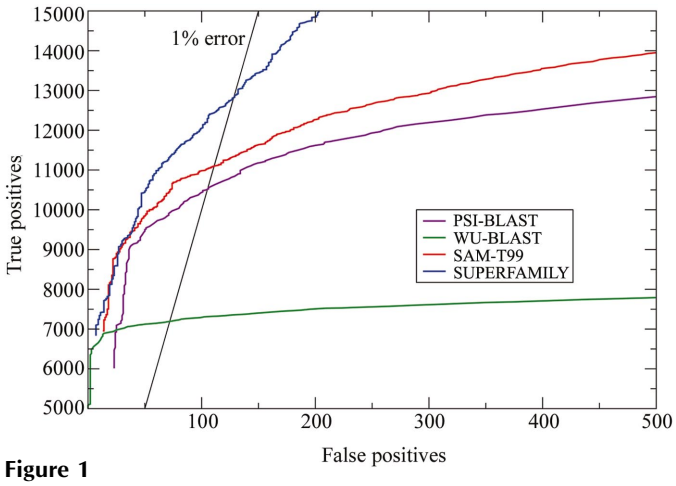

Figure 1
The number of true and false positives found by four different methods when scoring all sequences of structures in SCOP filtered to 40% sequence identity.

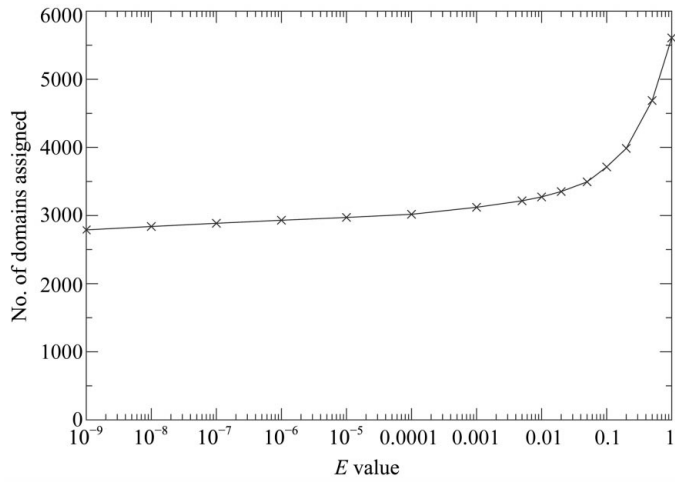

Figure 2
The number of domains assigned by SUPERFAMILY to the *Escherichia coli* genome at different error rates (using different *E*-value thresholds). In this case, the *E* values are calculated such that at 0.01 the expected error rate is 1%. The curve shows that in the critical region around the 1% error rate the number of domains assigned is not very sensitive to the threshold.

2.2. How it does it

SUPERFAMILY uses a library of HMMs representing all superfamilies in SCOP. HMMs are sequence profiles very similar to PSI-BLAST (Altschul *et al.*, 1997) profiles. Profiles are built from multiple sequence alignments and represent a group of sequences (in this case a superfamily) rather than a single sequence. The profile should embody the features that characterize the superfamily which it is supposed to represent. By comparing a sequence to a profile, far more distant relationships can be detected than by comparing two sequences, which is what pairwise methods such as BLAST (Altschul *et al.*, 1990) and FASTA (Pearson & Lipman, 1988) do.

2.3. Why use it?

To compare the ability of different methods to detect and classify domains in SCOP (Murzin *et al.*, 1995), a test was carried out, the results of which are shown in Fig. 1. The test comprises of an all-against-all search of sequences of structures in SCOP filtered to 40% sequence identity (Brenner *et*

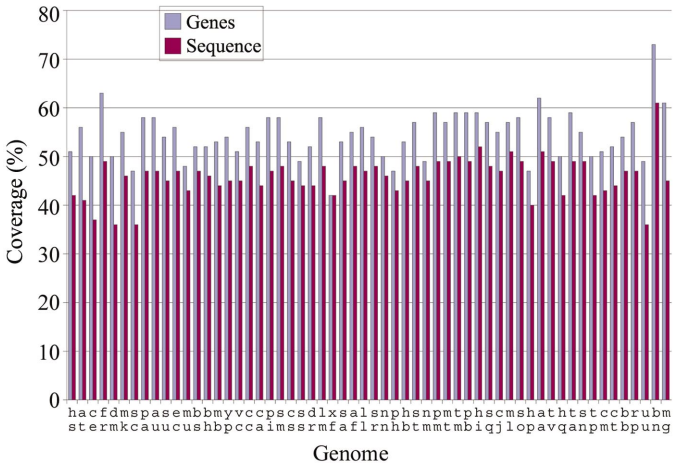

Figure 3
The SUPERFAMILY coverage of 58 complete genomes. The coverage of genes is the percentage of genes with at least one domain assignment; the coverage of sequence is the percentage of amino acids in all genes covered by domain assignments. The two-letter codes for the genomes shown in this graph are those defined by the SUPERFAMILY database. The full names are available at http://supfam.org/SUPERFAMILY/cgi-bin/genome_names.cgi.

al., 2000). The SCOP classification is then used to decide whether the relationships detected are true or false.

It is clear from Fig. 1 that profile methods (*SAM-T99*, *PSI-BLAST*) perform far better in this test of remote homology detection than pairwise methods (*WU-BLAST*). SUPERFAMILY is based upon *SAM-T99* (Karplus *et al.*, 1998), but improves upon and adds to it, specifically aiming at SCOP superfamily classification. For more information, please refer to Gough *et al.* (2001).

2.3.1. Accuracy and reliability. The *E*-value scores provided by the *SAM* software which are used for the assignments provide a theoretical value for the expected error rate. Large-scale tests show that these theoretical expectations are very close to the observed error rates (Gough *et al.*, 2001). Close examination of hundreds of assignments where the results are checked (Bujnicki *et al.*, 2001) indicate a 1% error rate.

Examination of the assignments made by SUPERFAMILY at different error rates (*E*-value threshold) in Fig. 2 shows that in the critical region the number of assignments are not highly sensitive to the threshold chosen.

3. Applications to structural genomics

The HMM library was designed for genome analysis leading to evolutionary studies (Apic *et al.*, 2001*a*; Teichmann *et al.*, 2001), but also has applications in structural genomics.

3.1. Target selection

The aim of structural genomics projects is to solve new structures which give us a more complete view of the world of protein structure. There are several different views of the structural world, which lead to alternative approaches to the selection of target proteins for the projects for structure determination. Some approaches which may be aided by the use of the SUPERFAMILY database are described here.

3.1.1. Novel folds. There are a limited number of common protein folds in nature (Chothia, 1992), many of which have at least one representative structure solved. There are, however, many folds which have not yet been determined and it is the aim of some projects to determine the structure of proteins with novel folds, with a view to generating complete coverage of fold space. Any target sets which have been designed to find novel folds can be scanned against the HMM library to see if they might belong to a superfamily for which there is already a structural representative. The effect of this is to remove distant homologues to known folds, but since a negative result is inconclusive, the aim is not to identify novel folds directly.

Of the new structures which are deposited in the PDB every week, half of those with no sequence homology to any other PDB sequence using *BLAST* (E value > 0.1), can be assigned by the HMM library to a superfamily with a known structural representative. Of those which cannot be assigned, very roughly half have novel folds. This test was independently carried out by *LiveBench* (Bujnicki *et al.*, 2001).

3.1.2. Domain combinations. The HMM library has been used to assign structural domains to sequences in all of the completely sequenced genomes (see Fig. 3). Assignments currently cover approximately 40% of the amino acids in eukaryote genomes (50% of the sequences) and 45% of the prokaryote genomes (55% of the sequences). Genome sequences may have had some of their domains assigned but not all, hence the discrepancy between the coverage of amino acids and number of sequences.

In the PDB there are structures of proteins with different combinations of domains. By examining the assignments of domains to genome sequences, it is possible to observe protein sequences with combinations of domains for which a structure has not yet been solved (Apic *et al.*, 2001*a*,*b*). These novel combinations provide targets for structural genomics projects and are available at http://www.mrc-lmb.cam.ac.uk/genomes/DomCombs. Although the novel aspect of these targets is not the fold but the combination of folds, these targets offer certainty in this aspect. It is not possible (using any method) to find targets which are certain to have a novel fold.

3.1.3. Targeted attacks. The aim of some projects is to achieve maximum structural coverage of a particular genome, *e.g. Mycobacterium tuberculosis* (http://www.doe-mbi.ucla.edu/TB/pubs.php). As explained above, approximately half of all genome sequences are assigned to a superfamily for which there is a known structural representative. The genome analysis provides a good starting point for target selection on any project targeted at a particular genome.

3.2. General inquiry

As well as target selection, the model library can be used more generally to obtain information about a protein which could be relevant to structural studies.

3.2.1. Evolutionary relationships. If the structure of a protein is known or if there is a similarity to a protein of known structure, the SCOP database provides the classification at the superfamily level which links the domains of proteins to others with a common ancestor. As mentioned before, if no relationship to a structure is known the model library may be able to detect one. SCOP also subgroups superfamily domains into families which are more closely related and usually have the same function.

3.2.2. Genome occurrence. Once the superfamily is known, it is possible through the genome analysis to see the occurrence of the members in the different genomes. For a given superfamily, all of the members in every genome that have been assigned by SUPERFAMILY are listed. The distribution across genomes may reveal interesting features such as particular species which are missing the superfamily in question or which for some reason have a much greater number of members than others. The distribution across the different kingdoms of life may also be of interest.

3.2.3. Sequence alignments. One of the services provided on the World Wide Web is multiple sequence alignment. There are alignments of PDB sequences belonging to the same superfamily and it is possible for the user to add their own

sequences to the alignment. All of the genome sequences from the different organisms which have been assigned to the same superfamily can also be aligned. Multiple sequence alignments of homologous proteins reveal patterns of evolutionary conservation which represent the structural and functional constraints on the protein. There is an automatic statistical analysis of the multiple alignments designed to detect features and aid such analysis.

Thanks to Cyrus Chothia for discussions and to the Medical Research Council for funding.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.* **215**, 403–410.
Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). *Nucleic Acids Res.* **25**, 3389–3402.
Apic, G., Gough, J. & Teichmann, S. A. (2001*a*). *J. Mol. Biol.* **310**, 311–325.
Apic, G., Gough, J. & Teichmann, S. A. (2001*b*). *Bioinformatics*, **17**, Suppl. **1**, S83–S89.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Brenner, S. E., Koehl, P. & Levitt, M. (2000). *Nucleic Acids Res.* **28**, 254–256.
Bujnicki, J. M., Elofsson, A., Fischer, D. & Rychlewski, L. (2001). *Bioinformatics*, **17**, 750–751.
Chothia, C. (1992). *Nature (London)*, **357**, 543–544.
Eddy, S. R. (1996). *Curr. Opin. Struct. Biol.* **6**, 361–365.
Gough, J. & Chothia, C. (2002). *Nucleic Acids Res.* **30**, 268–272.
Gough, J., Karplus, K., Hughey, R. & Chothia, C. (2001). *J. Mol. Biol.* **313**, 903–919.
Hughey, R. & Krogh, A. (1996). *Comput. Appl. Biosci.* **12**, 95–107.
Karplus, K., Barrett, C. & Hughey, R. (1998). *Bioinformatics*, **14**, 846–856.
Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. (1994). *J. Mol. Biol.* **235**, 1501–1531.
Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *J. Mol. Biol.* **247**, 536–540.
Pearson, W. R. & Lipman, D. J. (1988). *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
Smith, T. (2000). *Nature Struct. Biol.* **7**, 927.
Teichmann, S. A., Rison, S. C., Thornton, J. M., Riley, M., Gough, J. & Chothia, C. (2001). *J. Mol. Biol.* **311**, 693–708.