Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W249–W251***
doi:10.1093/nar/gki363

Fragnostic: walking through protein structure space

Iddo Friedberg* and Adam Godzik

The Burnham Institute, 10901 N. Torrey Pines Road, La Jolla, CA 92037, USA

Received February 8, 2005; Revised and Accepted February 23, 2005

ABSTRACT

The Fragnostic (http://ffas.burnham.org/Fragnostic) web tool implements a novel and useful view of protein structure space. We mined a non-redundant subset of the PDB for common fragments shared between proteins inhabiting different SCOP folds. Subsequently, we formulated an inter-fold similarity measure based on fragment sharing. Fold space is described as a graph whose nodes are folds between which the edges are drawn depending on the extent of fragment sharing. In this fashion, Fragnostic helps discover meaningful relationships between proteins belonging to different folds, based on sharing similar fragments in the proteins comprising those folds. Distant fold similarity information is supplemented by annotations taken from Gene Ontology, SCOP and CATH. Overall, Fragnostic is a tool which helps discover structural and functional relationships between proteins which are distantly related or seemingly unrelated.

BACKGROUND

The two popular protein classification schemes, CATH (1) and SCOP (2), partition the protein structure universe hierarchically, proceeding from coarse-grained to fine-grained partitions. The initial, coarse-grained partitioning of structure space is based on the secondary structure content. Because there are two well-ordered secondary structure elements, we have four possible classes as the topmost partitioning rank in those databases (SCOP and CATH actually use a few more, *ad hoc* classes). Classes are then more finely partitioned into folds (SCOP) or topologies (CATH), based on manual assignment. There may be between 100 and 200 folds per class. We know from experience that many proteins which are assigned to different folds share a structural/functional similarity. When proteins are categorically assigned to different folds, we lose important information about possible similarities between individual proteins assigned to different folds. Furthermore, because fold assignment is manual and sometimes arbitrary, there are cases where a fold–fold similarity between proteins inhabiting two different folds is glaringly obvious. These anomalies arise from the categorical assignment of proteins in a hierarchical classification scheme. We named the gap between the few classes and the many folds the 'granularity gap'. This granularity gap acts as a barrier preventing us from seeing obvious and not-so-obvious similarities between proteins from different folds, as was elaborated upon in studies conducted by Harrison *et al*. (3) and Choi *et al*. (4).

BRIDGING THE GRANULARITY GAP

One way of bridging the granularity gap is to re-establish the relationships between fold populations using similarities in a sub-domain level. We have chosen to address this problem using short fragments shared between populations of proteins in different folds. In another place (I. Friedberg and A. Godzik, submitted for publication) we describe in detail the generation and analysis of a fragment dataset. Briefly, we used a non-redundant set of solved structures, PDB-SELECT25 (5), to generate a dataset of 2.5×10^7 fragments of lengths 5, 10, 15 and 20 residues. Fragments were generated using a sliding window along each protein's sequence. Those fragments were aligned using FFAS03 (6), a sensitive profile–profile alignment program. The high scoring profile-based alignments were then screened by aligning them structurally, and only the alignments with a C-α RMSD of $\leqslant$1 Å were retained. After this two-step screening process, we had a dataset of 1.25×10^5 fragment pairs. The fragments were derived without any assumptions regarding their secondary structure content, an 'agnostic' approach; hence, 'Fragnostic'. We proceeded to implement a distance measure between folds, fragment based fold similarity (FBFS), based on fragment sharing.

- Given n folds, indexed $(1, \ldots, n)$.
- Each fold will have a set of fragments shared with other folds: $(X_1, X_2, \ldots, X_n)$.
- X_i being the set of all fragment pairs which are shared in fold i. $|X_i|$ is the number of those pairs.
- $X_{i,j}$ is the set of all fragment pairs shared between fold i and fold j and $|X_{i,j}|$ is a number of such pairs.

*To whom correspondence should be addressed. Tel: +1 858 646 3100; Fax: +1 858 713 9925; Email: idoerg@burnham.org

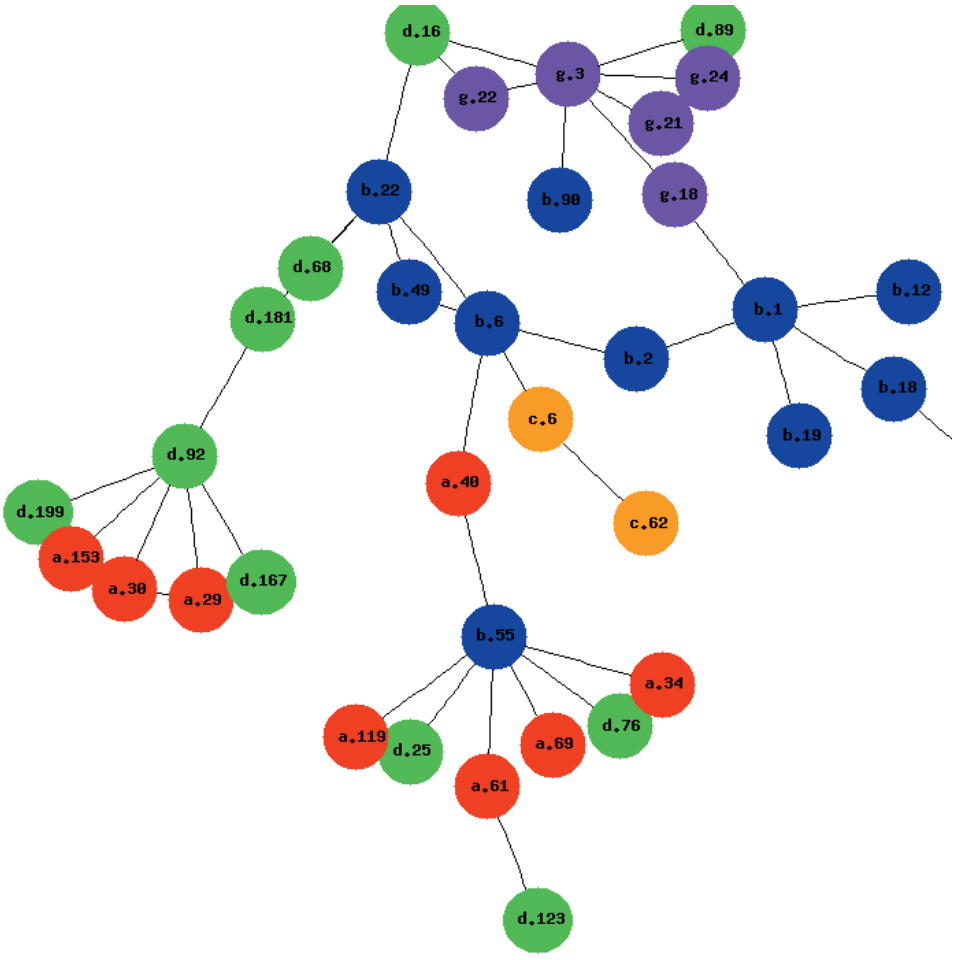

Figure 1. Part of a Fragnostic graph for fragment length 10, FBFS threshold of 0.2 and number of fragment threshold of 1. Circles are the SCOP fold populations, color coded according to SCOP class. Red, all alpha; blue, all beta; orange, alpha/beta; green, alpha + beta; and purple, small.

- FBFS is then defined as follows:

$$\mathrm{FBFS}(i,j) = \begin{cases} i \neq j: \max\left[\dfrac{|Xi,j|}{|Xi|}, \dfrac{|Xi,j|}{|Xj|}\right] \\ i = j: 1 \end{cases} \quad \mathbf{1}$$

Having FBFS as a distance measure, we generated four weighted graphs, using fragment lengths of 5, 10, 15 and 20 residues. Each vertex represents a population of PDB-SELECT25 proteins in a given fold. Two vertices may be connected by a weighted edge, with the weight determined by the FBFS score.

IMPLEMENTATION

The Fragnostic web tool lets the user examine the relationship between fold populations, based on the graph representation outlined above. The user enters a fragment length, an FBFS threshold level and a number of shared fragments threshold level. The latter was entered to correct a positive bias which may exist in the case of folds with small populations. Fragnostic then generates a graph. Each vertex is shown as a circle, color-coded according to the SCOP class. The SCOP concise classification scheme code (SCCS) is shown in the vertex. SCCS is a four-position code assigned by SCOP to a family, with the first position (a letter) denoting the class, the second the fold, the third the superfamily and the fourth the family. Positions 2–4 of the SCCS are numbers, e.g a.4.3.23. As each vertex is composed of a population of proteins with a common fold, only the first two positions of the SCCS are shown (a.4). Placing the cursor over the vertex will show its fold's SCOP-assigned title. Two vertices are connected by an edge if the FBFS score between the two connected vertices is higher than the threshold provided by the user. Figure 1 shows a part of such a graph. Clicking on a vertex will display a table showing the SCOP domains from PDB-SELECT25 which belong to the vertex's fold. The table entry is linked to a 3D model of that domain, viewed using the Rasmol program (7). The model is displayed as a cartoon, and the regions which are covered by fragments shared with other folds are colored. Colors range from blue to red, the 'hotter' (redder in spectrum) the color, the more fragments are shared in that region with other folds (Figure 2). Using Rasmol—a simple yet powerful protein visualization tool—the user can further analyze the protein. The page is linked to the folds which are connected to the current one and to their connecting edges (see below). Clicking on an edge will produce a table of all the fragment

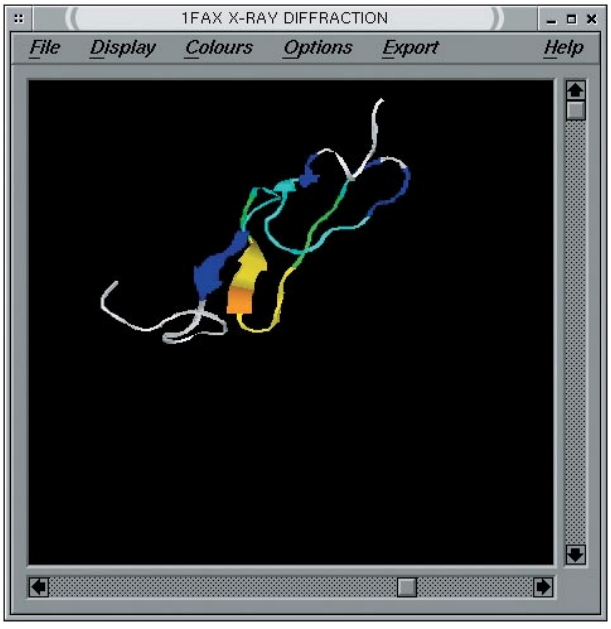

Figure 2. Coagulation factor X, light chain (PDB: 1FAX:L), which belongs to the knottins SCOP fold. The non-white areas are composed of length-10 fragments, shared with other folds.

alignments making up this edge. Whenever so annotated, a table entry will have Gene Ontology (GO) (8) terms associated with it, and/or Enzyme Commission (EC) classification number. The GO terms were taken from the PDB to GO mapping provided by The European Bioinformatics Institute (EBI). There may be multiple mappings between the chains and GO terms. This is because some protein chains have multiple functions, participate in more than one metabolic pathway, or are found in more than one cellular compartment. Care was taken, however, not to enter two GO terms when one clearly subsumes the other in an 'is-a' relationship. Thus if the term 'phosphodiesterase' appears associated with a given chain, 'esterase' will not be mentioned.

CONCLUSIONS

We present Fragnostic as novel method for walking through protein structure space. Rather than replacing SCOP with a new classification, it complements the existing classification by showing connections between known SCOP folds. Fragnostic is a powerful tool for revealing hidden inter-fold connections based on shared fragments. Fragnostic is also suitable for confirming hypotheses of structural or functional connections between proteins from different folds. In the future, we aim to permit querying using any SCOP entry, not only those in PDB-SELECT25. We are currently developing a fragment-to-structure matching method, so that the fragments—or rather a clustered library thereof—can be used as a structural motif library. Fragnostic was written using Zope (zope.org) for web content management and GraphViz (AT&T Laboratories) for displaying the graphs. The fragment dataset and associated information were generated using Biopython (biopython.org) and are maintained in a MySQL (MySQL AB) database.

ACKNOWLEDGEMENTS

We thank Drs Yuzhen Ye and Ana Rojas for providing invaluable advice for the design of the Fragnostic site. We thank Mr Bruce Worcester for his careful reading of this manuscript. This study was supported by NIH Grant P01GM63208-02. Funding to pay the Open Access publication charges for this article was provided by NIH Grant P01GM63208-02.

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. 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.
3. Harrison,A., Pearl,F., Mott,R., Thornton,J. and Orengo,C. (2002) Quantifying the similarities within fold space. *J. Mol. Biol.*, **323**, 909–926.
4. Choi,I.G., Kwon,J. and Kim,S.H. (2004) Local feature frequency profile: a method to measure structural similarity in proteins. *Proc. Natl Acad. Sci. USA*, **101**, 3797–3802.
5. Hobohm,U. and Sander,C. (1994) Enlarged representative set of protein structures. *Protein Sci.*, **3**, 522–524.
6. Rychlewski,L., Jaroszewski,L., Li,W. and Godzik,A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.*, **9**, 232–241.
7. Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.
8. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.