Functional diversity within protein superfamilies

James Casbon and Mansoor Saqi *

Bioinformatics Group, The Genome Centre, Barts and The London, Queen Mary's School of Medicine and Dentistry, Charterhouse Sq., London EC1 6BQ

Summary

Structural genomics projects are leading to the discovery of relationships between proteins that would not have been anticipated from consideration of sequence alone. However the assignment of function via structure remains difficult as some structures are compatible with a variety of functions. In this study we explore the relationships between structural diversity and functional diversity within distantly related members of SCOP superfamilies. We use the Gene Ontology functional classification scheme and Greens path entropy to measure functional diversity. We observe a negative correlation between the functional entropy of a superfamily and the size of the conserved core.

1 Introduction

A major obstacle to the exploitation of the huge amount of genome data now available is the lack of any functional annotation for many of the proteins. Typically some 30-40% of open reading frames cannot be assigned function on the basis of close sequence similarity to a protein of known function [15]. Such open reading frames are usually designated as 'hypothetical protein' or 'protein of unknown function'. Despite the development of powerful algorithms for the detection of remote sequence signals that facilitate probing below the so-called twilight zone of sequence similarity, functional annotation still remains a problem.

Many new structures are being solved through structural genomics [18, 26]. A recent analysis of solved target structures revealed that for 29% of domains, the 3-D structure revealed relationships not apparent from sequence [25]. This increasing amount of 3-dimensional information should impact significantly on our understanding of key properties that determine function and will aid the recognition of distant sequence relationships via structure. If a hypothetical protein shares structural similarity to that of a functionally characterised protein we might expect this to narrow down the possible functional roles of the protein and thereby aid in functional annotation.

Protein structure classification databases such as the Structural Classification of Proteins (SCOP) enable us to explore the characteristics of proteins that adopt the same global structures [16]. SCOP is a hierarchical categorisation in which a structural domain is classified according to class (secondary structure content), fold (broadly the spatial arrangement of the secondary structural elements), superfamily and family. Proteins at the superfamily level are believed to be related although this may not be apparent from consideration of sequences alone. Some folds are associated with a wide range of functions, while others seem less functionally versatile. For example, the TIM barrel fold contains 28 superfamilies and is associated with 4 of the

*Corresponding author, m.saqi@qmul.ac.uk

6 possible enzyme commission numbers [10]. Knowledge that a hypothetical protein adopts a functionally diverse fold such as a TIM barrel will not immediately narrow down the functional space. In contrast, the globins show a very specific repertoire of function despite large sequence diversity [3].

Here the relationship between structural and functional diversity at the level of protein superfamilies is explored. Many authors have developed pairwise measures of similarity for sequence [17, 23, 1], structure [11, 28] and function [13, 21]. The relationships between metrics have also been examined [7, 21]. The approach investigated in this paper is to examine, rather than pairwise similarity, the breadth of diversity of function among homologous groups of proteins. The analysis is restricted to superfamilies which show large numbers of sequences with less than ten percent sequence similarity, as these have the capacity for differing functions [27]. Two classification schemes are used: SCOP for structural classification and Gene Ontology (GO) for functional classification. A novel method for characterising the functional repertoire of a protein family is suggested, and the range of functional diversity exhibited under the metric is discussed. It is shown that the functional diversity of a superfamily shows some correlation with the numbers of proteins in the superfamily and the size of the conserved core of the superfamily.

2 Methods

2.1 Dataset

The SCOP database (version 1.67) is used as a classification of protein structure. The ASTRAL database is then used to select a non-redundant subset of SCOP domains showing no more than ten percent sequence identity with each other [5]. The dataset was formed by choosing from ASTRAL all superfamilies with more than ten members at this level of sequence diversity. This dataset contains 1260 domains distributed in 58 superfamilies. It is envisaged that this dataset will grow considerably as high throughput experimental structure determination projects progress.

2.2 Structural Diversity

Two measures of structural diversity are examined: the average RMSD (root mean square deviation) between members of the superfamily and the size of the core conserved structure of the superfamily. In order to measure the average RMSD, all against all structural alignments for domains in a given superfamily are performed using the SAP program [24]. This program reports 3 RMSD scores: a weighted RMSD, an unweighted RMSD over a 'best' set of closely aligned atoms and an RMSD for the whole alignment. The weighted RMSD is used for the measure as this will not be overly affected by outliers.

To measure the size of the conserved core, the number of positions marked as core in a structure-based multiple alignment of the superfamily is divided by the average length of a domain in the superfamily. The multiple alignments were constructed as described in [6] requiring that no two domains shared more than 10% sequence identity as defined by ASTRAL. Core positions were defined as those positions where the gap content is less than twenty percent and the average

separation is less than three Angstroms. This measure is termed the core size — if most of the structure is conserved across all domains in the superfamily, this value will be close to 1.0.

2.3 Functional diversity

To measure functional diversity, diversity amongst GO terms is used. To obtain GO terms for domains in the dataset, the program InterProScan [29] was used to assign GO terms. This program scans a query sequence against all databases in InterPro [2]. Significant hits are reported, with a corresponding InterPro record. InterPro mappings to GO are used to convert the InterPro records into GO terms. We consider only the molecular function ontology in this study.

The GO terms for the superfamily are enumerated and Green's path entropy function is used to measure the functional diversity [9]. Unlike Shannon's entropy which has no method for incorporating knowledge of relationships between states, Green's path entropy allows for the entropy to be considered where the relationship between states is pre-determined, in this case by GO.

For a given leaf l in a tree specifying a unique path $P_l = u_0, ...u_n, l$, the path entropy is defined as:

$$H(l) = \sum_{u_i \in P_l} \log d(u_i)$$

Where $d(u)$ is the outward degree of node u. The entropy of a given tree is then the average of these path lengths, or given a weighting over the leaves $w(l_j)$, which we can assume $\sum_j w(l_j) = 1$, the leaf weighted tree entropy is the expected path length under the weighting, $\sum_j p(l_j)H(l_j)$. A slight modification of the path entropy function is introduced by weighting the entropic contribution of a decision according to the depth in the tree. For a leaf l the path entropy becomes:

$$H(l) = \sum_{u_i \in P_l} f(i) \log d(u_i)$$

The tree entropy remains the weighted sum of path entropies.

A possible weighting scheme is $f(i) = K^i$ where $K < 1$ is a constant. This gives a relative weighting of K to a decision at depth $n + 1$ compared to a decision at depth n. This function reduces to the normal classification entropy for $K = 1$. For an example of why this weighting is desirable, see figure 1.

In order to measure the functional diversity, for each domain we take the GO terms assigned above and all their parent terms and combine them to form a tree. GO is not strictly a tree structure, however, we induce a tree structure by only observing parent terms forming the shortest route to the root. This tree represents the functional range of the superfamily. To measure the functional diversity, the entropy of this tree is measured. We weight the leaves by the proportion of domains observed with that term (or terms), and calculate the entropy.

However, the annotation stage may, and usually does, return more than one GO term for each domain. This is not surprising given that GO contains three main ontologies for process, function and location. Nevertheless, it may even return more than one term in each of the main ontologies. Some of these can be explained by the fact that InterProScan may report several different levels of detail, e.g. a protein may be reported to be both a 'binding' and an 'ATP-

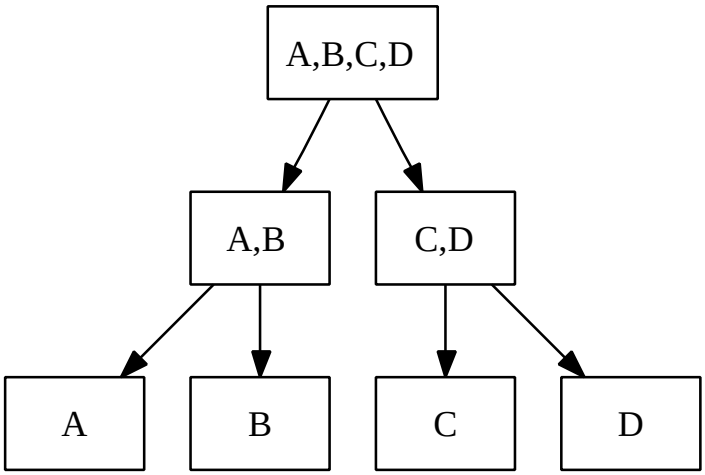

Figure 1: An example of the effect of depth weighting. Consider as an example the classification shown. Imagine we were to observe a set of objects X with labels $\{A, B\}$, and a set of objects Y with labels $\{A, C\}$. Taking X first, we prune the leaves C and D and the new leaf C, D and calculate the classification entropy. Clearly, there is a classification entropy of 1. Now looking at Y, we prune B and D and calculate the classification entropy, again it is 1. Both X and Y have a classification entropy of 1 despite the fact that X is clearly a more related set of examples lying in a more constrained subtree. In order to overcome this problem, we use the depth weighted classification entropy. For this example example, consider the depth weighted classification entropy with $K = 1/2$. Now, for X it is 1/2, whereas for Y it is 1. Thus the depth weighted entropy reflects the level in the classification at which the branching is made.

binding' protein. Therefore, a preliminary comparison of terms associated to each domain is made, and any terms which are parents of other terms are removed.

Even after removing related terms, a given protein would still have more than one annotation in an ontology. This is because gene ontology terms are usually only 'atomic', a functional description requires more than one term. For example, a globin may be described as both 'heme binding' and 'oxygen binding'. The functional description is the combination of both terms.

It is undesirable that two GO terms representing a single domain should contribute to the functional diversity of the superfamily. If the functional description is both terms, then ideally we would like there to exist a single node in the tree labelled with both terms where we can place the domain. A domain should exist at only one point in the hierarchy, else it will increase the functional diversity measure artificially. To achieve this we can use Greens tree product to create a tree that contains all combinations of terms across each level in GO (but still observing the hierarchy) by taking the product of the GO tree with itself [9]. This generates all possible combinations of terms. However, the product of a classification tree with itself is a special case of the product. When taking the product any nodes which represent the product of a term with itself become simple the term (i.e., term (A,A) becomes term A) and the terms are required to be ordered so that term (B,A) is the same as term (A,B). The product can be taken multiple times if more than two GO terms are given for a domain. Lastly, if any annotations belonging to a domain are parents of another term on the tree, a 'sink' leaf is introduced to represent the fact that the functional annotation is not specific enough in relation to other terms, since each domain must appear at the leaf. The result is a tree where a domain's annotations appear at one leaf only. An example of a such a tree showing the functional range of a superfamily can be seen in figure 2.

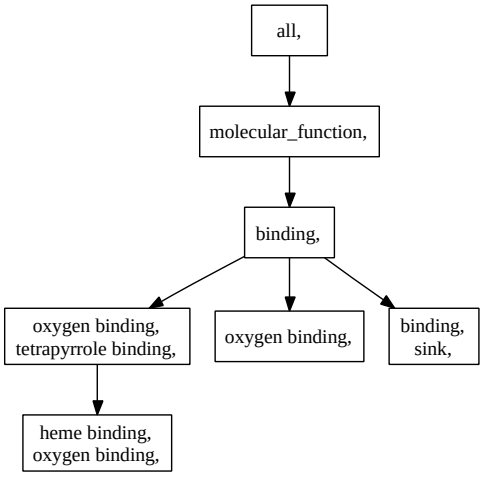

Figure 2: An example of a GO tree representing the "Globin-like" superfamily. A combination of functions can be observed in some domains, as well as a 'sink' leaf introduced for those domains that were only annotated as binding.

Superfamily	sunid	functional diversity
P-loop containing nucleoside triphosphate hydrolases	52540	4.33
NAD(P)-binding Rossmann-fold domains	51375	3.27
Nucleic acid-binding proteins	50249	2.70
Ribosomal protein S5 domain 2-like	54211	2.66
PLP-dependent transferases	53383	2.64
PH domain-like	50729	0.56
Thiamin diphosphate-binding fold (THDP-binding)	52518	0.42
RNA-binding domain, RBD	54928	0.42
Acyl-CoA N-acyltransferases (Nat)	55729	0.18
DEATH domain	47986	0.00

Table 1: Most and least diverse superfamilies in the dataset using a depth weighting of 0.75

3 Results

Across the superfamilies in the dataset there exists a large range of functional diversity. Figure 3 shows that the diversity ranges from highly conserved to very diverse. Most superfamilies however, are, by our measure, in the range 0.5-3.

In table 1 the five most functionally diverse and the five least functionally diverse SCOP superfamilies (out of the 58 superfamilies in our dataset) according to our metric are shown. The Enzyme Commission (EC) number presents a high level view of function where the first digit of EC number represents the class of the enzyme and can take 6 possible values [12]. From the top five, two superfamilies, the NAD(P)-binding Rossmann-fold domains and the P-loop containing nucleotide triphosphate hydrolases are associated with 4 different 1st digit EC numbers [8]. The PLP-dependent transferases are associated with 3 different 1st digit EC numbers and the Ribosomal protein S5 domain 2-like superfamily with a single 1st digit EC number. Functional diversity among the PLP-dependent transferases, a coenzyme binding domain (PLP, pyridoxal-phosphate is a cofactor) has been studied by Bray et al [4]. From the five least functionally diverse superfamilies, the Thiamin diphosphate-binding fold (THDP-binding) is

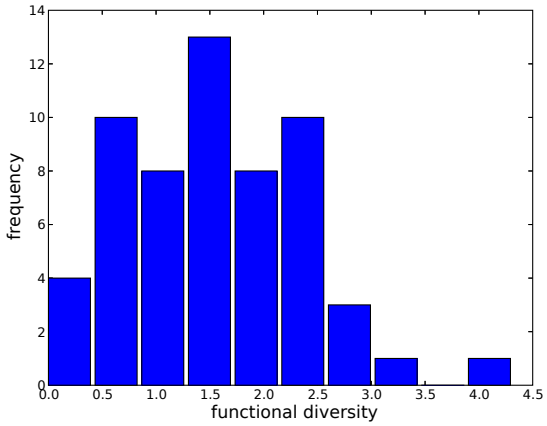

Figure 3: Histogram showing numbers of superfamilies with a given functional diversity (depth weighting of 0.75)

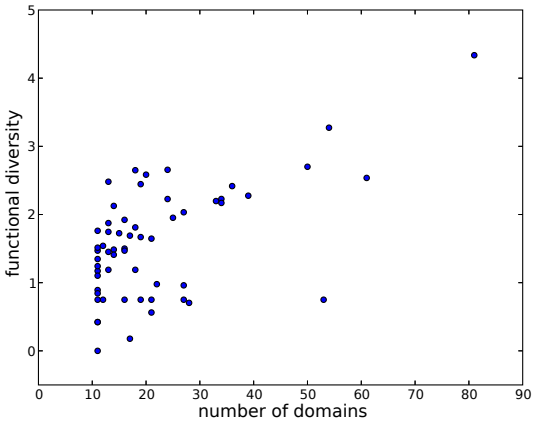

Figure 4: Scatter showing number of domains against functional diversity (depth weighting of 0.75)

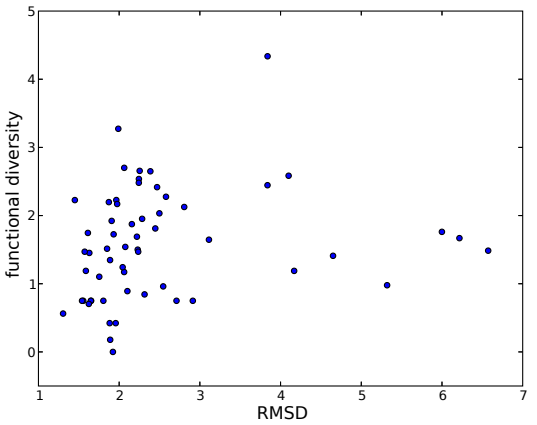

Figure 5: Scatter showing average RMSD of superfamily against functional diversity (depth weighting 0.75)

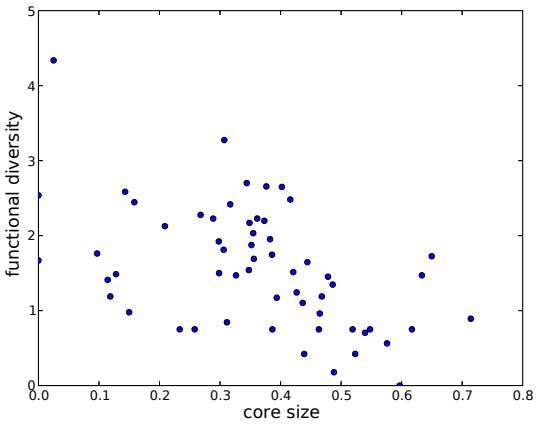

Figure 6: Scatter showing core size of superfamily against functional diversity (depth weighting of 0.75)

associated with 3 1st digit EC numbers and the Acyl-CoA N-acyltransferases (Nat) with 1.

The most functionally diverse superfamilies, namely the P-loop containing nucleoside triphosphate hydrolases, the NAD(P)-binding Rossmann-fold domains and the Nucleic acid-binding proteins are large sequence diverse families with 81, 54 and 50 domains in our dataset. The Ribosomal protein S5 domain 2-like superfamily and the PLP-dependent transferases superfamily have 24 and 18 domains in our dataset. Among the least functionally diverse superfamilies, the Thiamin diphosphate-binding fold (THDP-binding) superfamily, the RNA-binding domain, and the DEATH domain are considerably smaller each with 11 domains in our dataset. However, also having low functional diversity are the PH domain-like superfamily and the the Acyl-CA N-acyltransferases which have 21 and 17 domains in our dataset. We see that some of the most functionally diverse superfamilies are among the more populated superfamilies in our sequence diverse dataset and some of the least functionally diverse have far fewer domains.

Figure 4 shows a plot of superfamily size against the functional diversity of the superfamily. The plot shows a correlation between superfamily size and functional diversity, as confirmed by a correlation coefficient of 0.59. This trend was also observed by Shakhnovich et al. [19]. However, all but ten superfamilies have between ten and thirty domains. Amongst these domains, the correlation is much less pronounced, with a correlation coefficient of 0.18.

Figures 5 and 6 show plots of functional diversity against average RMSD and core size respectively. The figures show that, although there is no correlation between RMSD and functional diversity, there is a correlation between the core size and functional diversity. This correlation is weak and statistical in nature, but is confirmed by the correlation coefficient on -0.50. Moreover, this relationship remains for superfamilies with between ten and thirty domains (correlation coefficient -0.4).

4 Discussion

The function of a protein is not a trivial matter to describe. The level of abstraction best suited to functional annotation is not clear, and it may only be possible to define a function in a very specific context [22]. Despite this, in order to understand genomics data computationally, a

controlled ontology is a valuable tool. In order to proceed with the development of the method we have used the available data and tools (InterProScan) which are based on sequence similarity and other methods. We remain aware of the limitations of transfer of functional annotation by computational methods and that precise functional annotation is generally only revealed from experiment.

We have shown a method for describing and measuring the functional repertoire of a protein family based on a given ontology. Nevertheless, other measures of functional diversity could be adopted. A recent approach by Shakhnovich et al also calculates a functional flexibility score by averaging the Shannon entropy over each level of GO, but this definition does not fully account for the underlying hierarchy of function described by the gene ontology [20, 19]. A parwise distance for functional similarity within GO has been developed, but the distance measure cannot quantify the entropy of a set of GO labels [21]. Other groups have examined the lowest node in a hierarchy, whether GO or enzyme commission, that describes the function of a group of proteins [10, 13]. However, applying Greens path entropy to the Gene Ontology functional classification considers the hierarchy of function in calculating the entropy and also handles multiple functional labels through the tree product.

We remain aware that an automated functional annotation scheme could produce errors which would, in turn, affect the measure of functional diversity. It should also be noted that annotating function at the level of the domain may be problematic in the case of multi-domain proteins, or even multi function domains.

Our results have shown a correlation between functional diversity and the size of the conserved core of the superfamily. This core measure is arbitrary and, again, other measures could be used. For instance, another measure we examined was the average length of conserved residues across pairwise alignments. Although these results were not shown, they agreed with the results using the measure based on the multiple alignment as shown in the text. It does appear, however, that RMSD is not a useful measure when considering functional diversity. Perhaps this is because RMSD measures how conserved the core residues are, and not how many of the residues are in the core.

Global studies of the structure-function relationship are limited in their use when considering a particular example of a protein family. These results do, however, relate to previous studies investigating structure-function relationships. Matsuo and Bryant found that the size of the conserved core was greater between homologs than analogs, and that presence of this conserved core was discriminatory between the two (and that RMSD was not) [14]. We find that the size of the homologous core structure affects the functional diversity of homologous structures. It would be interesting to relate the functional specialisation with patterns of conservation in the core.

5 Acknowledgements

JC thanks the Special Trustees of the Royal London Hospital for funding.

References

[1] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, 1997.

[2] R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M.D. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N.J. Mulder, T.M. Oinn, M. Pagni, F. Servant, C.J. Sigrist, and E.M. Zdobnov. InterPro–an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics*, 16(12):1145–50, 2000.

[3] D. Bashford, C. Chothia, and A.M. Lesk. Determinants of a protein fold. Unique features of the globin amino acid sequences. *J Mol Biol*, 196(1):199–216, 1987.

[4] J.E. Bray, A.E. Todd, F.M. Pearl, J.M. Thornton, and C.A. Orengo. The CATH Dictionary of Homologous Superfamilies (DHS): a consensus approach for identifying distant structural homologues. *Protein Eng*, 13(3):153–65, 2000.

[5] S.E. Brenner, P. Koehl, and M. Levitt. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res*, 28(1):254–6, 2000.

[6] J. Casbon and M.A. Saqi. S4: structure-based sequence alignments of SCOP superfamilies. *Nucleic Acids Res*, 33 Database Issue:D219–22, 2005.

[7] C. Chothia and A.M. Lesk. The relation between the divergence of sequence and structure in proteins. *EMBO J*, 5(4):823–6, 1986.

[8] R.A. George, R.V. Spriggs, J.M. Thornton, B. Al-Lazikani, and M.B. Swindells. SCOPEC: a database of protein catalytic domains. *Bioinformatics*, 20 Suppl 1:I130–I136, 2004.

[9] C.D. Green. A Path Entropy Function for Rooted Trees. *Journal of the ACM*, 20(3), 1973.

[10] H. Hegyi and M. Gerstein. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol Biol*, 288(1):147–64, 1999.

[11] L. Holm and C. Sander. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res*, 25(1):231–4, 1997.

[12] IUBMB. *Enzyme Nomenclature: Reccomendations of the Nomencalture Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classficiation of Enzymes*. Academic Press, New York, 1992.

[13] P.W. Lord, R.D. Stevens, A. Brass, and C.A. Goble. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–83, 2003.

[14] Y. Matsuo and S.H. Bryant. Identification of homologous core structures. *Proteins*, 35(1):70–9, 1999.

[15] A. Muller, R.M. MacCallum, and M.J. Sternberg. Benchmarking PSI-BLAST in genome annotation. *J Mol Biol*, 293(5):1257–71, 1999.

[16] A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–40, 1995.

[17] S.B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.*, 48:443–453, 1970.

[18] A. Sali. 100,000 protein structures for the biologist. *Nat Struct Biol*, 5(12):1029–32, 1998.

[19] B.E. Shakhnovich, E. Deeds, C. Delisi, and E. Shakhnovich. Protein structure and evolutionary history determine sequence space topology. *Genome Res*, 15(3):385–92, 2005.

[20] B.E. Shakhnovich and J.M. Harvey. Quantifying structure-function uncertainty: a graph theoretical exploration into the origins and limitations of protein annotation. *J Mol Biol*, 337(4):933–49, 2004.

[21] B.E. Shaknovich. Improving the Precision of the Structure Function Relationship by Considering Phylogenetic Context. *PLOS Comp. Biol.*, 1(1), 2005.

[22] J. Shrager. The fiction of function. *Bioinformatics*, 19(15):1934–6, 2003.

[23] T.F. Smith, C. Burls, and M.S. Waterman. The statistical distribution of nucleic acid similarities. *Nucl Acids Res*, 13:645–656, 1985.

[24] W.R. Taylor. Protein structure comparison using iterated double dynamic programming. *Protein Sci*, 8(3):654–65, 1999.

[25] A.E. Todd, R.L. Marsden, J.M. Thornton, and C.A. Orengo. Progress of structural genomics initiatives: an analysis of solved target structures. *J Mol Biol*, 348(5):1235–60, 2005.

[26] D. Vitkup, E. Melamud, J. Moult, and C. Sander. Completeness in structural genomics. *Nat Struct Biol*, 8(6):559–66, 2001.

[27] C.A. Wilson, J. Kreychman, and M. Gerstein. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol*, 297(1):233–49, 2000.

[28] A.S. Yang and B. Honig. An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J Mol Biol*, 301(3):665–78, 2000.

[29] E.M. Zdobnov and R. Apweiler. InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, 17(9):847–8, 2001.