Quantifying Structure-Function Uncertainty: A Graph Theoretical Exploration Into the Origins and Limitations of Protein Annotation

Boris E Shakhnovich[1] **J. Max Harvey.**
[1]Bioinformatics Program, Boston University, Boston MA, 02215

Keywords: Protein Domain, Annotation, Graph Theory, Database, Structure-Function, Evolution.

Introduction.

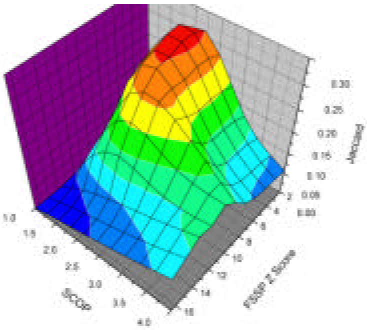
Since the advent of investigations into structural genomics, research has focused on correctly identifying domain boundaries, as well as domain similarities and differences in the context of their evolutionary relationships. As the science of structural genomics ramps up adding more and more information into the databanks, questions about the accuracy and completeness of our classification and annotation systems appear on the forefront of this research. A central question of paramount importance is how structural similarity relates to functional similarity. In this paper we begin to rigorously and quantitatively answer these questions by first exploring the consensus between the most common protein domain structure annotation databases CATH, SCOP and FSSP. Each of these databases explores the evolutionary relationships between protein domains using a combination of automatic and manual, structural and functional, continuous and discrete similarity measures. In order to thoroughly examine the issue of consensus, we build a generalized graph out of each of these databases and hierarchically cluster these graphs at interval thresholds. We then employ a distance measure to find regions of greatest overlap. Using this procedure we were able not only to enumerate the level of consensus between the different annotation systems, but also to define the graph-theoretical origins behind the annotation schema of Class, Family and Superfamily by observing that the same thresholds that define the best consensus regions between FSSP, SCOP and CATH correspond to distinct, non-random phase-transitions in the structure comparison graph itself. To investigate the correspondence in divergence between structure and function further, we introduce a measure of functional entropy that calculates divergence in function space. First, we use this measure to calculate the general correlation between structural homology and functional proximity. We extend this analysis further by quantitatively calculating the average amount of functional information gained from our understanding of structural distance and the corollary inherent uncertainty that represents the theoretical limit of our ability to infer function from structural similarity. Finally we show how our measure of functional "entropy" translates into a more intuitive concept of functional annotation into similarity EC classes.

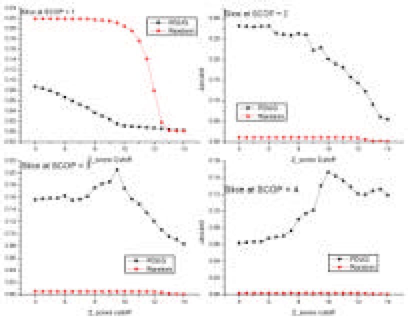
Databases as graphs

Through graph-morphing procedures for SCOP[1], CATH[2] and FSSP[3] we end up with three *weighted* graphs, one for each database. The nodes in each graph are the protein domains and the edges are the relationships defined by distances or proximity from each database. We proceed to cluster these graphs at regular interval cutoffs. For example, for FSSP we build a graph at each threshold from **Z**=2 to 16 with step .5. In order to do this, we pick a cutoff and keep all edges that are larger than this cutoff[4].

Compare graphs.

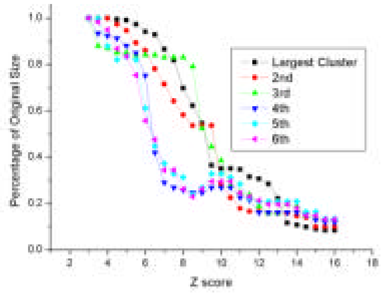
After TP, FP, TN and FN quantities have been defined, the distance measure between two graphs is merely a calculation of how many true positives the two graphs share with respect to false negatives and false positives. This measure is meant to calculate the level of agreement between the two graphs with respect to how many domain pairs they classify in the same cluster. Four Slices of the 3-D graph depicted in The cusps and maxima are easily discernable from these slices. At SCOP cutoff 1 Jaccard is actually smaller than the random control indicating that this level of annotation is probably not indicative of real evolutionary homology and may not indicate meaningful annotation. At SCOP Cutoff 2,3,4 the Jaccard distance between FSSP and SCOP is many thousands





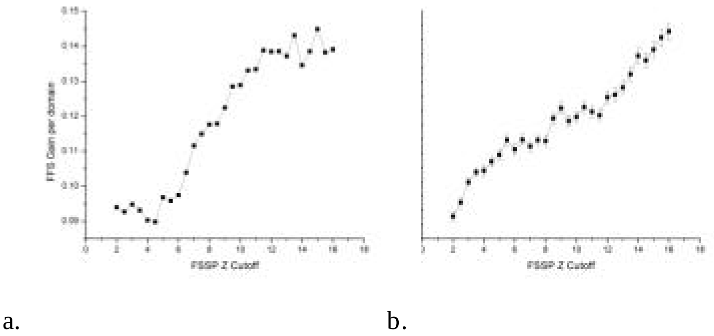
standard deviations away from random. At SCOP cutoff 2 (Fold level) the cusp occurs at Z=6, at SCOP cutoff 3 (Superfamily level) the maximum occurs at Z = 9 and SCOP cutoff 4 (Family level) the maximum occurs at Z = 10-11.

Phase transitions



The size of the largest clusters in FSSP graph plotted against the similarity cutoff threshold at which the graph is clustered. It is worth noting that the size of the largest cluster in the random graph is larger than the largest cluster in FSSP until the end of the phase transition at Z = 12. This is due to the power-law nature of the FSSP graph[4]. The size of the first six largest clusters plotted together as percentage of their original size. The computation was done by ordering the sizes of the clusters at each cutoff and plotting the largest six. The largest six clusters account for vast majority of the domains that are not orphans (singletons). It is worth observing that all the phase transitions occur between Z=6 and Z=9. The behavior of the size of the largest cluster (Fig. 6) and its difference with random bears a striking resemblance to the maxima we just observed on the distance landscapes between the three databases (Figs. 3,4). We can see that there are two very pronounced phase transitions in the size of the largest cluster. The first is from FSSP Z=6 to Z= 9 and the second is from Z=10 to Z=14. These represent the starting and ending points where the largest cluster "suddenly" breaks up into much smaller clusters the largest of which is almost fifty percent of the "parent". The size of the largest cluster in the random graph is always much larger than the size of the largest cluster in the real graph up until Z > 12. Because of this we will argue that the third and final non-random transition occurs at around Z = 11. The behavior of the other clusters closely mirrors that of the largest cluster thus showing that the phase transition is not just the function of the major superfolds but of the majority of the PDUG graph. It is interesting that the first three largest clusters transition at around Z=9 while the smaller three transition closer to Z = 6.

Function Uncertainty



a,b. The FFS[5] gain per domain with respect to structural similarity threshold. FFS of each cluster is compared to that expected by random for a cluster that size and added to the gain at that threshold (Eqs 6, 7). The final FFS gain is normalized by the number of domains annotated in the graph. The majority of the functional information is gained from Z = 6 to Z = 11, before and after those thresholds the information content obtained from structural comparison plateaus. Thus we can quantify the amount of function information gained by correctly annotating a domain to its Fold as .095 bit per domain while correctly identifying the Superfamily yields around .15 bits per domain of functional information. The intrinsic uncertainty with which we can expect annotation of function at a given structural similarity. For example, at Z = 6 (Fold level) on average the domain function cannot be annotated to be more precise than 1.6 bits per level on the GO tree. Note that there are two plateaus where the FFS does not significantly change with respect to Z score: the first starting from Z = 5 to Z = 8 and the other starting from Z =9 all the way to Z = 11 showing an intrinsic correlation between structure and function at the Fold and Superfamily Level of annotation. This once again confirms the theoretical origins of this annotation by showing the conservation of function at those levels of structural comparison.

References:

1. Lo Conte, L., Brenner, S. E., Hubbard, T. J., Chothia, C. & Murzin, A. G. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res* 30, 264-7.
2. Orengo, C. A., Pearl, F. M. & Thornton, J. M. (2003). The CATH domain structure database. *Methods Biochem Anal* 44, 249-71.
3. Dietmann, S., Park, J., Notredame, C., Heger, A., Lappe, M. & Holm, L. (2001). A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. *Nucleic Acids Res* 29, 55-7.
4. Dokholyan, N. V., Shakhnovich, B. & Shakhnovich, E. I. (2002). Expanding protein universe and its origin from the biological Big Bang. *Proc Natl Acad Sci U S A* 99, 14132-6.
5. Shakhnovich, B. E., Dokholyan, N. V., DeLisi, C. & Shakhnovich, E. I. (2003). Functional fingerprints of folds: evidence for correlated structure-function evolution. *J Mol Biol* 326, 1-9.