doi:10.1016/S0022-2836(02)01362-1 J. Mol. Biol. (2003) 326, 1–9

JMB

Available online at www.sciencedirect.com
SCIENCE DIRECT®

COMMUNICATION

Functional Fingerprints of Folds: Evidence for Correlated Structure–Function Evolution

Boris E. Shakhnovich[1], Nikolay V. Dokholyan[2], Charles DeLisi[1] and Eugene I. Shakhnovich[3]*

[1]*Bioinformatics Program
Boston University, Boston
MA 02215, USA*

[2]*Department of
Biochemistry and Biophysics
University of North Carolina
at Chapel Hill
School of Medicine
Chapel Hill, NC 27599, USA*

[3]*Department of Chemistry
and Chemical Biology
Harvard University
12 Oxford Street
Cambridge, MA 02138, USA*

**Corresponding author*

Using structural similarity clustering of protein domains: protein domain universe graph (PDUG), and a hierarchical functional annotation: gene ontology (GO) as two evolutionary lenses, we find that each structural cluster (domain fold) exhibits a distribution of functions that is unique to it. These functional distributions are functional fingerprints that are specific to characteristic structural clusters and vary from cluster to cluster. Furthermore, as structural similarity threshold for domain clustering in the PDUG is relaxed we observe an influx of earlier-diverged domains into clusters. These domains join clusters without destroying the functional fingerprint. These results can be understood in light of a divergent evolution scenario that posits correlated divergence of structural and functional traits in protein domains from one or few progenitors.

Keywords: protein evolution; protein function; structure–function relationship

Most models dealing with molecular evolution concede the importance of the relationship between protein structure and function.[1,2] Understanding this relationship is central not only to evolutionary biology, but also to structural genomics.[3] However, despite many efforts, the establishment of a clear relationship between protein structure and function has been elusive. This is in part due to the fact that neither structural nor functional inter-relationships are well defined. We propose that the structure–function relationship can be understood by considering an evolutionary perspective. One of the most intriguing pieces of data is the uneven distribution of sequences over protein folds.[4] To this end, we explore two views on protein evolution that have previously been suggested to account for the uneven distribution of sequences in fold space: that of convergent evolution.[5–7] and divergent evolution.

Convergent evolution posits that different folds evolved independently and the same ("most popular") protein structures are recycled many times by newly emerged proteins developing under changed functional or environmental pressures.[6] According to this model, new proteins are not related by evolution to their orthologues. New proteins spawn by chance, but some structures are more populated (have more sequences) than others because they are suggested to be more advantageous (thermodynamically, kinetically, or evolutionarily). Such a scenario would suggest limited relationship between structure and function.[8]

An alternative scenario is that of divergent evolution that suggests that a single or a few progenitor proteins give rise to many different, perhaps even unrelated offspring *via* processes of gene duplication and mutation.[9,10] These offspring can differ significantly from each other, either in sequence or structure, and can perform a varied array of functions many generations later.[11] It was shown recently that divergent evolution scenario implies important, observable structural relationships between domains: namely a scale-free organization of the protein universe that relays the history of how proteins are related to each other and would suggest a strong correlative relationship between structure and function.[12]

In Ref. 12 all protein domains were clustered according to their structural similarity as measured by DALI *Z*-scores.[4,13] The resulting graph, where nodes represent protein domains, and edges connect

Abbreviations used: PDUG, protein domain universe graph; GO, gene ontology.
E-mail address of the corresponding author: eugene@belok.harvard.edu

the nodes that correspond to domains which have structural similarity Z-scores greater than some threshold value Z_{min}, represents the protein domain universe graph (PDUG). PDUG provides a "hierarchical" way to annotate structure in terms of Z_{min}. If Z_{min} is decreased, less similar structures join disjoint clusters. The striking observation is that the PDUG is a scale-free network in the number of edges per node as well as in the sizes of domain clusters.[14–16] This special, highly non-random hierarchical structure of the PDUG can be quantitatively explained by a simple model that supposes only divergence of domain structures from few progenitors.

It is important to note that divergent evolution implies a structure–function relationship that complements the non-random structural hierarchy of the PDUG. As protein structures diverged from progenitor proteins, so did functions. This relationship is necessitated by the requirement that both the ancestor protein domain and all its descendants remain functional[17] and structurally stable during the progression of evolution. If evolution of function is similar in mechanism and time to that of structure, we can expect a structure–function relationship that can be observed in the context of a hierarchical functional annotation that allows comparison of protein functions at various levels of specificity of description. The level of hierarchical description is important, as it is the focal lens of functional evolution. Such hierarchical functional description is provided to the bioinformatics community by the GO consortium.[18]

The main result of this work is a striking finding that the corollary relationship between structural evolution and acquisition of new function by protein domains necessitated by a divergent evolution scenario can be quantitatively observed on the PDUG. Looking at PDUG through a hierarchical description of structural comparisons we find that we can characterize different clusters by the "functional fingerprint" that they display. A functional fingerprint is the distribution of functions within a particular cluster. We find that this distribution is quite unique to a given fold family on levels of functional annotation with high specificity of description. If we relax the Z_{min} threshold, we can see an influx of protein domains into structurally similar clusters. Even with the influx of as many as one hundred percent of the earlier diverged domains into the ancestral cluster, the functional fingerprint is not destroyed (does not become random), but is complemented with the same and similar functions. This preservation of unique functional fingerprints through evolutionary dynamics further highlights the close relationship between structure and function necessitated by divergent evolution.

The importance of independent functional hierarchical description

Our simplistic divergent evolution model that explains the non-random behavior of the PDUG[11] is based solely on the premise that a protein has an ancestor that is its closest structural homologue. This model describes well the structural non-homogeneity observed on the PDUG. The model characterizes the "oldest" proteins as those having the largest number of descendants and consequently the number of descendants for each protein depends on the protein's evolutionary age. We can therefore deduce from our divergent evolution model that the older clusters and proteins are more populated and have more connections on the PDUG.

To see whether there is mutual evolution between structure and function, we functionally annotate proteins. In general, proteins have very diverse functional descriptors. Some of these descriptors are unique such as Methionine synthase, b12-binding domains or methylmalonyl-coa-mutase, that describe only a single type of protein. On the other hand, all proteins can be broken up into just six or seven major functional categories such as enzyme, ligand binding, transporter. It seems apparent from the simple limiting cases presented above that the elucidation of a functional relationship between proteins depends on the system of description. Some medium specificity of functional description must be used if we are to quantitatively measure a functional relationships between proteins. Since we do not know the coarseness of the needed annotation, we clearly need a hierarchical system.

A hierarchical system of functional annotation was recently developed by the GO (gene ontology) consortium.[18] The GO system of annotation is well suited for measuring functional relationships between proteins because it defines a machine language where we can compare protein functions with little ambiguity based on their unique GO identifiers at different levels of specificity of annotation. The GO hierarchical language is organized as a directed acyclic graph (DAG). Each node in this graph is an annotation, a functional descriptor that we can assign to a gene or gene product. As the graph is traversed down, more precise functional descriptions populate the nodes. In this graph, the parent-leaf relationship of the nodes has an "all children are a subset of the parent" conjecture. For example, all adolases are enzymes as are CoA ligases because there is an edge from enzymes to both categories. We independently map protein function onto the whole of the PDUG (Figure 1).

Data and methods

We use protein domains as identified by Dietmann and Holm in the FSSP database of protein domains. Structural similarity between each pair of protein domains is characterized by their DALI Z-score. In order to create a graph from domain and comparison data, we define a structural similarity threshold Z_{min} and connect any two domains that have DALI Z-score $Z_c > Z_{\text{min}}$ by an edge. Thus, we create the PDUG. It is crucial

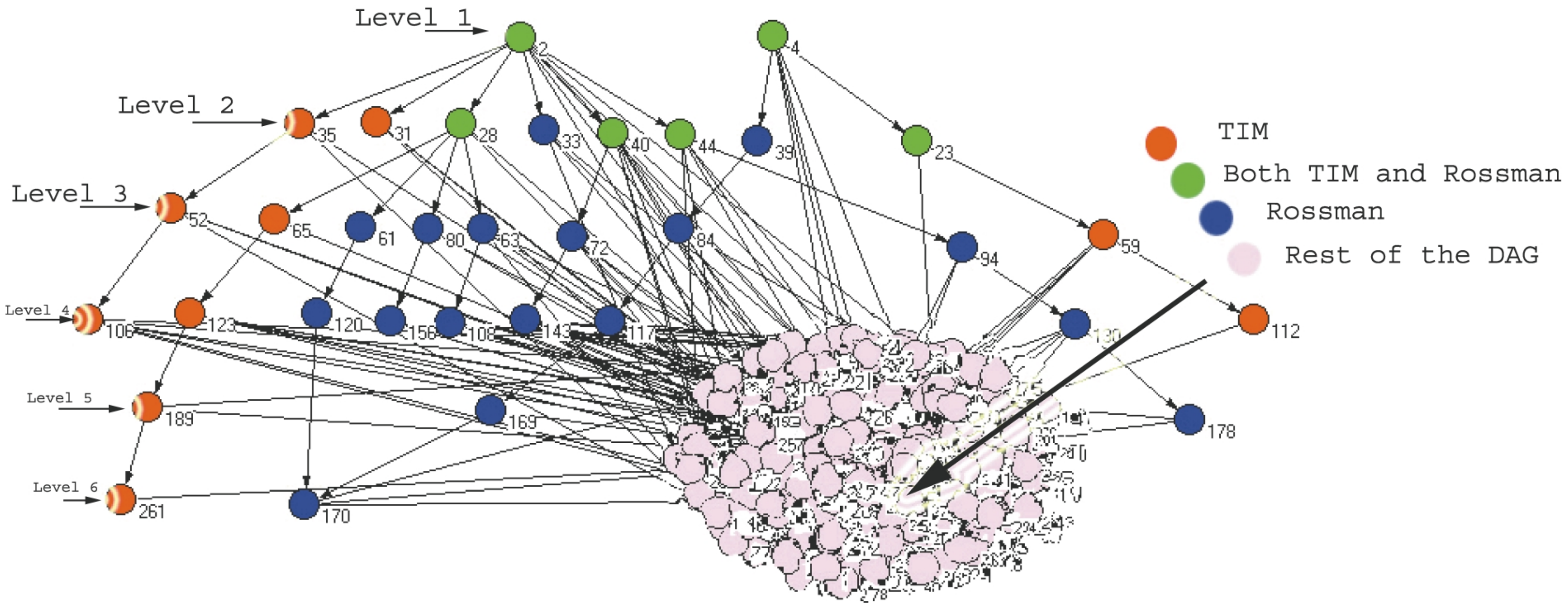

Figure 1. The mapping of the clusters onto the hierarchical GO tree. Nodes are functional categories and edges are subset relations. The specificity depends on the level of hierarchical description. The majority of proteins within each cluster i.e. the functional core either follow a single connected GO path or end up at the same ontology node down the hierarchy. For example, most Rossman Fold proteins end up in node 170 (GTPase), and a vast majority of proteins in the TIM fold end up in node 261 (amylase). Notably, the GO ontology is a very large directed acyclic graph and proteins in each fold follow only a minuscule subset of all possible paths, as shown by the overwhelming majority of nodes that end up outside the followed paths on the GO directed acyclic graph.

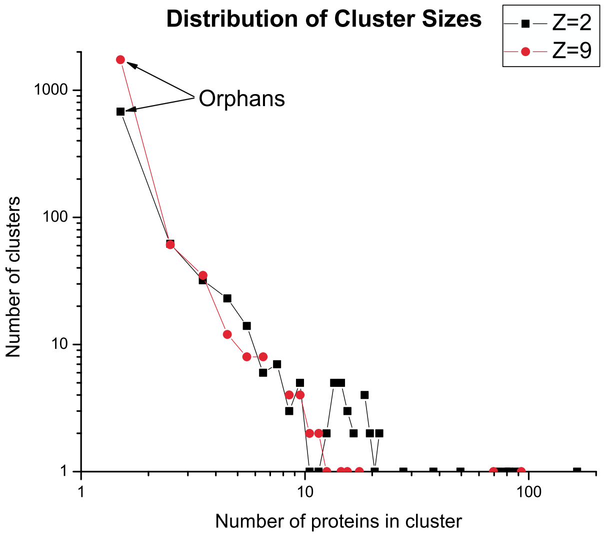

Figure 2. The distribution of cluster sizes at $Z_{\min} = 9$ and $Z_{\min} = 2$ cutoff. The axes are on a Log–Log scale. The distribution shows resemblances to a power law, where many clusters contain only single proteins (orphans) and few clusters contain many proteins.

to note that, in contrast to weighted graphs considered in, the PDUG, is un-weighted graph where each edge that made it above threshold is considered equally. Clustering of this un-weighted graph represents its partitioning into disjoint clusters which can be carried out exactly using the classical depth-first search algorithm. Each disjoint cluster represents a family of structurally related proteins in which each protein is presented only once. Clusters are named according to the largest number of SCOP annotations in that cluster. For example, although at $Z_{\min} = 9$ the largest cluster contains Rossman fold, P-loop fold and others, we will call this the "Rossman fold" cluster for brevity. Disjoint PDUG clusters are, in principle, equivalent to the fold classification level of the SCOP database at some $Z_{\min}$. Figure 2 shows the distribution of cluster sizes for $Z_{\min} = 9$ and $Z_{\min} = 2$ cutoffs.

In order to carry out a completely machine based annotation, we use a direct mapping of the genes found in SwissProt Database that coded for the PDB entry of the protein domain in the PDUG. We mapped the SwissProt entries that correspond to the domains to the curated GO annotation by the Gene Ontology Consortium. Each such annotation was mined independently by the GO consortium primarily from literature searches†. This yielded a non-trivial mapping from PDB to GO, thus giving each protein its functional assignment. The assignment is non-trivial because some SwissProt entries had many functional annotations corresponding to large, multi-functional, multi-domain proteins, from which our domain was only one. In this case, we kept all functional annotations. Working with domains alleviates the problems of "flow of structure" inside the clusters, where a common domain may be shared by unrelated proteins.[19]

Using the annotations for protein domains, the overall GO tree was reconstructed by finding all paths "up" from the annotation and normalizing for the number of domains. For example, if a domain was annotated as an hydrolase and oxidoreductase, the first level would be "enzyme" with one hundred percent assurance, while the second level would be "hydrolase" and oxidoreductase each with fifty percent assurance. This could correspond to different functions of the domains in different genomes or multifunctional proteins. This method also resulted in less annotations on lower levels of the GO ontology as some proteins were not annotated to the last level. This in turn had an effect on the statistical significance of the smaller cluster like IGFold. The statistical significance for IGFold decreased with increasing annotation level due to the relative lack of annotation on those levels.

Divergent evolution observed

Our analysis provides a multidimensional view of the protein domain universe. In our representation, each protein has two dimensions at two threshold parameters: a structural cluster that it was assigned to in the PDUG at some threshold $Z_{\min}$-score, and the functional annotation that was assigned to the same protein from the GO mapping at some level of specificity. Our main goal is to find out if there exists a non-random mapping from the structural space to the functional space. Unique and different functional fingerprints for structural clusters is strong evidence for complementary structure function evolution.

We observe from ordering of the most common functional annotation in each cluster that there are some functions that dominate in structural clusters, and that they differ between clusters (Table 1). We find that these functions that dominate inside clusters are different than those that are just overrepresented for all proteins or even for orphans: single membered clusters in the PDUG. We observe a higher level of homogeneity of function in clusters compared to the distribution of function in the whole of the PDUG, which is aided greatly by the hierarchical nature of GO. We also find that this differentiation of function shows up distinctly at the fifth level of annotation (Figure 3(a)–(f)), while the differentiation at the first level is much less pronounced (Figure 3(g)–(i)). We discover that although each structural cluster (family) has a multi-functional fingerprint, these differ between clusters (Figure 3(a)–(f)).

In order to evaluate the statistical significance of functional fingerprints belonging to individual clusters, we compute the probability that the distribution of function within each cluster occurred by chance (Figure 4). Under this assumption, the probability of a functional fingerprint inside a

† http://www.geneontology.org

Table 1. Dominating function for a given fold at each level of GO. Orphans are all clusters with a single protein domain that has no discernible structural neighbor. Note that for globins, “oxygen transporter” occurs at both the second and third level of of annotation. Due to some protein domains not being annotated at lower level, the total number of annotated protein domains decreases with level. Only the largest five clusters were included in this table. This annotation reflects clustering at $Z_{\min} = 9$

Fold name	Functional level	Functional annotation of the largest group of homologous proteins	Number of members	Percent in the fold
All	I	Enzyme	996	38
	II	Hydrolase	374	13
	III	Purine nucleotide binding	175	7
	IV	Adenyl nucleotide binding	142	7
	V	ATP binding	142	9
	VI	ATPase	23	3
Orphans	I	Enzyme	523	34
	II	Hydrolase	214	12
	III	Purine nucleotide binding	93	6
	IV	Adenyl nucleotide binding	76	7
	V	ATP binding	76	8
	VI	ATPase	18	4
Rossman	I	Enzyme	131	57
	II	Oxidoreductase	55	24
	III	Oxidoreductase, CH–OH group of donors	32	16
	IV	Oxidoreductase, acting on the CH–OH group of donors, NAD or NADP as acceptor	30	17
	V	GTP binding	17	13
	VI	GTPase	7	13
Tim	I	Enzyme	94	85
	II	Hydrolase	32	31
	III	Hydrolase, acting on glycosyl bonds	27	29
	IV	Hydrolase, hydrolyzing *O*-glycosyl compounds	27	33
	V	Amylase	16	25
	VI	Alpha-amylase	14	66
IGFold	I	Signal transducer	21	39
	II	Antibody	16	33
	III	Transmembrane receptor	17	65
	IV	B-cell receptor	17	77
	V	Antibody	16	89
Globins	I	Ligand binding or carrier	17	54
	II	Oxygen transporter	16	44
	III	Oxygen transporter	16	100
Beta-prop	I	Enzyme	10	50
	II	Hydrolase	8	50
	III	Hydrolase, acting on glycosyl bonds	6	50
	IV	Hydrolase, hydrolyzing *O*-glycosyl compounds	6	54
	V	Alpha-sialidase	6	66
	VI	Exo-alpha-sialidase	6	66

cluster follows a multinomial distribution:

$$E\{n_i\} = \frac{N!}{\prod_i n_i!} \prod_i p_i^{n_i} \tag{1}$$

where N is the total number of domains in a given cluster, n_i is the number of domains observed with function i inside that cluster. Thus, $\sum n_i = N$. The value p_i is the probability of randomly picking ith functional annotation. To calculate p_i, we count the number of domains that were annotated with function F_i in the whole of the PDUG and divide by the total number of annotated domains at that level of GO. The variable i is used to count over all functions that are present in the calculated cluster. p_i is the probability of finding a certain function i if randomly taking a protein out of the whole protein domain universe.

The results (Figure 4) show that the statistical significance of the functional fingerprint increases as the functional level of annotation increases except at the sixth level of GO where the amount of information (the number of annotated proteins) is too small for significant statistics, and for IGFold that looses relative statistical significance due to the decreasing number of annotations at higher levels. As a measure of control, we randomly reshuffle edges in the PDUG, creating random clusters of the same sizes as in the original. Using the same techniques as described above, we determine the functional fingerprint of the newly reshuffled clusters. The statistical significance of this random control assignment is negligible (results not shown). Such control presents another evidence for the significance of the unique and different functional fingerprints of structural clusters in the PDUG, and thus for correlated structure–function evolution. Most importantly we find that with the increase in

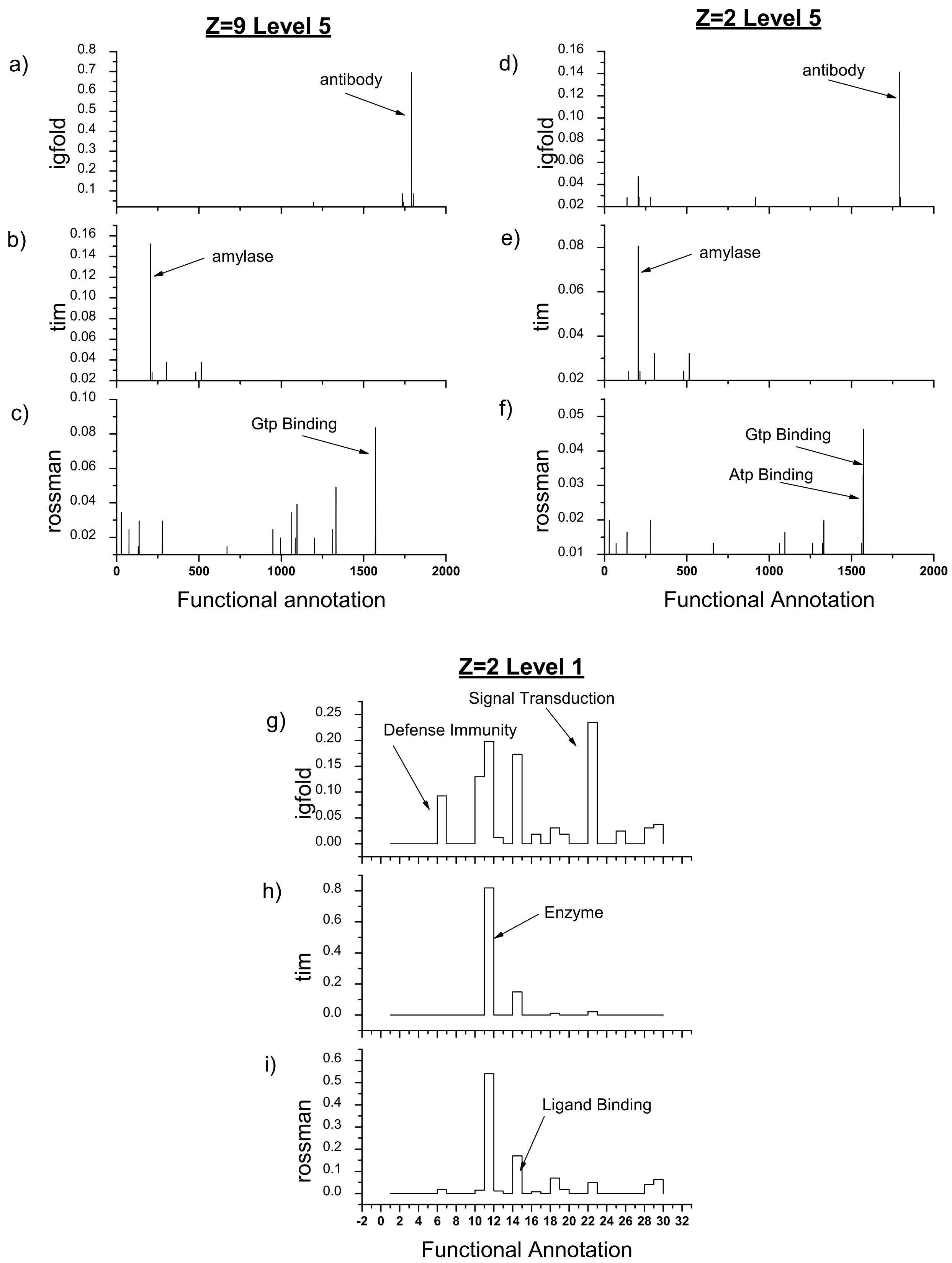

Figure 3. Functional fingerprints of folds. *X*-axis is the functional annotation category. *Y*-axis is the normalized number of proteins annotated with that function. (a)–(c) Functional annotation at the fifth level of GO ontology for $Z_{\min} = 9$. Notably, each cluster has its own, distinct functional fingerprint that is observably different than those of other clusters. (d)–(f) At the fifth level of annotation after proteins joined their ancestral clusters, for $Z_{\min} = 2$. The fingerprints are more diverse, however still differ significantly from each other. (g)–(i) At the first level of annotation with $Z_{\min} = 2$, the fingerprints overlap significantly, and hard to distinguish one from the other.

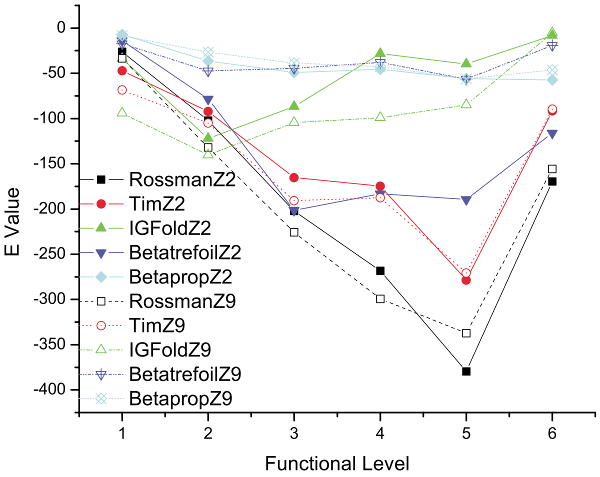

Figure 4. Multinomial distribution calculation of statistical significance of functional fingerprints at each level of GO for both $Z_{\min} = 2$ (solid lines and filled) and $Z_{\min} = 9$ (dashed lines and empty icons). *E*-value of each structural cluster is presented in terms of its functional level of GO. The *E*-value represents the probability of finding a particular functional fingerprint by chance at each level of GO ontology. The *E*-value is computed by applying a multinomial distribution equation at each level of functional description Equation (1). The probabilities are on a log scale. The probabilities are computed with the null hypothesis of a random distribution of protein domains at each level. While the relative *E*-value decreases with increasing functional cutoff, igfold does not follow the same trend because at higher levels, the decrease in the number of annotated proteins decreases the statistical significance. Even though this is true, the statistical significance still remains well above random sampling. Notably, the statistical significance of the increase in the number of protein domains between $Z_{\min} = 9$ and $Z_{\min} = 2$ is offset partly by the "diffusion of the fingerprint in functional space". In this way each cluster takes on a certain statistical significance based on the overall distribution of function (the multifunctionality of fold). However, most beta-trefoils added to a single functional category between $Z_{\min} = 9$ and $Z_{\min} = 2$ thus increasing their relative statistical significance.

the number of proteins in each cluster, the overall statistical significance does not decrease, but in most cases increases between $Z_{\min} = 9$ and $Z_{\min} = 2$. This could only happen if the functional categories that are added to these clusters are not random. In once case, Beta-propeller proteins, the overall statistical significance even significantly rises due to an extremely uneven sampling of function (mostly reinforcing sialidase).

Finally, we explore the chronological imprints of divergent evolution dynamics. In the framework of divergent evolution, the domains that diverged the earliest, connect at low $Z_{\min}$ thresholds and domains that diverged later will have higher $Z_{\min}$ thresholds (Figure 5). In the case of Figure 5 we observe a local diversification of function with evolution. We explore the extent to which relaxation of the structural threshold will affect our functional fingerprints inside the growing clusters. We observe that between $Z_{\min} = 9$ and $Z_{\min} = 2$ half of the proteins that diverged significantly enough as to become unrecognizable at $Z_{\min} = 9$ joined their ancestral fold cluster at $Z_{\min} = 2$. Approximately half of the proteins that were orphans at $Z_{\min} = 9$ joined structural clusters at $Z_{\min} = 2$ (Figure 6). Even when clusters grow by as much as one hundred percent it is striking how well the separation of functional fingerprints is preserved. Orphan domains joined their ancestral structural clusters with very similar functional annotation, i.e. newly added protein domains did not dilute the functional fingerprint of the clusters they joined (Figure 2). This can be quantified by the relative change in the probability of observing a functional fingerprint given random assignment of functions: Equation (1) between $Z_{\min} = 9$ and $Z_{\min} = 2$ for structural family clusters (Figure 4). A conspicuous effect of inclusion of new members into clusters is the addition of new non-homologous proteins to their cluster's central functional categories thereby increasing the overall statistical significance of the fingerprint. However, some proteins add to less populated cluster functions inside the fingerprint, suggesting a possible evolutionary drive toward diversification of function.[20]

Discussion

This study demonstrates strong evidence for divergent evolution of structure and function in protein domains. We clearly observe a non-random homogeneity of function within structural clusters: the functional fingerprint. More importantly, if we compare the structural clusters between each other, their functional fingerprints differ. We observe the phenomenon of older, more populated clusters diffusing more in the functional space than newer, less populated clusters. For example, the largest structural cluster that at $Z_{\min} = 9$ is mainly populated by proteins with the distinctive Rossman fold and P-loop fold, is mainly localized in the guanile nucleotide binding GO annotation heavily weighed towards GTP binding activity.

As we move to less populated, and therefore presumably younger clusters, we observe that the function is more localized. This is probably because the domain family had less time to diverge in structure and consequently function. The TIM barrel fold mainly has the function of hydrolases. IGFolds are obviously very specialized folds performing mostly receptor and defense functions. Interestingly, globins localize almost exclusively (95%) into the "oxygen transporter" functional category. The probability that a set of randomly chosen proteins falls into one particular category at the fifth level of the GO ontology is diminishingly small. For example, for all globins this chance turns out to be on the order of 10^{-80}.

It has been known for a long time that there are specialized folds. For example, the IGFolds are known to perform immunity/defense functions,

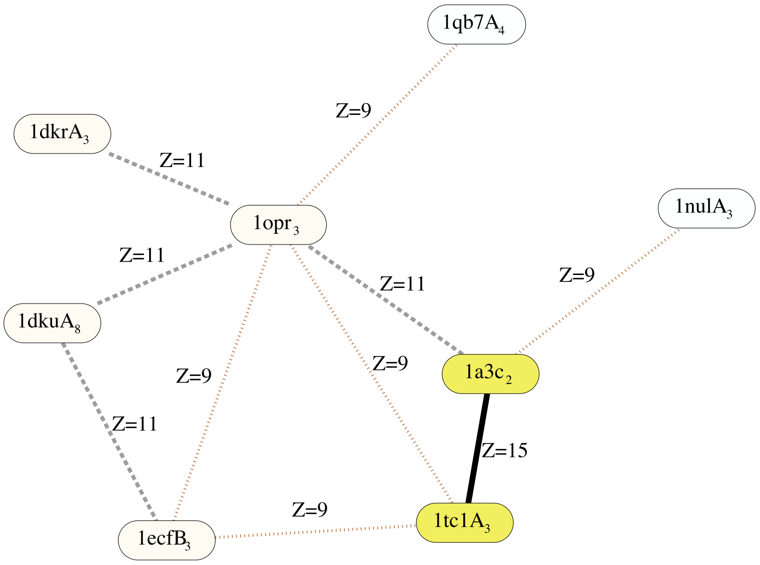

Figure 5. A schematic picture of how protein domains diffuse into structural clusters and define them at different thresholds. Nodes represent protein domains, links represent structural similarity at above threshold. Thick line represents domains connected at $Z_{\min} = 15$, dotted lines represent domains connected at $Z_{\min} = 11$ and thin lines represent domains connected at $Z_{\min} = 9$. Domains inside the PDUG have drastically different connectivity depending on the similarity threshold based on which we connect two nodes. It is evident that at a very high threshold $Z_{\min} = \infty$, all protein domains are orphans, i.e. they are connected to no other domains because their similarity is lower than the structural similarity threshold. At $Z_{\min} = 0$ all protein domains are connected because the threshold is non-discriminatory. As we change the threshold, we observe orphans diffuse into larger clusters of sufficiently similar domains. As the evolutionary threshold $Z_{\min}$ is relaxed, proteins that were orphans and had no structural neighbors join their ancestral clusters. All proteins in the PDUG have less that 25% sequence homology to each other. In this figure, the domains that joined at $Z_{\min} = 15$ are yellow, at $Z_{\min} = 11$ are in beige and at $Z_{\min} = 9$ are white. Here 1a3c is pyrimidine biosynthetic operon repressor from *Bacillus subtilis*, 1tc1 is hypoxanthine phosphoribosyltransferase from *Trypanosoma cruzi*, 1opr is orotate phosphoribosyltransferase from *Salmonella typhimurium*, 1dku is phosphoribosylpyrophosphate synthetase from *Bacillus subtilis*, 1ecf is glutamine phosphoribosylpyrophosphate from *Escherichia coli*, 1qb7 is adenine phosphoribosyltransferase from *Leishmania donovani* and 1nul is xanthine–guanine phosphoribosyltransferase from *Escherichia coli*. This is a great example, of diversification of function with evolution of structure within and throughout genomes.

and it is not surprising to find that its functional annotation differed significantly from all other structural clusters. Yet, we find significant homogeneity even in more ubiquitous and less specialized cluster like those containing TIM-Barrel and Rossman folds.

Secondly, as we attach more diverged protein domains to their ancestral clusters, the proteins attach with closely matching functional descriptions, complementing the functional fingerprint of their ancestors. Importantly when new functions are added to the fingerprint, we never observe these functions join two clusters equally. The function space is sampled unevenly by closely structures joining clusters as $Z_{\min}$ is relaxed. We also notice that there are functional categories that are more populated in the Protein Domain Universe. These are the functions that have many structurally similar, but sequentially different proteins performing them. It would be very interesting to investigate why some functions are more redundant than other ones. We could hypothesize that these are the older functions (those that older proteins started with) that evolved much earlier and consequently have close descendant proteins performing similar functions.

We go on to note that approximately half of the proteins are not orphans even at $Z_{\min} = 9$. An interesting phenomenon is observed as we increase the amount of structural evolutionary time that we look at by decreasing the threshold $Z_{\min}$. Higher $Z_{\min}$ represents a more recent snapshot of

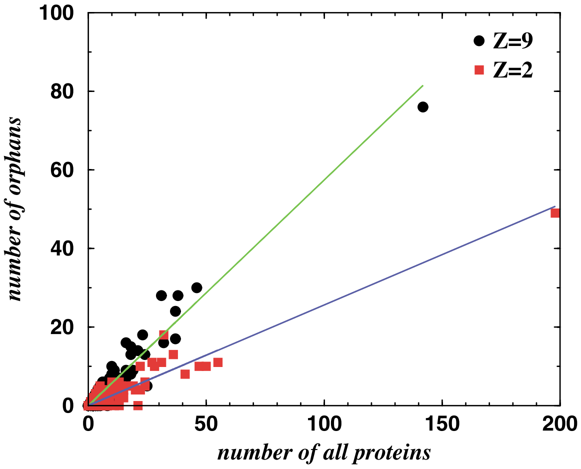

Figure 6. Scatter plot of number of orphans having a specific function *versus* all proteins having the same function. Each data point corresponds to a functional assignment. The total number of functional assignments is 2097. The upper right corner points represent the most populated function: ATPase. Their removal does not affect either the correlation coefficient or the statistical significance of the fit. At $Z_{\min} = 9$, the correlation coefficient is 0.96, and the regression coefficient is 0.57. At $Z_{\min} = 2$, the correlation coefficient is 0.91, and the regression coefficient is 0.25. This plot attests once again to the uneven sampling of function of protein domains joining their ancestral structures.

evolution. As we decrease $Z_{\min}$ from 9 to 2, we see half of the orphans join ancestral clusters (Figure 6), but we also see that the core of the functional annotation within each cluster grows almost proportionally (Figure 3). The functional core is the collection of non-homologous proteins that dominate functional annotations inside clusters. They are visibly seen as propagating together through the GO directed acyclic graph. This is in stark difference to random sampling of the protein domain universe where no such "core" can be found and where we observe a more random distribution of functional annotation across all levels of GO. Notably, as we decrease $Z_{\min}$, many functions peripheral to the nucleus diffuse into the fingerprint.

Distributions of function and clusterization of structure on the PDUG act as evolutionary snapshots. We work with two evolutionary lenses: the functional and structural similarity hierarchies. We look to find evidence for the mutual evolution of structure and function. According to divergent evolution, aside from the biochemical consideration of function structure correlation, there is also a biological pressure for proteins to retain close functional as well as structural similarity to their ancestors upon mutation and duplication. This implies a possibility to trace protein lineages *via* structural comparisons. Strikingly, we observe both the functional homogeneity within structural clusters and the complementary attachment of orphans with similar functions as we allow earlier diverged proteins to diffuse into their ancestral clusters.

Acknowledgements

We would like to thank R. C. Berwick and J. E. Straub for critical reading and support during the preparation of the manuscript. EIS is supported by NIH (GM52126). NVD is supported by NIH NRSA Grant (GM20251-01).

References

1. Aravind, L. & Koonin, E. V. (1999). Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *J. Mol. Biol.* **287**, 1023–1040.
2. Jordan, I. K., Kondrashov, F. A., Rogozin, I. B., Tatusov, R. L., Wolf, Y. I. & Koonin, E. V. (2001). Constant relative rate of protein evolution and detection of functional diversification among bacterial, archaeal and eukaryotic proteins. *Genome Biol.* **2** RESEARCH0053.
3. Baker, D. & Sali, A. (2001). Protein structure prediction and structural genomics. *Science*, **294**, 93–96.
4. Holm, L. & Sander, C. (1996). Mapping the protein universe. *Science*, **273**, 595–602.
5. Li, H., Tang, C. & Wingreen, N. S. (1998). Are protein folds atypical? *Proc. Natl Acad. Sci. USA*, **95**, 4987–4990.
6. Taverna, D. & Goldstein, R. A. (2000). The distribution of structures in evolving protein populations. *Biopolymers*, **53**, 1–8.
7. Csete, M. E. & Doyle, J. C. (1999). Reverse engineering of biological complexity. *Science*, **295**, 1664–1669.
8. Todd, A., Orengo, C. & Thornton, J. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113–1143.
9. Brenner, S. (2001). Natural progression. *Nature*, **409**, 459.
10. Dokholyan, N. V. & Shakhnovich, E. I. (2001). Understanding hierarchical protein evolution from first principles. *J. Mol. Biol.* **312**, 289–307.
11. Ponting, C. P. & Russell, R. B. (2000). Identification of distant homologues of fibroblast growth factors suggests a common ancestor for all beta-trefoil proteins. *J. Mol. Biol.* **302**, 1041–1047.
12. Dokholyan, N. V., Shakhnovich, B. E. & Shakhnovich, E. I. (2002). Expanding protein universe and its origin from the biological Big Bang. *Proc. Natl Acad. Sci. USA*, **99**, 14132–14136.
13. Holm, L. & Sander, C. (1995). Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.* **20**, 478–480.
14. Qian, J., Luscombe, N. M. & Gerstein, M. (2001). Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J. Mol. Biol.* **313**, 673–681.
15. Yanai, I., Camacho, C. J. & DeLisi, C. (2000). Prediction of gene family distributions in microbial genomes: evolution by gene duplication and modification. *Phys. Rev. Lett.* **85**, 2641–2644.
16. Huynen, M. A. & van Nimwegen, E. (1998). The frequency distribution of gene family sizes in complete genomes. *Mol. Biol. Evol.* **15**, 172–184.
17. Cooper, V. S., Schneider, D., Blot, M. & Lenski, R. E. (2001). Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli*. *J. Bacteriol.* **183**, 2834–2841.
18. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M. *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29.
19. Schug, J., Diskin, S., Mazzarelli, J., Brunk, B. P. & Stoeckert, C. J., Jr (1992). Predicting gene ontology functions from ProDom and CDD protein domains. *Genome Biol.* **12**, 648–655.
20. Kashiwagi, A., Noumachi, W., Katsuno, M., Alam, M. T., Urabe, I. & Yomo, T. (2001). Plasticity of fitness and diversification process during an experimental molecular evolution. *J. Mol. Evol.* **52**, 502–509.

Edited by J. Thornton

(Received 20 August 2002; received in revised form 12 November 2002; accepted 22 November 2002)