5

Towards Unifying Protein Evolution Theory

N.V. Dokholyan and E.I. Shakhnovich

5.1 Two Views on Protein Evolution

There are many proposed models for protein domain evolution. The main contention [1], however, is between divergent [2] and convergent [3] models of evolution. Divergent evolution proposes that there was an ancestor to all domains that we see today and consequently all domains are historically related to each other. Convergent evolution claims that chance is largely responsible for the appearance of new kinds of domains. Convergent evolution explains the prevalence of certain types of structures by stating that those structures are more favorable and are often recycled for different functions. The distinction between convergent and divergent evolution is neither abstract nor purely academic. If divergent evolution is the primary model responsible for the appearance of new domains, then it is possible to reconstruct lineages of domains, follow structural change, and predict contextual functional ranges that occur due to well-defined evolutionary pressures. However, if convergent evolution is the main driving force behind the appearance of new structures and functions, in principle there should be no discernable structure–function relationship. An important step in understanding evolution is the quantification of dominant evolution scenario(s) as theoretical background for derivation of the structure–function correlation. Here, we describe recent advances in this direction.

5.2 Challenges in Functionally Annotating Structures

To outline the steps that are needed to create a theory of structure–function correlation, we would like to first outline the challenges in doing so. First, we have a problem of finding the "atomic unit" of functionality in proteins. This atomic unit is commonly taken to be a protein domain. This is so because domains can be both structurally sound and functional outside the protein that they are a part of. The most common ways of finding protein domains rely

heavily on personal "intuition" [4] and are often points of vigorous debate [5,6]. However, such an atomic unit must be agreed upon if a theory of structure–function relation is to succeed. An atomic unit is integral to addressing the structure–function correlation because of the need to "compare" these units. If we pick incorrect atomic units, comparisons will be plagued by problems like "flow of structure" and overlapping sets of nonunique functions.

The determination of the function for a hypothetical protein is currently based on three strategies [7]. The first strategy is based on finding sequence similarity to known proteins. Even at low sequence similarities, there may be a set of conserved amino acids constituting an active site or a conserved hydrophobic core [8–11]. In the case of a conserved active site, similar amino acids may indicate the function of a hypothetical protein. The second strategy involves the search for protein surface cavities using sequence and structural similarities to protein with known function. As in the first strategy, the extent of the success of this methodology depends strongly on the conservation of local sequence and structural motifs. In addition, the second strategy relies on the knowledge of the protein structure. The driving assumption for these strategies is the possible similarity of the active sites between proteins sharing the same or similar function [12,13]. Also, there have been several mechanisms proposed to search for local functional motifs by comparison to libraries of three-dimensional structural templates [13–15] and the analysis of the physical properties of protein surfaces. Teichmann and Thornton [7, 16] described two examples of correct functional annotation of hypothetical proteins: the HdeA protein from *Escherichia coli* [17] and the protein corresponding to gene 226 from *Methanococcus janaschii* [18, 19]. Finally, the third strategy is based on the crystalographic studies of the bound cofactors in the native protein structure. The main limitation of this strategy is that it requires experimental reconstruction of the three-dimensional structures of protein–ligand complexes. The efforts to annotate function exclusively based on structure and sequence, as discussed above, are complicated by the fact that different sequences may fold into similar structures [4] but have different functions. A notable example of functional diversity inside a structurally homologous family is the case of the P-loop NTPases. The structures of RecA (2reb) [20] and adenylate kinase (2ak3) [21] proteins are similar. Both are alpha and beta proteins. Both contain P-loop topology. Both are placed in the same SCOP family [4]. Yet, their functions are quite distinct. RecA is a DNA repair protein, while the adenylate kinase is a transfer protein transferring phosphate groups from AMP to ADP.

The problem is further complicated when some clearly homologous genes, found through high scoring sequence comparison, fold into the same structure, but perform different functions depending on the genome they are expressed in. For example, two sequence homologues of an alpha/beta hydrolase share more than 75% sequence identity. Yet, the *E. coli* version has the function of a serine-activating enzyme, while the human homologue is a lysosomal carboxypeptidase [22].

Another challenge in understanding the structure–function relationship in protein domains is that it is difficult to define quantitatively the metric of the "function space." While structural relationships between protein domains are easily quantifiable, so that several good measures of it exist – Z-score [23–25] and RMSD, the distance in functional space is poorly defined. A way to address this crucial issue is by using a hierarchical description of protein functionality and probabilistic models that aim to quantify functional proximity based on dominating functional annotations.

Finally, the idea of understanding the protein function from individual characteristics is exacerbated by some puzzling findings where evolution cannot be easily traced. These findings were attributed most often to convergent evolution. For example, Ponting and Russel [26] describe the case of the Ser/His/Asp catalytic triad [27], which has been identified in five completely different protein folds. These folds apparently do not share any common ancestor based on extensive sequence similarity searches. Therefore, these folds are not considered homologous, yet their functions are the same and the catalytic triad is also strikingly similar.

5.3 The Importance of the Tree of Life

Along with creating methods for measuring sequence, structure, and metabolic pathway variability, we also have to create a way to describe the context (see [9]) for each sequence and structure. A hierarchical description of the context exists now and is referred to as the "tree of life" [28, 29]. The tree of life describes the most probable divergence of organisms (or most parsimonious way to describe their interrelationship based on multiple sequence alignment of a single ubiquitous protein). An example of a tree based on this principle of parsimony was constructed in the mid-1970s by Woese and collaborators who used SSU rRNA molecule as a molecular chronometer. On the basis of this information, most trees of life are now drawn [30–33]. SSU rRNA is used because it is widespread in organisms and its structure is highly conserved [33].

There is, however, considerable debate about the precision of the phylogeny predicted by a tree drawn purely on rRNAs alone. This dissention stems partly from the fact that other genes give believably different trees for the same set of organisms [34–42]. Further, apparently different trees can be produced by naturalistic methods of traditional phylogeny [43–45]. It has been noted that due to the existence of lateral transfer – a phenomenon where genes are transferred via a plasmid or vector or some other method other than divergence to another organism – nets should be used instead of trees to represent interrelationships between organisms [46]. Indeed, the building of an accurate tree of life or a net of life is very important, both because it is an interesting academic challenge and because it is a necessary part in understanding the context of the structure–function relationship elucidation of such aspects as functional pressure on domain evolution.

5.4 Building the PDUG

To consider sequence, structure, and function [47] information in a unified, systematic way, Shakhnovich et al. [48] defined both gene families and fold families (sequence and structure homologies) quantitatively using the Protein Domain Universe Graph (PDUG) [49]. The PDUG is a graph where nodes are sets of closely related sequences folding into well-defined domains [50, 51] and edges are connections between the nodes that are based on structural similarity (Fig. 5.1), while sequence identity between any pair of sequences belonging to different nodes is less than 25%. To build the PDUG, Dokholyan et al. [49] took all sequences from NRDB90 [52] and all structural domains from HSSP [51]. Shakhnovich et al. [48] further used BLAST [53] sequence homology to detect all sequences in NRDB90 with more than 25% sequence identity to each HSSP domain. That set of sequences was combined into a single gene family. Using cross-indexing between Swiss-Prot [54] and InterPro [55] all equilogs (different sequences with the same function) belonging to every gene family were identified. Those equilogs are further used to reconstruct the functional flexibility score (FFS; see [6]).

Using this PDUG formalism, it is possible to explore global correlations between sequence, structure, and function determinants. For example, we can define a gene family based on micro-evolutionary considerations: the PDUG represents on many evolutionary scales the variability accessible to a given

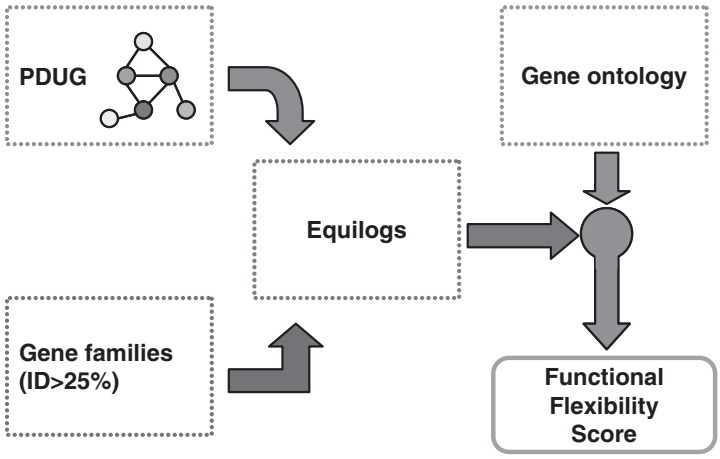

Fig. 5.1. A diagram of the scaled organization and intrinsic properties of the PDUG. The PDUG is built hierarchically: first, domains' structural similarities are compared to each other and from this information the structural graph is created. All the sequences from NRDB with more than 25% identity to the original sequence of each domain on the PDUG are collected into a gene family. All the equilogs (sequences with the same function) matching the gene family are collected and used to create a probabilistic GO tree from which the FFS is calculated using (5.1)

gene upon mutation, whether that variability is in sequence, function, or structure space. Unlike other definitions of gene families [50,56], the definition proposed by Shakhnovich et al. [48] is local, i.e., with respect to a particular gene. The gene family of a gene is, therefore, all the immediate sequence neighbors inside a PDUG node. Similarly, the fold family of a structure is all the structural neighbors of that domain on PDUG (Fig. 5.1). By defining the cutoff value for the sequence or structure comparison, it is possible to control the variability for that gene, thus implicitly controlling the allowed evolutionary divergence time on which structure–function determinants are calculated. This approach turns out to be invaluable in gleaning new insights into structure–function correlation, coevolution, and definition of important properties in the PDUG.

5.5 Properties of the PDUG: Power Laws on Very Different Evolutionary Scales

The properties of the PDUG have been recently addressed in several works [49, 57, 58]. The properties of the PDUG largest cluster were determined [49]. The size of the largest cluster in the PDUG and random control graph were determined and compared as a function of the structural similarity score $Z_{\min}$ [49] defined by FSSP [24]. The random control graphs were constructed by maintaining the same number of proteins and connections as in the actual PDUG, but reshuffling the connections between the nodes. Control random graphs represent an evolution process without any driving force, i.e., any node can be connected to another node by chance. Dokholyan et al. [49] found a pronounced transition of the size of the largest cluster in the PDUG at $Z_{\min} = Z_c \approx 9$. Random graphs featured a similar transition, but at a higher value of $Z_{\min} = Z_c \approx 11$. The distribution of cluster sizes depends significantly on whether $Z_{\min} > Z_c$ or $Z_{\min} < Z_c$ for both the PDUG and random graphs. It was also observed that the probability density $P(M)$ of cluster sizes M for both the PDUG and random graphs followed a power-law at their respective Z_c: $P(M) \propto M^{-2.5}$. The observed power-law behavior of $P(M)$ is simply a consequence of criticality at Z_c as it is featured prominently both for the PDUG and random graphs. The power-law probability density of cluster sizes is a generic percolation phenomenon that has been observed and explained in both percolation [59, 60] and random graph theories [61].

To define more concretely the structural properties of the PDUG, Dokholyan et al. [49] computed the probability $P(k)$ of the number of edges per node k taken at $Z_{\min} = Z_c$ for individual clusters. It is known that $P(k)$ distinguishes random graphs from various graphs observed in science and technology [61]. In stark contrast with the equivalent random graph, the PDUG is scale-free with $P(k) \propto k^{-1.6}$ with a high degree of statistical significance (p-value less than 10^{-8}). The power law fit of $P(k)$ is most accurate at $Z \approx Z_c$, and noticeably deteriorates above and below Z_c. Similar calculations were also

performed on the number of equilogs inside each PDUG node and the number of sequence members coding for individual domains [48]. These measures represent three different levels of evolutionary divergence and characterize the topology of the PDUG. Strikingly, it was discovered that not only do all these values similarly follow a power-law, but they also obey the same exponent $P(k) \propto k^{-1.6}$ This behavior turns out to be the fingerprint of divergent evolution and these data will help differentiating between the histories of divergent and convergent evolution, as well as building proper evolutionary models.

The power-law fit at $Z_{\min} > Z_c$ quickly becomes meaningless as the range of values of connectivity k rapidly diminishes as greater $Z_{\min}$ lead to mostly disconnected domains. At $Z_{\min} < Z_c$ the power law fit also becomes problematic in the whole range of k because at large values of k (50–100) $P(k)$ shows some nonmonotonic behavior that can be interpreted as a maximum at large k (the data are insufficient to conclude that with certainty). However, the remarkable property of a maximum $P(k)$ at $k = 0$ i.e., dominance of orphans remains manifest at all $Z_{\min}$ values. This is in striking contrast with random graph that is not scale-free at any value of $Z_{\min}$ and where $P(k)$ allows almost perfect Gaussian fit with maximum at higher values of k. This power-law behavior also turns out to be in contrast to that generated by convergent evolution models (see [8]). The criterion for selecting cutoff value based on transition in the giant component also turns out to be highly useful in determining proper functional clusters (see [10]).

It is worth noting that the exponent -1.6 in the connectivity distribution was recently corrected and is suggested to be closer to -1.0 [57]. The correction arises from the consideration of the exponential finite size effects, which significantly contribute to the power-law regime [62, 63]. In addition, recent examination of the origin of the scale-free properties of the PDUG suggested that the PDUG is not modular, i.e., it does not consist of modules with uniform properties. Instead, it was found the PDUG to be self-similar at all scales [57].

5.6 Functional Flexibility Score: Calculating Entropy in Function Space

Shakhnovich et al. [48] defined the function determinant of a gene family as entropy in function space. When this measure is calculated in the context of the PDUG, the Gene Ontology (GO) [24] is utilized to define the functional variability (functional flexibility score or FFS) of a set of genes. The FFS is a measure of the total amount of information needed to describe all the functionality of a gene family. Interestingly, the FFS statistically correlates with the logarithm of the total number of sequences in a gene family [48]. Contrary to merely counting the number of members in a gene family, the FFS appears to be a more robust (with respect to possible bias in the databases,

as well as uneven sampling of phylogeny) measure of gene family diversity because it normalizes on the number equilogs, i.e., sequences that diverged far enough to represent functionally diverse proteins. The FFS is a characteristic of both the sequence topology of the PDUG as well as a functional determinant of neighborhoods.

To calculate functional entropy, all sequences were combined into a single set [48]. These sequences were then matched to InterPro [55] equilogs, proteins with different sequences but the same function. The complete GO tree was further reconstructed from the annotations of equilogs and the number of equilogs of the family, which is assigned a particular functional annotation normalized by total number of annotations at each level, is calculated (Fig. 5.1). Thus it is possible to calculate the average amount of information per level needed to fully describe the function of each gene family using the following equation [48]:

$$\mathrm{FSS} = -\frac{1}{\mathrm{Max}(L)} \sum_{l} \sum_{i \in \{\text{nodes on level } l\}} p_i \log(p_i)\,. \tag{5.1}$$

Here, $\mathrm{Max}(L)$ is the maximal number of levels of annotation, the summation is taken over all levels l and over all nodes i filled by the gene family on the GO tree, and p_i is the fraction of the family that is annotated with function i.

5.7 Lattice Proteins and Its Random Subspaces: Structure Graphs

How important is the observation of power law in the organization of the PDUG? Is it a generic feature of proteins as compact polymers or a result of their evolutionary selection? To address this question, Deeds et al. [64] turned to a simple yet exact lattice model of compact 27-mers, whose fully compact conformations have been fully enumerated to yield 103,346 conformations. The structural comparison between all pairs of compact structures can be carried out in a similar way to the DALI method by calculating the number of native contacts that two conformations have in common. Then the lattice structure graph (LSG) is constructed in a similar way to the PDUG [64]. The evaluation of the node connectivity distribution for the complete lattice graph all 103,346 structures and various randomly selected (in a manner consistent with the convergent evolution scenario) subgraphs, as well as subgraphs of the 3,500 most designable lattice structures, yielded LSGs that feature Gaussian rather than power-law distribution of $p(k)$, in sharp contrast with the real PDUG. These results suggest the evolutionary origin of power-law degree distribution in the PDUG.

5.8 Divergence and Convergence Explored: What Power Laws Tell Us about Evolution

To advance toward a resolution of the debate distinguishing convergent and divergent evolution, Deeds et al. [65] explored the predictions of convergent models. The simplest such model assumes that nodes are discovered completely randomly – that is, the likelihood of adding a particular node has nothing to do with the number of structural neighbors it has. In this class of convergent models, as organisms evolve and speciate, nodes were added to each proteome randomly, producing organisms whose structural domains represent a random subset of the existing protein domains taken from all organisms. As with the entire PDUG, one can create a network from the structural similarity between the domains within a particular proteome, and under this type of convergent model, the resulting graph would be a random subgraph of the PDUG. Thus, unbiased convergence leads to the null hypothesis that proteome-specific subgraphs will be random subgraphs of the extant protein universe.

Given a template graph of N_0 nodes with a distribution of edges per node described by $p_{N_0}(k)$, the average distribution of edges per node in a subgraph of N nodes chosen completely randomly should follow:

$$p_N(k) = C\left[\sum_{s=k}^{\mathrm{Max}k_{N_0}} \binom{s}{k}\left(\frac{N}{N_0}\right)^k\left(1-\frac{N}{N_0}\right)^{s-k}\right] \tag{5.2}$$

where $\mathrm{Max}k_{N_0}$ is the maximally connected node in the template graph and C is taken to normalize $p_N(k)$. This expression is similar in justification to those used in percolation theory to estimate the behavior of scale-free networks when a fraction of nodes are eliminated [66–68]. It is clear from the theory that the degree distributions of random subgraphs exhibit a power-law region that persists for a characteristic length at each subgraph size. Deviations to higher k's in subgraphs constitute an exponential tail to the distribution, indicating that random subgraphs are very unlikely to contain nodes with connectivity greater than this value. The maximum k in random subgraphs is thus a "fingerprint" of random subgraphs of a given size.

On the basis of the understanding of the behavior of random subgraphs, (5.2), Deeds et al. [65] turned to actual proteomes to determine whether they represented random subgraphs of the PDUG. The authors created the organismal subgraphs using homology to determine which of the domain structures in the PDUG occur in each organism. For each subset of nodes found in each proteome, Deeds et al. [65] calculated the degree distributions for these subgraphs.

In addition, Deeds et al. [65] performed a power-law regression on the degree distributions for the subgraphs of 59 fully sequenced bacterial proteomes, all of which were well fit by a power law. This raises the possibility

that these subgraphs may indeed be random; however, the maximum k in the organismal subgraphs was considerably larger than that observed in arbitrary random subgraphs of the same size. The probability that organismal PDUGs are random subgraphs of the complete PDUG was estimated by comparison of their actual maximum k with distribution of maximum k in random subgraphs based on (5.2). Such probability was found to be extremely low in most organisms except few smallest ones where it reached about 0.01.

This finding suggests that the null hypothesis corresponding to a convergent evolution model is highly unlikely and, thus, that unbiased models have a low likelihood of explaining protein structural evolution. Even given the inability of convergent models to explain this behavior, it is not necessarily clear that this nonrandomness could be observed in a divergent model similar to those that have been proposed [49, 69]. To explore this possibility, Deeds et al. [65] modified an earlier model described in detail by Dokholyan et al. [49] to include speciation (Fig. 5.2). Simulations of the speciation model that included generation of 3,500 model proteins and four organisms were performed, with speciation occurring after 1,000 and 2,000 steps, and then the degree distributions of the subgraphs in each organism were compared to that of the overall graph produced by the model. For ten realizations of the model, the model organism subgraphs exhibited power-law degree distributions that deviated to larger maximum k than expected at random. The p-values for the model

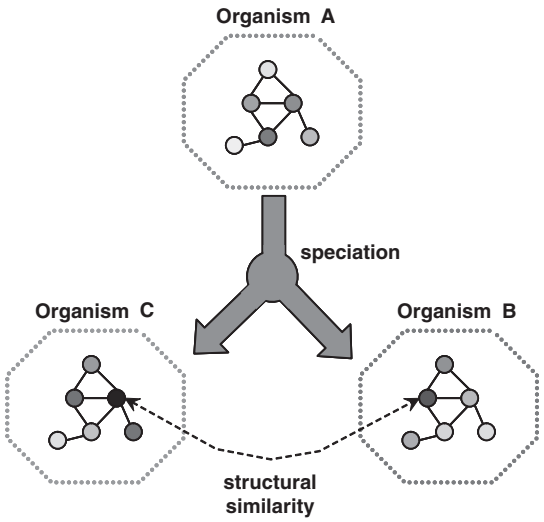

Fig. 5.2. A diagram of the speciation model. The PDUG evolves from all nodes belonging to one organism, and after a specified gene duplication event, an organism A undergoes speciation to create two organisms B and C with identical graphs. All the new nodes, however, are added to one organism or the other, given that the duplication events that give rise to new nodes will only occur within a single proteome. After the proteomes evolve independently for a number of steps, speciation occurs again, and so on

subgraphs indicated that the bulk of these graphs did not have a high probability of being random (although the probabilities were higher than observed for the actual organisms). Although this model is certainly not detailed enough to explain the evolution of real protein structures, it does demonstrate that a simple, dynamic process can produce model organism subgraphs that are similar to those observed in real organisms and that with careful modeling the debate between convergence and divergence can be resolved. Most importantly, these data show that convergent evolution is highly unlikely and that divergence is probably the dominating force in protein domain evolution.

5.9 Context Is Important

The above examples describing differences in function for the same protein between different genomes (see [2]) show that understanding the structure–function correlation involves buttressing structure information with "contextual" information. We see that functionality may vary for the same structure depending on the genome, or the metabolic pathway. This is intuitively understandable if we look at structure–function correlation from an evolutionary standpoint [7, 22, 26, 70, 71]. Domains that were put under different evolutionary pressures (which in our case translates directly into being in a different context) evolve different functions. Consequently, we think that to understand the relationship between structure and function, it is not enough only to enumerate the possibilities, but it is also necessary to understand the progression that has led to the state of protein domain universe as we observe it now.

In a recent work, Shakhnovich [72] has demonstrated that using phylogenetic information it is possible to dramatically and quantitatively improve functional annotation. He introduced, besides the structural similarity measure (Z-score) and the functional similarity (FFS), a measure of phylogenetic distance (P) and presented a 3-dimensional landscape in (Z, FFS, P) space. Dramatically, this landscape was found to be well-shaped. This finding points out that pronounced structure–function correlation is observed only for domains that are phylogenetically close, while for phylogenetically distant domains correlation between structure and function (i.e., between Z and FFS) essentially vanishes. This study shows clearly that "context" information, such as the origin and phylogenetic history of a protein domain, in some cases may influence the precise function of the gene more than the structure.

5.10 Not All Functions Are Created Equal and Neither Are Structures

With the PDUG and the newly developed techniques, it is possible to perform structure–function studies on a global scale capturing evolutionary relationships that are not easily revealed by anecdotal studies alone. An interesting

problem to address is the coevolution of gene family size, functionality, and structure. For example, we can ask whether some functions require smaller gene families than others by computing correlations between particular functionality and the FFS. Since the FFS reflects the size of the gene family, this is equivalent to asking whether there is any bias in the kind of function performed by domains encoded by large gene families versus small ones. Previous research has shown that some functionalities may allow many analogous functions increasing the FFS of the family, while others tend to have stricter requirements. For example, some functions such as the kinase activity have varied specificities within a relatively small number of sequence mutations, [73] while others such as globins have much less functional flexibility despite, in some cases, substantial sequence divergence [74]. Another example is the eightfold (β/α) barrel structure, first observed in triose-phosphate isomerase, occurs ubiquitously in nature. It is nearly always an enzyme and most often involved in molecular or energy metabolism within the cell. This extreme example of the "one fold-many functions" paradigm illustrates the difficulty of assigning function through a structural genomics approach for some folds. Another example is the beta-propeller fold that appears as a very fascinating architecture based on four-stranded antiparallel and twisted beta-sheets, radially arranged around a central tunnel. Similar to the β/α-barrel (TIM-barrel) fold, the beta-propeller has a wide range of different functions. Some proteins containing beta-propeller domains have been implicated in the pathogenesis of a variety of diseases, such as cancer, Alzheimer's, Huntington's, Lou Gehrig's diseases, arthritis, familial hypercholesterolemia, retinitis pigmentosa, osteogenesis, hypertension, and microbial and viral infections. While some studies exist that suggest that gene families encoding enzymes and enzymatic folds are larger [75,76], they do not provide us with an overall picture, if one exists, of general biases. Such biases may reveal important evolutionary pressures that determine the codependence of structure and function.

Shakhnovich et al. [48] computed the FFS of all domains on the PDUG and in turn assigned a functional category at the first level of GO (Fig. 5.1). It was found that as the FFS of the domain increases, the percentage of enzymes in the bin decreases, and consequently the percentage of domains with signal transduction activity increases. Other functions remain relatively constant, at least to the first approximation of these data. These results point to the tendency of larger gene families (tests were also done directly between gene family size and FFS with the same results) to encode domains with signal transduction activity and less diverse gene families to encode domains with enzymatic activity. These results are unexpected in the light of the studies mentioned above [75–77] and more research has to be done in this area to understand the biological reasons and evolutionary mechanisms that have enabled us to observe these trends. However, the obvious avenue of research from here would involve the determination of the interdependency between the allowed functional diversity for the gene family and the particular function.

An approach to begin addressing this question is to look at physical characteristics of the structure. From a physical perspective, the potential of a gene to obtain new function may depend on its ability to accept mutations without destroying the three-dimensional structure of a protein domain that it encodes. The PDUG enables us to begin testing such hypothesis. As mentioned before, we know from some studies that some folds such as TIM-Barrels [75] and beta-propellers [77] encode large sequence families that are functionally diverse. However, until now, no specific structural characteristic could be identified that corresponded to the allowed functional diversity. Using the PDUG and the FFS, it was possible to attribute functional diversity to structure, which is supported with previously reported results of different functional categories having different FFS [78]. For example, most $\alpha+\beta$ protein domains are involved in binding functions such as DNA binding, ATP grasp, and FAD binding activity.

These results could mean that transcription factors are not as selective as once thought and that noise in form of nondeleterious mutations to the fold may give rise to the necessary transcriptional noise in the expression of proteins. Of course, this finding sheds no light on the intrinsic qualities of all alpha proteins versus $\alpha + \beta$ proteins that enables the latter to support more sequences. It also says nothing about what biological mechanisms and pressures enabled separations between commonly used folds and orphans. Clearly coevolution of structure and function occurs.

5.11 Concluding Remarks

The presented overview provides strong evidence that methodologies based on the use of the PDUG work in a variety of applications. Such an approach makes it possible to address from a unique single prospective such seemingly disconnected issues as character of protein domain evolution, structure–function relationship, relation between structural and functional properties of proteins, and certain properties of genes that they encode. We are confident that the PDUG approach is likely to yield further important fundamental unifying insights into structural genomics, evolution, and structural biology.

References

1. N.V. Dokholyan, E.I. Shakhnovich in *Power Laws, Scale-free Networks and Genome Biology*, ed. by E.V. Koonin, Y.I. Wolf, G.P. Karev (Eurekah.com and Springer, Berlin Heidelberg New York, 2005)
2. M.Y. Galperin, D.R. Walker, E.V. Koonin, Genome Res. **8**, 779 (1998)
3. R.F. Doolittle, Trends Biochem. Sci. **19**, 15 (1994)
4. A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, J. Mol. Biol. **247**, 536 (1995)

5. A.S. Siddiqui, U. Dengler, G.J. Barton, Bioinformatics **17**, 200 (2001)
6. U. Dengler, A.S. Siddiqui, G.J. Barton, Proteins **42**, 332 (2001)
7. S.A. Teichmann, A.G. Murzin, C. Chothia, Curr. Opin. Struct. Biol. **11**, 354 (2001)
8. N.V. Dokholyan, E.I. Shakhnovich, J. Mol. Biol. **312**, 289 (2001)
9. N.V. Dokholyan, E.I. Shakhnovich, Proceedings of the International School of Physics "Enrico Fermi" **145**, 227 (2001)
10. U. Bastolla, M. Porto, H.E. Roman, M. Vendruscolo, Proteins **58**, 22 (2005)
11. U. Bastolla, M. Porto, H.E. Roman, M. Vendruscolo, Gene **347**, 219 (2005)
12. R.B. Russell, P.D. Sasieni, M.J. Sternberg, J. Mol. Biol. **282**, 903 (1998)
13. J.A. Irving, J.C. Whisstock, A.M. Lesk, Proteins **42**, 378 (2001)
14. A.C. Wallace, N. Borkakoti, J.M. Thornton, Protein Sci. **6**, 2308 (1997)
15. R.B. Russell, J. Mol. Biol. **279**, 1211 (1998)
16. S. Jones, J.M. Thornton, J. Mol. Biol. **272**, 133 (1997)
17. K.S. Gajiwala, S.K. Burley, J. Mol. Biol. **295**, 605 (2000)
18. K.Y. Hwang, J.H. Chung, S.H. Kim, Y.S. Han, Y. Cho, Nat. Struct. Biol. **6**, 691 (1999)
19. B. Stec, H. Yang, K.A. Johnson, L. Chen, M.F. Roberts, Nat. Struct. Biol. **7**, 1046 (2000)
20. R.M. Story, I.T. Weber, T.A. Steitz, Nature **355**, 318 (1992)
21. K. Diederichs, G.E. Schulz, J. Mol. Biol. **217**, 541 (1991)
22. A.E. Todd, C.A. Orengo, J.M. Thornton, J. Mol. Biol. **307**, 1113 (2001)
23. L. Holm, C. Sander, J. Mol. Biol. **233**, 123 (1993)
24. L. Holm, C. Sander, Nucleic Acids Res. **25**, 231 (1997)
25. L. Holm, C. Sander, Trends Biochem. Sci. **20**, 478 (1995)
26. C.P. Ponting, R.R. Russell, Annu. Rev. Biophys. Biomol. Struct. **31**, 45 (2002)
27. G. Dodson, A. Wlodawer, Trends Biochem. Sci. **23**, 347 (1998)
28. W.F. Doolittle, Science **284**, 2124 (1999)
29. A.J. Roger, O. Sandblom, W.F. Doolittle, H. Philippe, Mol. Biol. Evol. **16**, 218 (1999)
30. W.F. Doolittle, J.R. Brown, Proc. Natl. Acad. Sci. USA **91**, 6721 (1994)
31. N.R. Pace, Science **276**, 734 (1997)
32. C.R. Woese, Microbiol. Rev. **51**, 221 (1987)
33. C.R. Woese, O. Kandler, M.L. Wheelis, Proc. Natl. Acad. Sci. USA **87**, 4576 (1990)
34. K.S. Makarova et al., Genome Res. **9**, 608 (1999)
35. D. Tumbula et al., Genetics **152**, 1269 (1999)
36. J.A. Lake, R. Jain, M.C. Rivera, Science **283**, 2027 (1999)
37. R.F. Doolittle, Nature **392**, 339 (1998)
38. M.C. Rivera, R. Jain, J.E. Moore, J.A. Lake, Proc. Natl. Acad. Sci. USA **95**, 6239 (1998)
39. M. Ibba et al., Science **278**, 1119 (1997)
40. M. Ibba et al., Proc. Natl. Acad. Sci. USA **96**, 418 (1999)
41. T.D. Edlind et al., Mol. Phylogenet. Evol. **5**, 359 (1996)
42. J.R. Brown, W.F. Doolittle, Proc. Natl. Acad. Sci. USA **92**, 2441 (1995)
43. C.J. Bult et al., Science **273**, 1058 (1996)
44. R.P. Hirt et al., Proc. Natl. Acad. Sci. USA **96**, 580 (1999)
45. J.A. Lake, Proc. Natl. Acad. Sci. USA **91**, 1455 (1994)
46. E. Hilario, J.P. Gogarten, Biosystems **31**, 111 (1993)

47. B.E. Shakhnovich, N.V. Dokholyan, C. Delisi, E.I. Shakhnovich, J. Mol. Biol. **326**, 1 (2003)
48. B.E. Shakhnovich, E. Deeds, C. Delisi, E.I. Shakhnovich, Genome Res. **15**, 385 (2005)
49. N.V. Dokholyan, B. Shakhnovich, E.I. Shakhnovich, Proc. Natl. Acad. Sci. USA **99**, 14132 (2002)
50. C.A. Orengo, F.M. Pearl, J.M. Thornton, Methods Biochem. Anal. **44**, 249 (2003)
51. S. Dietmann et al., Nucleic Acids Res. **29**, 55 (2001)
52. L. Holm, C. Sander, Bioinformatics **14**, 423 (1998)
53. S.F. Altschul et al., Nucleic Acids Res. **25**, 3389 (1997)
54. B. Boeckmann et al., Nucleic Acids Res. **31**, 365 (2003)
55. R. Apweiler et al., Bioinformatics **16**, 1145 (2000)
56. E.L. Sonnhammer, S.R. Eddy, E. Birney, A. Bateman, R. Durbin, Nucleic Acids Res. **26**, 320 (1998)
57. N.V. Dokholyan, Gene **347**, 199 (2005)
58. G. Tiana, B.E. Shakhnovich, N.V. Dokholyan, E.I. Shakhnovich, Proc. Natl. Acad. Sci. USA **101**, 2846 (2004)
59. S. Havlin, D. Ben-Avraham, Adv. Phys. **36**, 695 (1987)
60. D. Stauffer, A. Aharony, *Introduction to Percolation Theory* (Taylor and Francis, Philadelphia, 1994)
61. B. Bollobas, *Random Graphs* (Academic, London, 1985)
62. N.V. Dokholyan et al., Physica A **266**, 55 (1999)
63. N.V. Dokholyan et al., J. Stat. Phys. **93**, 603 (1998)
64. E.J. Deeds, N.V. Dokholyan, E.I. Shakhnovich, Biophys. J. **85**, 2962 (2003)
65. E.J. Deeds, B. Shakhnovich, E.I. Shakhnovich, J. Mol. Biol. **336**, 695 (2004)
66. R. Albert, A.-L. Barabasi, Rev. Mod. Phys. **74**, 47 (2002)
67. R. Cohen, K. Erez, D. Ben-Avraham, S. Havlin, Phys. Rev. Lett. **85**, 4626 (2000)
68. R. Cohen, D. Ben-Avraham, S. Havlin, Phys. Rev. E **66**, 036113 (2002)
69. E.V. Koonin, Y.I. Wolf, G.P. Karev, Nature **420**, 218 (2002)
70. J.M. Thornton, C.A. Orengo, A.E. Todd, F.M. Pearl, J. Mol. Biol. **293**, 333 (1999)
71. A.G. Murzin, Curr. Opin. Struct. Biol. **8**, 380 (1998)
72. B.E. Shakhnovich, PLoS Computat. Biol. **1**, e9 (2005)
73. G. Manning, G.D. Plowman, T. Hunter, S. Sudarsanam, Trends Biochem. Sci. **27**, 514 (2002)
74. D. Bashford, C. Chothia, A.M. Lesk, J. Mol. Biol. **196**, 199 (1987)
75. N. Nagano, C.A. Orengo, J.M. Thornton, J. Mol. Biol. **321**, 741 (2002)
76. M.Y. Galperin, E.V. Koonin, Protein Sci. **6**, 2639 (1997)
77. V. Fulop, D.T. Jones, Curr. Opin. Struct. Biol. **9**, 715 (1999)
78. B.E. Shakhnovich, J.M. Harvey, J. Mol. Biol. **337**, 933 (2004)