Chapter 10

Scale-Free Evolution:

From Proteins to Organisms

Nikolay V. Dokholyan* and Eugene I. Shakhnovich

Introduction

One of the most intriguing problems in molecular biology is the origin of the vast population diversity of protein families.[1-4] Following the assumption that the protein families are populated at random, one would expect a multinomial distribution of the family populations.[5] However, it has been discovered[6-9] that distribution of the family populations is by far nonexponential, but has a long tail, which signifies that some specific mechanisms govern populations of protein families. To explain such diversity, there emerged two views of *convergent* and *divergent* evolution (Fig. 1).

In the *convergent* evolution scenario,[10] it is postulated that the present population distribution of protein fold families is the result of convergent processes in the course of evolution which were selectively populating folds. In the simplest scenario, it is presumed that evolution has reached equilibrium in the protein structural space.** Due to the underlying physical nature of evolutionary processes, i.e., the physical nature of amino acid interactions that underlie the properties of specific folds, the expected equilibrium distribution of family population follows the Boltzman distribution. Thus, more "designable" folds, that can be encoded by many sequences, have a higher representation in genomes.[11,13-18] This assumption, called the "designability principle", is based on phenomenological considerations[13] and on observations drawn from exhaustive enumeration of all sequences in simplified two- and three-dimensional lattice protein models. In the course of evolution more designable folds become more populated than less designable folds, which results in the uneven distribution of observed populations of protein families.[17,19]

There have been several arguments[10,13-15,20] based on various observations favoring convergent evolution. Teichmann et al proposed that structural similarities arise solely due to physical interactions that favor particular packing and chain topologies.[20] Functional pressure was proposed to be the paladin of protein structural convergence. One of the most striking example is that of the Ser/His/Asp catalytic triad,[10,21] which is found in a number of folds that have no significant sequence similarity. Antifreeze proteins (AFP) provide a crucial defense for

** Strictly speaking the equilibrium may have not been reached, nevertheless protein families can still be populated according to some "attractive" features such as designability.[11,12]

*Nikolay V. Dokholyan—Department of Biochemistry and Biophysics, The University of North Carolina at Chapel Hill, School of Medicine, Chapel Hill, North Carolina 27599, U.S.A. E-mail: dokh@med.unc.edu

Power Laws, Scale-Free Networks and Genome Biology, edited by Eugene V. Koonin, Yuri I. Wolf and Georgy P. Karev.

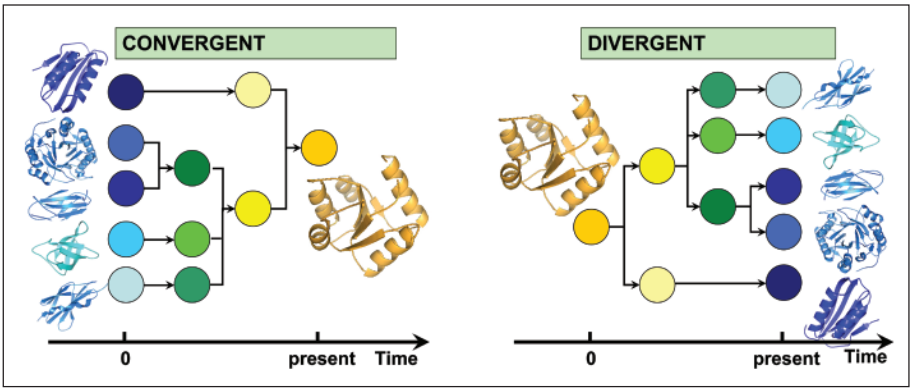

Figure 1. Schemes of the convergent and divergent evolution scenarios.

organisms against sub-zero temperatures. The beetles *Dendroides Canadensis* and *Tenebrio Molitor* AFP have dissimilar sequences from known plant and fish AFP,[22,23] indicating functional convergence of AFP proteins. Another striking example is of Zn-dependent carboxypeptidases that cleave off the C-terminal amino acid residues from proteins and peptides.[24,25] Two families of nonhomologous carboxypeptidases—thermolysin and mitochondrial processing peptidase[24]—show functional and structural similarity, although the topologies—the arrangement of the secondary structure elements in these folds—are different. In general, most of the observations of convergent evolution have relied on finding functional similarity (and structure of the active site) between proteins with low sequence similarity.[20,26-28]

The arguments of protein hereditary relation based solely on their sequence similarity are questionable. It was shown that a sequence is not a robust feature of proteins.[29] In fact, a protein structure often remains stable after a single amino acid substitution. Amino acid substitutions may accumulate in the course of evolution: some of them will be destabilizing, some will be stabilizing. However, as long as a protein itself is stable in its environment, there may be no reasons for it to be eliminated from the genome of a corresponding organism. One can argue that amino acid substitutions, that affect protein folding kinetics and function, may be more damaging for the viability of a protein in a cell. However, many single-domain proteins have just a few amino acids—*protein folding nucleus*[30-32]—that govern fast folding kinetics. In a lowest approximation, multi-domain protein folding kinetics is also governed by a few amino acids—he nuclei of each individual domain. Thus, even though the evolutionary may exert pressure to preserve important amino acids for protein folding kinetics, the small number of them may not prevent proteins diverge strongly in the course of evolution.

The evolutionary pressure to preserve functionally important amino acids (e.g., the active site) may depend on the number of alternative proteins in cells that are capable of performing the same function. Even if a protein were to lose completely its function, it may still survive in a cell and acquire a new function later on in the course of evolution. For example, both enoyl-CoA hydratase and 4-chlorobenzoyl-CoA dehalogenase show significant sequence and structure similarities, but they catalyze different reactions.[33] In addition, in proteins, that play a structural role in the cell, such as fibronectin,[34] functionally important amino acids are the same as those that stabilize these proteins. In proteins, with a binding site, the number of amino acids that constitute a binding site is small. Therefore, the evolutionary pressure to preserve functionally important amino acids may not affect strongly the ability of proteins to diverge during evolution.

It is possible that the evolutionary pressure to preserve a protein's sequence in a more "designable" fold family is not strong enough for protein sequences to diverge from each other

up to the point when their sequence similarity becomes of the order of randomly chosen proteins. Thus, the sequence may not be a robust measure of hereditary relation between proteins.

In the *divergent* evolution scenario, the present day proteins are the pra-children of a small set of prebiotic proteins. They diverge from a few "original" proteins by duplication, deletion, and accumulating amino acid substitutions.[10,35-37] Crucial support for divergent evolution came when the gene duplication was documented.[38-40] The principal advantage of the divergent over convergent evolution scenario is that the former does not rely on the "designability" principle.* Of course, there still are examples of proteins[10] that have similar structure function but vastly different sequence, which are disputed to be an indication of the convergent evolution. However, since the sequence may not be indicative of hereditary relation, such argument is unsupported. If structure is more conserved in the course of evolution, it is important to retrace the evolutionary relation based on structure for those proteins that have sequence similarity below the accepted level for the homologous proteins (approximately 25%).

Gerstein and Levitt pioneered the structural census of protein sequences[41] and discovered that most popular folds—that are most often used in proteins of various organisms—constitute the largest fold families.[41,42] This fact, however, can also be interpreted from *both* convergent and divergent evolution scenarios. Based on convergent evolution scenario, those folds that are most adaptive to a new function are more populated than more "rigid" folds are populated. From the divergent evolution perspective, the more often a given fold is used in the cell, the more often it is expressed and, therefore, the more often it varies in the course of evolution.

The absence of the experimental crux makes it challenging tests strongly support one scenario versus another. Despite the large number of examples that favor divergent or convergent scenarios, there is no unified biophysical theory that would combine into a single theoretical framework for understanding apparently disconnected observations within a single evolutionary concept.

Protein Evolutionary Relationships from Structure Similarities

One step towards a unifying theory of protein evolution is the reconstruction of protein relationships based on their structural similarity. There have been several efforts made in quantifying structural similarities between proteins.[4,43,44] The ambiguity in all of these efforts arises from complications in rigorous quantitative definition of structural similarity. Semi-intuitive definitions of folds have been employed to construct two popular databases, SCOP[4] and CATH.[43] The main drawback of these databases is that they are somewhat subjective.

The FSSP database based on the DALI structure comparison algorithm[44] defines a quantitative measure of structural similarity, the Z-score. However, selection of the threshold value Z_{min} of the Z-score, beyond which proteins are considered structurally similar, also introduces an element of ambiguity into FSSP-based family classification. In a recent paper,[45] Getz and coauthors provided a quantitative relationship between FSSP, CATH and SCOP classifications. These authors noted that the matrix of pairwise Z-scores can be viewed as a weighted graph, where each two proteins that have similarity $Z > 2$ ($Z = 2$ is the minimal Z-score reported in FSSP) are connected by an edge that carries weight corresponding to the Z-score similarity between these two proteins. Getz et al[45] employed clustering algorithms, developed for weighted graphs, to identify fold families. However, clustering of weighted graphs is not exact as it may depend on the chosen algorithm and other factors. Another well-known problem with structural classification of whole proteins presented in FSSP is so-called "floats" where two structurally unrelated proteins having a common "promiscuous" domain are identified as structurally similar. It is, therefore, crucial to reconstruct protein structural relationships taking into account the problem of "floats".

* The divergent evolution and prevalence of more designable structures do not contradict each other.[12]

Protein Structure-Function Relation from Evolutionary Perspective

Functional annotation of proteins is crucial for our understanding of how the cooperative organization of proteins in cells relates to the specific cell anatomy and function.[20] Understanding cell anatomy and function is, in turn, important for understanding the evolution of organisms. On a practical side, the ability to alter a cell's function and/or development may aid rational drug development. However, one of the challenging tasks of structural genomics is the determination of protein function based on its structure.

The determination of the function of a hypothetical protein is currently based on three strategies.[20] The first strategy is based on identifying any sequence similarity to known proteins. Even at low sequence similarities, there may be a set of conserved amino acids constituting an active site. These amino acids may indicate the function of a hypothetical protein.[46,47] The principal limitation of this strategy is the extent to which functionally important amino acids are conserved. It has been demonstrated on five various fold families,[29] that evolutionary pressure to preserve functionally important amino acids may not be as strong as the pressure to preserve amino acids responsible for protein stability in cells. Therefore, the determination of "true" conservation of amino acids due to their functional role may be arduous.

The second strategy for functional assignments of hypothetical proteins is the search for protein surface cavities using sequence and structural similarities to proteins 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. The driving assumption for such strategy is the possible similarity of the active sites between proteins sharing the same or similar function.[48-50] There have been several mechanisms proposed to search for local functional motifs by comparison to libraries of three-dimensional structural templates[27,28,49,51,52] and the analysis of the physical properties of protein surfaces.[53] Teichmann et al[20] described two examples of structural genomics leading the functional annotation of hypothetical proteins: the HdeA protein from *Escherichia coli*[54] and the protein corresponding to gene 226 from *Methaococcus janaschii*.[55,56]

The third strategy is based on the crystallographic studies of bound cofactors in the native protein structure. The main limitation of this strategy is that it requires experimental reconstruction of the three-dimensional structure of protein-ligand complexes, which may be unsuccessful. Even in successful cases, the time scale for the experimental structure determination is much larger than that by using bioinformatics approaches described above.

Due to the severe limitations of all three strategies, it is, thus, crucial to develop a novel technique to rigorously relate protein structure to protein function. Shakhnovich et al proposed a strategy that is based on the assumption of evolutionary relation of proteins that may be so distant that neither structural nor sequence similarities directly are able to identify the function of a given protein. This strategy is to identify a divergent evolutionary pathway—a set of structurally similar proteins that link two dissimilar proteins.[57]

Protein Evolutionary Relations within and between Individual Proteomes

An overwhelming amount of various experimental observations, DNA sequencing data, and resolved protein structures in the past few decades open inviting opportunity to understand the cell machinery at a molecular level. This opportunity, however, is hampered by the fact that there is no unifying view that would serve as a framework for a theoretical basis to explain all available data from molecular to cellular levels of descriptions. Present knowledge offers us understanding of biological processes at various scales: from small molecules living at the Angstrom scale (10^{-10} m) to organisms living at the meter scale. It is an enticing challenge then to bridge these scales by developing a unifying theory.

One step to create a bridge between the nano- and hundred-nano-scales is to reconstruct cell organization at the molecular level. To construct such a bridge it is necessary to reconstruct the cell protein-protein interaction network. A large number of techniques have been developed for the systematic analysis of protein interactions,[58-60] such as yeast two-hybrid-based methods,[61,62] surface plasmon resonance biosensors,[63] isothermal titration calorimetry,[64] optical spectroscopy,[65] mass spectrometry of protein complexes,[66-68] protein chips,[69] and other methods that combine computational and experimental approaches.[70] These methods aim to reconstruct full-scale protein interaction networks in primitive organisms, such as yeast[66,67] and *Helicobacter pylori*.[62] These methods indeed offer novel insight on protein interactions, although their application is currently limited to the simpler unicellular organisms.

Computational methods are alternative approaches to experimental ones. Large amounts of available biological data and cost-effectiveness made computational approaches recently bloom. There have been undertaken several principal computational efforts. The phylogenetic profile method[71-73] is based on comparison of complete genomes of various organisms. Such comparison can be correlated with the set of specific functions present in one organism and absent in another. The principal drawbacks of this method are that (1) it can be used only for complete genomes, (2) some functions may be redundant and not represented by the same set of proteins, and (3) it can not be used for most common and essential proteins to most organisms. The conservation of gene neighborhoods has been utilized to predict functional genes in bacterial genomes.[74-76] The applicability of this approach though is limited to bacterial genomes. A search for domain fusion events[46,77-80] has been used to find the functional role of promiscuous domains incorporated in various larger proteins across the phyla. Such a search, though, is limited to multi-domain proteins. Other methods, such as *mirrortree*[81] and *in silico* two-hybrid methods[82] to search for protein interaction networks are sensitive to coverage of species under study, since they are dependent on multiple sequence alignments. It is crucial to develop a theoretical basis for techniques that would reconstruct the protein relations within individual proteomes and reconstruct the evolutionary relations between them based on available data.

Sequence Divergence

There are several principal facts about protein sequence-structure relation observed:[1,3,4,13,15,18,33,83-90] (i) proteins taken from various species and having sequence identity, *ID*, at least *ID* = 25-30% have similar three-dimensional structures (*native state*)[90-97] and are said to belong to the same fold family; (ii) some pairs of proteins sharing the same fold have sequence similarity as low as expected for random sequences *ID* ~8-9%;[44,87,98] (iii) within the same fold family, protein sequences have only 3-4% "anchored" amino acids.[87]

In 1987, eleven leading evolutionary biologists[99] made a statement asking the scientific community for the appropriate usage of the term "homology". Two proteins are said to be homologous if they possess a common evolutionary origin (e.g., ref. 100). Because many proteins that have high sequence similarity are homologous, this term has been used loosely in the discussion of any proteins with high sequence identity. Proteins that have no common ancestor, but possess structural similarity, are called analogs.

If the sequence identity of two structurally similar proteins is high (*ID* > 25-30%), there is a high probability that these proteins share a common ancestor, and thus, statistically, one would rarely be mistaken when calling these two proteins homologs. If the sequence similarity of two structurally similar proteins is low (*ID* < 25%), it is difficult to establish whether these proteins are homologs or analogs.[29,35] In fact, despite clever efforts,[100] it is still questionable whether there is a unique solution to the problem of determining whether two proteins with low sequence identity are homologs or analogs, i.e., whether they evolved in divergent or convergent evolution.

Two proteins are likely to be homologs that diverged from the same root if they still carry the same function (i.e., if the evolutionary time elapsed from their common divergence point is smaller than functional relaxation time τ_F). However, if two structurally similar proteins with low sequence identity have significantly different functions, then there is little information with which to identify them as homologs or analogs. These two proteins might be homologs, although one of them has evolved to possess a new function.[33] However, these two proteins can also be analogs and their similarity in structure is purely accidental or, for example, is due to a potential similarity of the structure of the binding site. The question then becomes – how can we retrace the history of these two proteins?

Our results[29] suggested that it may be impossible to retrace the history of two structurally similar proteins with low sequence identity based purely on sequence analysis. In this case, the ancestral relation classification terms—homologs and analogs—become meaningless. There are two reasons we believe this to be so. To explain these reasons, in reference 29 we proposed a model of evolution (*Energy Gap Model*) that attempts to reproduce the principal protein observations (i-iii) described above. The Energy Gap Model is based on the design of a set of structurally identical sequences by the *Z*-score minimization.[101-104] The idea is to find the similarities in the sequences of such a set and to recover those residues that are conserved across this set. The protein folding theory[105,106] suggests that *Z*-score minimization is equivalent to maximizing the energy gap between misfolded or unfolded conformations and the native state of a protein. It has been pointed out that such maximization results in stable and fast-folding proteins.[102,107] Thus, by designing sequences that have the same fold, we attempt to mimic evolution in diversifying protein sequences for the same fold family. In addition, the Energy Gap Model is a dynamical model, i.e., there is an implicit time scale that allows one to follow the evolution of sequences during the design procedure. The model is discussed in detail in reference.[29]

Why It May Be Impossible to Reconstruct Hereditary Relations Between Proteins Based Solely on Their Sequence Similarity?

Firstly, the correlation function $C(\tau)$, which measures the probability of an amino acid not to be affected by mutations in time τ, decays exponentially, so that beyond the correlation function relaxation time one can not relate the sequences—original, and the one observed at time τ later. Secondly, it did not make a difference if we started our design procedure from one sequence or from two unrelated sequences. These sequences diverged so much from each other in a short design simulation time, that one could not identify which initial sequence we used in the design procedure. Furthermore, our results[29] suggested that some degree of homology may occur even between sequences that converged from unrelated root to the same structure, i.e., in clear analogs. The reason for that is that as we showed in reference 29 some positions may feature conserved residues due to physical requirement of stability of a common fold. Physical conservation of certain classes of amino acids at some positions in protein folds may be reflected on the genetic level due to the specifics of genetic code. Such conservation in some cases may be confused with homology due to the origin of sequences in divergent evolution. A rigorous definition of analogs and homologs can therefore come only either from the understanding of the correlation times τ between consecutive mutations or by reconstructing the actual structural and/or functional evolutionary pathways. If the time scale is smaller than the typical time scale for the formation of a family of homologs, τ_0, then the homology is well-defined: the homologous sequences in this case have high sequence similarity, while the analogous sequences have low sequence similarity. At a longer time scale $\tau >> \tau_0$, unless there is a high sequence similarity between sequences, the notion of homology and analogy becomes meaningless.

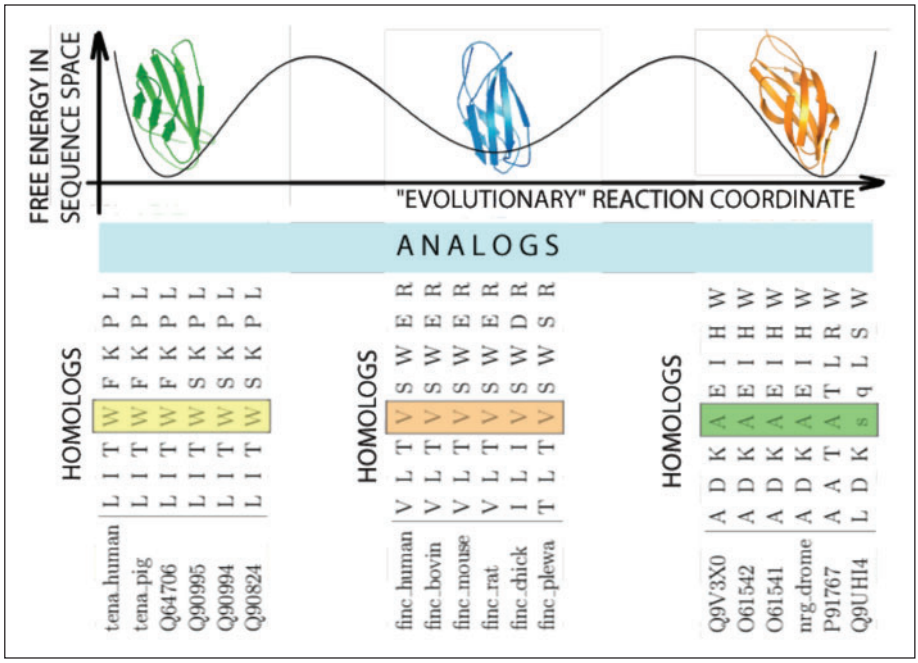

Figure 2. A schematic representation of the evolutionary processes that result in conservation patterns of amino acids. For a given family of folds, e.g., immunoglobulin (Ig) fold in this diagram, there are several alternative minima (3) in the hypothetical free energy landscape in the sequence space as a function of the "evolutionary" reaction coordinate (e.g., time). Each of these minima are formed by mutations in protein sequences at some typical time scales, τ_0, that do not alter the protein's thermodynamically and/or kinetically important sites, forming families of homologous proteins. Transitions from one minimum to another occur at time scales, $\tau=\tau_0 \exp(\Delta G/T)$ where ΔG is the free energy barrier separating one family of homologous proteins from another. At time scale τ mutations occur that would alter several amino acids at the important sites of the proteins in such a way that the protein properties are not compromised. At time scale τ the family of analogs is formed. In three minima we present three families of homologs (1TEN, 1FNF, and 1CFB) each comprised of six homologous proteins. We show 8 positions in the aligned proteins: from 18 to 28. It can be observed that at position 4 (marked by blocks) in each of the families presented in the diagram amino acids are conserved within each family of homologs, but vary between these families. This position corresponds to position 21 in Ig fold alignment (to 1TEN) and is conserved.

The Underlying Scenario of Protein Evolution

We conjecture that the hierarchical organization of structurally similar proteins may be the result of the separation of the evolutionary time scales, shown schematically in Figure 2. On a time scale τ_0, a set of mutations occur that do not affect those amino acids that play crucial thermodynamic, kinetic and/or functional roles. As a result, there is little variation in sequences at the important sites of proteins. If a mutation occurs at the thermodynamically, kinetically and/or functionally important sites, it usually substitutes amino acids with close physical properties so that core, nucleus and/or functional site are not disrupted and the protein folds into its family fold, is stable in this fold, and its function is preserved. At this time scale, a family of homologs is born.

Rarely, at time scale τ, correlated mutations or larger-scale sequence rearrangements occur[108-110] that modify *several* amino acids at the core, nucleus and/or functional site, so that the

stability and kinetics of proteins are not altered. Such a set of mutations can drastically modify the sequence of the protein. However, within the time scale τ_0, a family of homologs is born within which there is conservation of (already new) amino acids in the specific (important) sites of homologous proteins. Although there are alternations in the specific sites of the proteins at the time scale τ, these sites are more preserved than the rest of the sequence. The proposed view of protein evolution is consistent with the observations of the hierarchical organization of structurally similar proteins in families of homologs. Sets of families of homologs are organized, in turn, in super-families of (possible) analogs. The evolutionary time in our analysis is associated with the number of mutations that accumulate in the course of evolution. Because the rates may vary between families and even proteins, the relation of evolutionary time to physical time is not straightforward. Evolutionary time can be rigorously defined statistically as the number of mutations that occur in a fold family, averaged over all family members. The real time for one family may be different from that of another. These considerations complicate interpretation of sequence-based approaches to organismic phylogeny and calls for more robust, structure based approaches to phylogeny (Deeds, Hennesey, Shakhnovich, in preparation).

Support for such a scenario comes from several studies reporting observations of correlated mutations in proteins in the course of evolution.[108,109,111] In addition, Axe et al[112] have demonstrated that random substitution of core residues in ribonuclease barnase by hydrophobic residues preserves the activity of barnase in a significant number of cases. They produced barnase mutants in which 12 of 13 hydrophobic core residues have together been randomly replaced by hydrophobic alternatives. A strikingly high proportion (23%) of mutants maintained structural integrity enough to support enzymatic activity of barnase.

Murzin[33] proposed an elegant scenario of the evolution of protein architecture while maintaining its function. He argued that protein folding pathways may be altered by mutations. As a result, a local free energy minimum of the wild type protein may become a global free energy minimum of a mutant protein. The conformations at these states—global free energy minima of mutant and wild type proteins—may have no structural resemblance. However, these states may maintain the same function. As an example, Murzin argued that catalytic domain of the carboxypeptidase G_2[113] is structurally similar to aminopeptidase from *Aeromonas Proteolytica.*[114] However, these enzymes fold into two topologically different topoisomers.

Reconstructing Evolutionary Relations between Proteins

To overcome difficulties of reconstructing the evolutionary relation of proteins based on their sequences, we resort to the analysis of structural relationships between proteins. We employed a graph representation of the protein domain universe, in which we considered only protein *domains* that do *not* exhibit pairwise sequence similarity in excess of 25% and each such protein domain represented a node of the graph.[8] We used protein domains as identified by Dietmann and Holm in the FSSP database of protein domains.[115] Structural similarity between each pair of protein domains was characterized by their DALI Z-score.[115] We defined a structural similarity threshold $Z_{\min}$ and connected any two domains on our graph that had DALI Z-score $Z \geq Z_{\min}$ by an edge. This way we created the protein domain universe graph (PDUG). It is crucial to note that, in contrast to weighted graphs considered in reference 45, the PDUG, is an unweighted graph where each edge that made it above threshold is considered equally. Clustering of such an unweighted graph represents its partitioning into disjoint clusters which can be carried out exactly using the classical depth-first search algorithm.[116] Each disjoint cluster represents a family of structurally related proteins in which each protein is presented only once (Fig. 3). Disjoint PDUG clusters are, in principle, equivalent to the *fold* classification level of the SCOP database.[4]

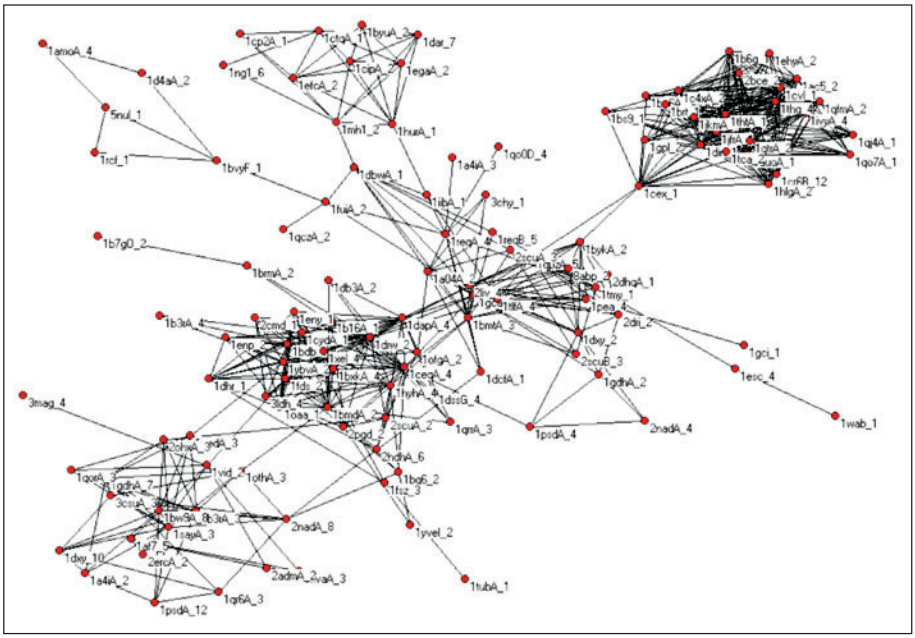

Figure 3. An example of a large cluster of TIM-barrel fold protein domains. Protein domains whose DALI similarity *Z*-score is greater than $Z_{min} = 9$ are connected by lines.

Properties of the Protein Domain Universe Graphs

We computed the size of the largest cluster in PDUG and random control graph as a function of Z_{min}.[8] We found a pronounced transition of the size of the largest cluster in PDUG at $Z_{min} = Z_c \approx 9$. The random graphs feature 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. We also found that the probability density *P(M)* of cluster sizes *M* for both the PDUG and random graphs follows 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[117,118] and random graph theories.[119] Gerstein and coworkers also reported a power-law distribution for fold family sizes derived from the SCOP database[7] and attributed the observed power-law distribution to a certain evolutionary mechanism. However, we showed in reference 8 that random graphs featured the same power-law distribution for fold family sizes and were simply explained by percolation theory.[117-119]

In order to characterize the structural properties of the PDUG we computed the probability $\wp(k)$ of the number of edges per node *k* taken at $Z_{min} = Z_c$ for individual clusters. It is known that $\wp(k)$ distinguishes random graphs from various graphs observed in science and technology.[120] In drastic contrast with the equivalent random graph, the PDUG is scale-free with $\wp(k) \propto k^{-1.6}$ with a high degree of statistical significance (*p*-value less than 10^{-8}). The power law fit of $\wp(k)$ is most accurate at $Z_{min} \approx Z_c$, and noticeably deteriorates above and below Z_c. The 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) $\wp(k)$ shows some nonmonotonic behavior which 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 $\wp(k)$ at $k = 0$ i.e., dominance of orphans remains manifest at all $Z_{\min}$ values. This is in striking contrast with random graph which is not scale-free at any value of $Z_{\min}$ and where $\wp(k)$ allows almost perfect Gaussian fit with a maximum at higher values of k.

The discovery of the scale-free character of the protein domain universe is striking. It has immediate evolutionary implication by pointing out the possible origin of all proteins from a single or a few precursor folds—a scenario parallel to the origin of the Universe from Big Bang. An alternative scenario, whereby protein folds evolved de novo and independently, would have resulted in random PDUG rather than the observed scale-free one.

The rigorous method of clustering protein *structures*[8] provides a number of insights. First of all, using graph theory for protein structure classification removes the ambiguities that are inherent in the highly useful, albeit manual, approaches to structural classification of proteins.[4,43] Perhaps not surprisingly we observed that the structure of the graph representing the protein domain universe depends on the $Z_{\min}$ threshold value of Z-score above which protein domains are considered structurally similar and are connected by an edge of the graph. However, at a certain critical value $Z_{\min} = Z_c$, the structure of the PDUG becomes remarkably universal, simple and amenable to theoretical understanding from an evolutionary standpoint.

An important component of the analysis presented in reference 29 is random control where PDUG was compared with random graph. Our results showed that random weighted graph having the same weight (Z-score) distribution as PDUG featured same cluster size distribution. Since clusters in PDUG can be associated with fold level classification of protein structure, this observation suggested that nonuniform distribution of nonhomologous proteins over folds may not be due to special features of "most popular" protein folds as suggested previously by some researchers.[15,85] However that does not necessarily imply that observed protein folds are not selected based on their physical properties.[121] It is possible that the divergent evolution scenario described here occurs only on these selected folds while unfeasible ones are not observed in nature. However the analysis presented in reference 29 points out that explanation of the nonuniform distribution of nonhomologous proteins over observed folds does not require invoking the "designability principle"[15] or related conjectures about the nonuniform density of sequences in space of protein folds.[85]

We discovered that the structure of the PDUG is by far nonrandom, but rather represents a scale-free network featuring power-law distribution of the number of edges per node. The most striking qualitative aspect of the observed distribution is the much greater number of "orphans" (i.e., domains that are not structurally similar to any other domains) compared with random graph control. Importantly this qualitative feature remains prominent at any value of threshold $Z_{\min}$ despite the fact that power-law fits of $\wp(k)$ gets worse when $Z_{\min}$ deviates from Z_c. A natural explanation of this finding is from a divergent evolution perspective. The model of divergent evolution presented in reference 8 is in qualitative agreement with PDUG as it produces a large (compared with random graph) number of orphans (Fig. 4).

Besides reproducing the scale-free behavior of the PDUG, the divergent evolution model also quantitatively captures more specific graph properties of PDUG. In particular it was shown that the distribution of clustering coefficients[122] of nodes of PDUG is almost exactly matched by the divergent evolution model. This is in contrast to the random control where the scale-free PDUG has been randomly rewired while connectivity of each node is kept intact (Deeds & Shakhnovich, unpublished results).

Orphans are created in the model mostly via gene duplication and their subsequent divergence from a precursor. This may be meaningful biologically because duplicated genes may be under less pressure and hence prone to structural and functional divergence. The divergent

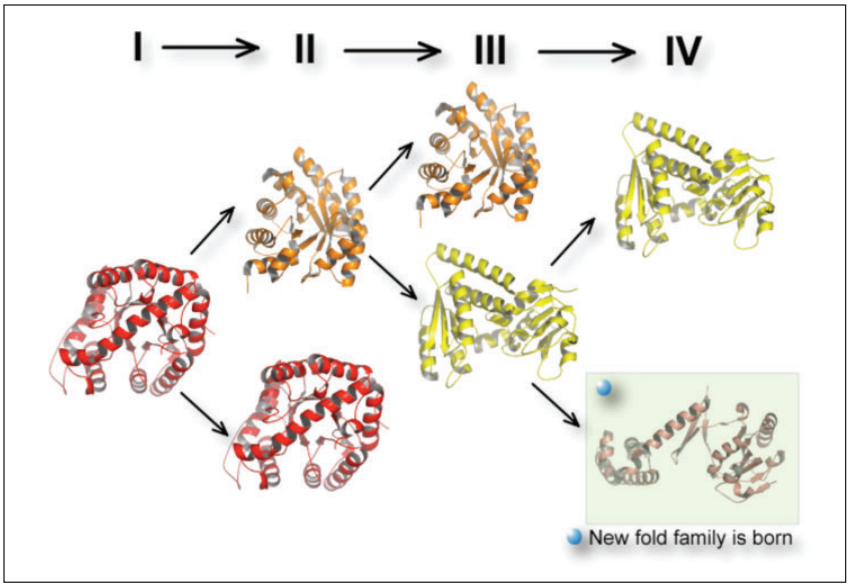

Figure 4. A cartoon representation of the divergent evolution model presented in reference 8. In this model, proteins diverge through a number of gene duplication events and point mutations (I→II→III→IV). While a single amino acid substitution may not significantly alter the protein structure, a number of them may result in drastic changes in protein structure. If these changes result in a functional and, most importantly, stable protein, a new fold family is born. In the model, each protein is represented by a node. Nodes representing proteins with significant structural similarity are connected by edges with a weight. If in the course of evolution an edge's weight lowers below the threshold value, the nodes become disconnected. At each evolutionary step, a randomly chosen node is duplicated and an edge with a weight (chosen from uniform distribution) is created that connects progeny to its parent. If the weight is below a threshold value, the nodes become disconnected and a new protein family is born. In addition, at each evolutionary step, after gene duplication a newly created node may become connected to its pra-parent.

evolution model presented in reference 8 is a schematic one as it does not consider many structural and functional details and its assumptions about the "geometry" of the protein domain space in which structural diffusion of proteins occurs may be simplistic. However, its success in explaining the qualitative and quantitative features of PDUG supports the view that all proteins might have evolved from a few precursors.

An important aspect of the model proposed in the reference 8 is that it provides only a conceptual framework for reconstructing protein structural space. The fine details of evolution contain crucial ingredients that underlie selective pressure in the model proposed in reference 8. Recently Deeds et al[123] uncovered how the features of an underlying protein structural space might impact protein structural evolution using lattice polymers as a completely characterized model of this space. In reference 123 we developed a measure of the structural comparison of lattice structures in analogy to the one used to understand structural similarities between real proteins. We used this measure of structural relatedness to create a graph of lattice structures and compared this graph (in which nodes were lattice structures and edges were defined using structural similarity) to the graph obtained for real protein structures. In reference 123 we found that the graph obtained from all compact lattice structures exhibited a distribution of structural neighbors per node consistent with a random graph. We also found that subgraphs of 3500 nodes chosen either at random or according to physical constraints, such as selective

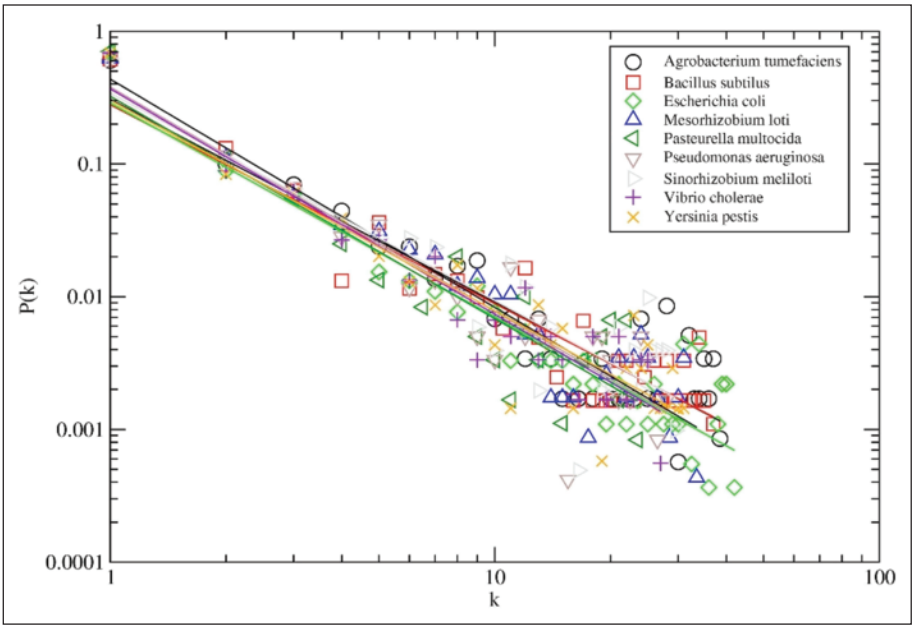

Figure 5. Degree distributions for nine bacterial subgraphs.[124] The degree distributions were shifted by a degree of 1 to allow display of orphan (degree 0) nodes on a log–log plot. Notably all proteomes exhibit similar to PDUG *P(k)* behavior: the power-law fits of these organisms yielded exponents that were approximately -1.6.

protein designability,[11] also represented random graphs. We developed a divergent evolution model based on the lattice space which produces graphs that were capable of recapitulating the scale-free behavior observed in similar graphs of real protein structures. Indeed, in contrast to this universal behavior, we observed subgraphs with power-law degree distributions only as the result of a very specific evolutionary sampling procedure. This not only demonstrated that scale-free graphs may be derived from such spaces but also that the rules underlying divergent graph evolution models are sufficient to produce this behavior.

Evolution of Proteins and Organisms

An important observation has been made by Deeds et al[124] who determined the structural content of 59 fully sequenced bacterial proteomes. Deeds et al identified structural proteomes—subgraphs of PDUG that belong to a specific organism through mapping of the PDUG representative domains on to a homologous domain of a given organism. Each such proteome contains a subset of domains from the PDUG. Strikingly, Deeds et al[124] found that these subgraphs are themselves scale-free networks (Fig. 5).

Deeds et al[124] explored two convergent evolutionary models to explain the scale-free organization of proteoms and concluded that such models were unlikely to explain the PDUG structural patterns. Addition of speciation events to a divergent model, however, resulted in model organisms that exhibit nonrandom subgraphs similar to those observed for real organisms. Deed et al[124] demonstrated that any divergent model must include some ingredient of speciation in order to account for the nonrandom overlap between structural proteomes. Such analysis of structural proteomes allowed authors to discount convergent models of structural

evolution in favor of a specific divergent view that includes both organismal and structural evolution.

An important consequence of this study was the observed correlation between the Darwinian divergent evolution of organisms and the evolution of proteins. Such correlation couples two basic biological scales—microscopic (proteins) and macroscopic (organisms). The fact that these scales are coupled suggests a truly scale-free evolution of molecules and organisms. It also signifies of the single unifying law that governs evolution of proteins and organisms.

Reconstruction of Protein Structure-Function Relation

Most evolutionary models concede some form of relationship between protein structure and function.[125,126] Understanding this relationship is central not only to evolutionary biology, but also to structural genomics.[127] However, despite many efforts, the establishment of a clear relationship between structure and function has been elusive. This is in part due to the fact that neither structural nor functional relationships are well defined. The structure-function relationship between proteins can be understood in light of an evolutionary prospective. Two views on protein evolution have previously been suggested to account for the uneven distribution of sequences in fold space: that of convergent evolution[17,128,129] and divergent evolution.

Convergent evolution posits that different folds evolved independently and the same ("most popular") protein structures are recycled many times by proteins having different functions.[17] According to this model, new proteins may not be related by evolution to their orthologues as new proteins with similar function are rediscovered anew in many organisms. New proteins spawn by chance, but some structures are more populated than others because they are suggested to be more advantageous (thermodynamically, kinetically, evolutionarily). Such a scenario would suggest little relationship between structure and function.[130]

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.[131] These offspring can differ significantly from each other, either in sequence or structure, and can perform a varied array of functions many generations later.[132] 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 relationship between structure and function.[8]

It is important to note that divergent evolution implies a structure-function relationship that mirrors the structural hierarchy of PDUG. As protein structures diverged from progenitor proteins, so did functions. This relationship is necessitated by the requirement that the protein domain and all its descendants remain both functional[133] and structurally stable during the progression of evolution. Since evolution of function is similar in spirit and timeframe to that of structure, the structure-function relationship can also 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 Gene Ontology (GO) consortium.[134]

The main result of reference 135 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 at certain levels of functional annotation provided by GO. If we relax the Z_{min} threshold, we can also see an influx of protein domains into structurally similar clusters. These newly joined domains do not destroy the functional fingerprint of these clusters. 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

The simplistic divergent evolution model[8] that explains the nonrandom behavior of the PDUG is based solely on the premise that a protein has an ancestor that is its closest structural homologue. This model fits the data 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 argue from our divergent evolution model that the older clusters and proteins are more populated and have more connections in PDUG. Of course, there is a significant stochastic component evolution of proteins that may drastically affect both family populations and their connectivity.

To detect mutual evolution between structure and function, in reference 135 we independently annotated proteins based on their function. By considering the function of all the proteins that are annotated and disregarding sequence homologues, we found that proteins have, in general, diverse functional descriptors. These descriptors are unique such as Methionine synthase, b12-binding domains or methylmalonyl-coa-mutase. 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 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 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 consortium.[134] 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. 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. In reference 135 we independently mapped protein function onto the whole of PDUG.

In order to carry out a completely machine based annotation, we used a direct mapping of the genes found in SwissProt Database that coded for the PDB entry of the protein domain in PDUG. We mapped the SwissProt entries to the curated annotation of SwissProt by the Gene Ontology Consortium. Each such annotation was mined independently by the GO consortium primarily from literature searches (http://www.geneontology.org). This yielded a nontrivial mapping from PDB to GO, thus giving each protein its functional assignment. The assignment is nontrivial because some SwissProt entries had many functional annotations corresponding to large, multifunctional, 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.[136] Flow of structure can

happen when proteins A and B share a common domain C. Proteins A and B could then have highly nonrandom structural similarity, but different functions due to the noncommon domain being active. This way, domains may be erroneously classified as functionally equivalent while this may not actually be the case.

Divergent Evolution Observed

In reference 135 we presented strong evidence for divergent evolution of structure and function in protein domains by relating proteins' structures to their functions. We observed a homogeneity of function within structural clusters: the functional fingerprint. Functional fingerprints differ between the structural clusters. We observed the phenomenon of older, more populated clusters diffusing more in the functional space than newer, less populated clusters. For example, the largest structural cluster mainly populated by proteins with the distinctive Rossman fold, is mainly localized in the guanile nucleotide binding GO annotation.

When we considered less populated and therefore presumably younger clusters, we observed 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. Immunoglobulin (Ig) folds are very specialized folds performing mostly B-cell receptor functions. Interestingly, globins localize 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 probability was found to be of the order of 10^{-80}.

It has been known for a long time that there are specialized folds. For example, the Ig folds are known to perform immunity/defense functions, and it is not surprising to find that its functional annotation differed significantly from all other structural clusters. We still found significant homogeneity even in more ubiquitous and less specialized folds such as TIM barrel and Rossman. Using the analysis of reference 135 it may be possible to identify a fold family by deciphering its functional fingerprint.

In reference 135 as we attached newly diverged protein domains to their ancestral clusters, the proteins attached with matching functional descriptions and complemented the functional fingerprint of their ancestors. We also noticed that there are functional categories that are more populated in the PDUG. These were the functions of many structurally similar, but sequentially different proteins. We therefore asked why some functions are more redundant than other ones. We speculated that these are the older functions (those that older proteins started with) that evolved much earlier and consequently have close descendant proteins performing similar function.

The PDUG functional annotations[135] revealed an interesting phenomenon related to the origin of orphans: as we increased the amount of structural evolutionary time, that we controlled by decreasing the threshold Z_{min} (higher Z_{min} represents a more recent snapshot of evolution), the orphans join ancestral clusters. Approximately half of the proteins are not orphans even at $Z_{min} = 9$. As we decreased Z_{min} from 9 to 2, we found that a half of the orphans join ancestral clusters, however the nucleus of the functional annotation within each cluster also grew almost proportionally. The functional nucleus is the collection of nonhomologous 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 "nucleus" can be found and where we observe a more random distribution of functional annotation across all levels of GO. Notably, as we decreased Z_{min}, many functions peripheral to the nucleus diffuse into the fingerprint.

Conclusion

Refinement of the methodologies of protein structure determination yielded a massive amount of important information about protein structure. Due to the fundamental developments in the field of molecular evolution, this information unveiled a peculiar picture of the protein structural, sequence, and functional spaces. In particular, graph-theoretical approaches enable us to decipher specific characteristics of these spaces.

It has been suggested[29] that protein thermodynamics is one of the important evolutionary driving forces that shape the protein sequence space and govern the architecture of the protein structural space. This force relates protein sequence and structural spaces.

One striking observation is the scale-free organization of the PDUG[8]—protein structural space—which is signified by hierarchical relations between structurally similar proteins. The emergence of power-law scaling of the PDUG connectivity $\wp(k)$ is the result of evolutionary dynamics that is as robust at the scale of specific proteomes or at the scale of all organisms. The correlation between structural organization of proteomes and appearance of new organisms (speciation)[124] also suggest a truly universal "scale-free" evolutionary dynamics, whereby the appearance of new protein fold families is parallel to appearance of new species.

Distributions of function and structure over the PDUG act as two evolutionary lenses. It is evident that the evolution of structure and function is mutual and governed by the same underlying principles.[135] Since according to divergent evolution, aside from the biochemical consideration of function structure correlation, there is also 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 and further identify a possible function of putative proteins.

References

1. Levitt M, Chothia C. Structural patterns in globular proteins. Nature 1976; 261:552-558.
2. Koonin EV, Wolf YI, Karev GP. The structure of the protein universe and genome evolution. Nature 2002; 420:218-223.
3. Orengo CA, Jones DT, Thornton JM. Protein superfamilies and domain superfolds. Nature 1994; 372:631-634.
4. Murzin AG, Brenner SE, Hubbard T et al. Scop - A structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995; 247:536-540.
5. Feller W. An introduction to probability theory and its applications. 1968.
6. Yanai I, Camacho CJ, Delisi C. Prediction of gene family distributions in microbial genomes: Evolution by gene duplication and modification. Phys Rev Lett 2000; 85:2641-2644.
7. Qian J, Luscombe NM, Gerstein M. Protein family and fold occurrence in genomes: Power-law behaviour and evolutionary model. J Mol Biol 2001; 313:673-681.
8. Dokholyan NV, Shakhnovich B, Shakhnovich EI. Expanding protein universe and its origin from the biological Big Bang. Proc Natl Acad Sci USA 2002; 99:14132-14136.
9. Karev G, Wolf Y, Rzhetsky A et al. Birth and death of protein domains: A simple model of evolution explains power law behavior. BMC Evol Biol 2002; 2:18.
10. Ponting CP, Russell RR. The natural history of protein domains. Annu Rev Biophys Biomol Struct 2002; 31:45-71.
11. England JL, Shakhnovich EI. Structural determinant of protein designability. Phys Rev Lett 2003; 90:art-218101.
12. England JL, Shakhnovich BE, Shakhnovich EI. Natural selection of more designable folds: A mechanism for thermophilic adaptation. Proc Natl Acad Sci USA 2003; 100:8727-8731.
13. Finkelstein AV, Gutun AM, Badretdinov AY. Why are the same protein folds used to perform different functions. FEBS Lett 1993; 325:23-28.
14. Govindarajan S, Goldstein RA. Why are some protein structures so common? Proc Natl Acad Sci USA 1996; 93:3341-3345.

15. Li H, Helling R, Tang C et al. Emergence of preferred structures in a simple model of protein folding. Science 1996; 273:666.
16. Rykunov DS, Lobanov MY, Finkelstein AV. Search for the most stable folds of protein chains: III. Improvement in fold recognition by averaging over homologous sequences and 3D structures. Proteins 2000; 40:494-501.
17. Taverna DM, Goldstein RA. The distribution of structures in evolving protein populations. Biopolymers 2000; 53:1-8.
18. Buchler NEG, Goldstein RA. Surveying determinants of protein structure designability across different energy models and amino-acid alphabets: A consensus. J Chem Phys 2000; 112:2533-2547.
19. Tiana G, Shakhnovich B, Dokholyan NV et al. Imprint of evolution on protein structures. Proc Natl Acad Sci USA, 2004; 101:2846-2851.
20. Teichmann SA, Murzin AG, Chothia C. Determination of protein function, evolution and interactions by structural genomics. Curr Opin Struct Biol 2001; 11:354-363.
21. Dodson G, Wlodawer A. Catalytic triads and their relatives. Trends Biochem Sci 1998; 23:347-352.
22. Duman JG, Li N, Verleye D et al. Molecular characterization and sequencing of antifreeze proteins from larvae of the beetle Dendroides canadensis. J Comp Physiol [B] 1998; 168:225-232.
23. Duman JG. Antifreeze and ice nucleator proteins in terrestrial arthropods. Annu Rev Physiol 2001; 63:327-357.
24. Makarova KS, Grishin NV. Thermolysin and mitochondrial processing peptidase: How far structurefunctional convergence goes. Protein Sci 1999; 8:2537-2540.
25. Makarova KS, Grishin NV. The Zn-peptidase superfamily: Functional convergence after evolutionary divergence. J Mol Biol 1999; 292:11-17.
26. Chothia C, Hubbard T, Brenner S et al. Protein folds in the all-beta and all-alpha classes. Annu Rev Biophys Biomol Struct 1997; 26:597-627.
27. Wallace AC, Borkakoti N, Thornton JM. TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. Protein Sci 1997; 6:2308-2323.
28. Russell RB. Detection of protein three-dimensional side-chain patterns: New examples of convergent evolution. J Mol Biol 1998; 279:1211-1227.
29. Dokholyan NV, Shakhnovich EI. Understanding hierarchical protein evolution from first principles. J Mol Biol 2001; 312:289-307.
30. Abkevich VI, Gutin AM, Shakhnovich EI. Specific nucleus as the transition-state for protein-folding - evidence from the lattice model. Biochemistry 1994; 33:10026-10036.
31. Fersht AR. Nucleation mechanisms in protein folding. Curr Opin Struct Biol 1997; 7:3-9.
32. Dokholyan NV, Buldyrev SV, Stanley HE et al. Identifying the protein folding nucleus using molecular dynamics. J Mol Biol 2000; 296:1183-1188.
33. Murzin AG. How far divergent evolution goes in proteins. Curr Opin Struct Biol 1998; 8:380-387.
34. Pankov R, Yamada KM. Fibronectin at a glance. J Cell Sci 2002; 115:3861-3863.
35. Lupas AN, Ponting CP, Russell RB. On the evolution of protein folds: Are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? J Struct Biol 2001; 134:191-203.
36. Russell RB, Ponting CP. Protein fold irregularities that hinder sequence analysis. Curr Opin Struct Biol 1998; 8:364-371.
37. Russell RB. Domain Insertion. Protein Eng 1994; 7:1407-1410.
38. Muller HJ. Bar Duplication. Science 1936; 83:528-530.
39. Ohno S. Evolution by Gene Duplication. Springer-Verlag: Berlin, 1970.
40. Ohno S, Wolf U, Atkin NB. Evolution from fish to mammals by gene duplication. Hereditas 1968; 59:169-187.
41. Gerstein M, Levitt M. A structural census of the current population of protein sequences. Proc Natl Acad Sci USA 1997; 94:11911-11916.
42. Qian J, Stenger B, Wilson CA et al. PartsList: A web-based system for dynamically ranking protein folds based on disparate attributes, including whole-genome expression and interaction information. Nucl Acids Res 2001; 29:1750-1764.

43. Orengo CA, Bray JE, Buchan DWA et al. The CATH protein family database: A resource for structural and functional annotation of genomes. Proteomics 2002; 2:11-21.
44. Holm L, Sander C. Protein-structure comparison by alignment of distance matrices. J Mol Biol 1993; 233:123-138.
45. Getz G, Vendruscolo M, Sachs D et al. Automated assignment of SCOP and CATH protein structure classifications from FSSP scores. Proteins 2002; 46:405-415.
46. Sprinzak E, Margalit H. Correlated sequence-signatures as markers of protein-protein interaction. J Mol Biol 2001; 311:681-692.
47. Dietmann S, Fernandez-Fuentes N, Holm L. Automated detection of remote homology. Curr Opin Struct Biol 2002; 12:362-367.
48. Russell RB, Sasieni PD, Sternberg MJE. Supersites within superfolds. Binding site similarity in the absence of homology. J Mol Biol 1998; 282:903-918.
49. Irving JA, Whisstock JC, Lesk AM. Protein structural alignments and functional genomics. Proteins 2001; 42:378-382.
50. Brocchieri L, Karlin S. Conservation among HSP60 sequences in relation to structure, function, and evolution. Protein Sci 2000; 9:476-486.
51. Bradley P, Kim PS, Berger B. TRILOGY: Discovery of sequence-structure patterns across diverse proteins. Proc Natl Acad Sci USA 2002; 99:8500-8505.
52. Andrade MA, Perez-Iratxeta C, Ponting CP. Protein repeats: Structures, functions, and evolution. J Struct Biol 2001; 134:117-131.
53. Jones S, Thornton JM. Prediction of protein-protein interaction sites using patch analysis. J Mol Biol 1997; 272:133-143.
54. Gajiwala KS, Burley SK. HDEA, a periplasmic protein that supports acid resistance in pathogenic enteric bacteria. J Mol Biol 2000; 295:605-612.
55. Hwang KY, Chung JH, Kim SH et al. Structurebased identification of a novel NTPase from Methanococcus jannaschii. Nat Struct Biol 1999; 6:691-696.
56. Stec B, Yang HY, Johnson KA et al. MJ0109 is an enzyme that is both an inositol monophosphatase and the 'missing' archaeal fructose1,6-bisphosphatase. Nat Struct Biol 2000; 7:1046-1050.
57. Shakhnovich BE, Harvey JM, Comeau S et al. ELISA: Structurefunction inferences based on statistically significant and evolutionarily inspired observations. BMC Bioinformatics 2003; 4:34.
58. Lakey JH, Raggett EM. Measuring protein-protein interactions. Curr Opin Struct Biol 1998; 8:119-123.
59. Legrain P, Wojcik J, Gauthier JM. Protein-protein interaction maps: A lead towards cellular functions. Trends Genet 2001; 17:346-352.
60. Valencia A, Pazos F. Computational methods for the prediction of protein interactions. Curr Opin Struct Biol 2002; 12:368-373.
61. Fields S, Song OK. A novel genetic system to detect protein protein interactions. Nature 1989; 340:245-246.
62. Rain JC, Selig L, De Reuse H et al. The protein-protein interaction map of Helicobacter pylori. Nature 2001; 409:211-215.
63. Schuck P. Reliable determination of binding affinity and kinetics using surface plasmon resonance biosensors. Curr Opin Biotechnol 1997; 8;498-502.
64. Doyle ML. Characterization of binding interactions by isothermal titration calorimetry. Curr Opin Biotechnol 1997; 8:31-35.
65. Ahmadian MR, Hoffmann U, Goody RS et al. Individual rate constants for the interaction of Ras proteins with GTPase-activating proteins determined by fluorescence spectroscopy. Biochemistry 1997; 36:4535-4541.
66. Gavin AC, Bosche M, Krause R et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 2002; 415:141-147.
67. Ho Y, Gruhler A, Heilbut A et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 2002; 415:180-183.
68. Back JW, de Jong L, Muijsers AO et al. Chemical cross-linking and mass spectrometry for protein structural modeling. J Mol Biol 2003; 331:303-313.
69. Zhu H, Bilgin M, Bangham R et al. Global analysis of protein activities using proteome chips. Science 2001; 293:2101-2105.

70. Tong AHY, Drees B, Nardelli G et al. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science 2002; 295:321-324.
71. Gaasterland T, Ragan MA. Microbial genescapes: Phyletic and functional patterns of ORF distribution among prokaryotes. Microb Comp Genomics 1998; 3:199-217.
72. Pellegrini M, Marcotte EM, Thompson MJ et al. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc Natl Acad Sci USA 1999; 96:4285-4288.
73. Bono H, Okazaki Y. Functional transcriptomes: Comparative analysis of biological pathways and processes in eukaryotes to infer genetic networks among transcripts. Curr Opin Struct Biol 2002; 12:355-361.
74. Tamames J, Casari G, Ouzounis C et al. Conserved clusters of functionally related genes in two bacterial genomes. J Mol Evol 1997; 44:66-73.
75. Dandekar T, Snel B, Huynen M et al. Conservation of gene order: A fingerprint of proteins that physically interact. Trends Biochem Sci 1998; 23:324-328.
76. Overbeek R, Fonstein M, D'Souza M et al. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci USA 1999; 96:2896-2901.
77. Marcotte EM, Pellegrini M, Thompson MJ et al. A combined algorithm for genome-wide prediction of protein function. Nature 1999; 402:83-86.
78. Marcotte EM, Pellegrini M, Ng HL et al. Detecting protein function and protein-protein interactions from genome sequences. Science 1999; 285:751-753.
79. Enright AJ, Iliopoulos I, Kyrpides NC et al. Protein interaction maps for complete genomes based on gene fusion events. Nature 1999; 402:86-90.
80. Tsoka S, Ouzounis CA. Prediction of protein interactions: Metabolic enzymes are frequently involved in gene fusion. Nat Genet 2000; 26:141-142.
81. Goh CS, Bogan AA, Joachimiak M et al. Coevolution of proteins with their interaction partners. J Mol Biol 2000; 299:283-293.
82. Pazos F, Valencia A. In silico two-hybrid system for the selection of physically interacting protein pairs. Proteins 2002; 47:219-227.
83. Chothia C. Proteins - 1000 Families for the Molecular Biologist. Nature 1992; 357:543-544.
84. Finkelstein AV, Badretdinov AY, Gutin AM. Why do protein architectures have boltzmann-like statistics. Proteins 1995; 23:142-150.
85. Finkelstein AV, Gutin A, Badretdinov A. Why are some protein structures so common? FEBS Lett 1993; 325:23-28.
86. Davidson AR, Sauer RT. Folded proteins occur frequently in libraries of random amino-acid-sequences. Proc Natl Acad Sci USA 1994; 91:2146-2150.
87. Rost B. Protein structures sustain evolutionary drift. Fold Des 1997; 2:S19-S24.
88. Chothia C, Gerstein M. Protein evolution - How far can sequences diverge? Nature 1997; 385:579.
89. Grishin NV. Estimation of evolutionary distances from protein spatial structures. J Mol Evol 1997; 45:359-369.
90. Holm L. Unification of protein families. Curr Opin Struct Biol 1998; 8:372-379.
91. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 1991; 9:56-68.
92. Flaherty KM, Mckay DB, Kabsch W et al. Similarity of the 3-dimensional structures of actin and the atpase fragment of A 70-Kda heat-shock cognate protein. Proc Natl Acad Sci USA 1991; 88:5041-5045.
93. Holmes KC, Sander C, Valencia A. A new ATP-binding fold in actin, hexokinase and Hsc70. Trends Cell Biol 1993; 3:53-59.
94. Orengo CA, Michie AD, Jones S et al. CATH - a hierarchic classification of protein domain structures. Structure 1997; 5:1093-1108.
95. Dodge C, Schneider R, Sander C. The HSSP database of protein structure sequence alignments and family profiles. Nucl Acids Res 1998; 26:313-315.
96. Sanchez R, Pieper U, Melo F et al. Protein structure modeling for structural genomics. Nat Struct Biol 2000; 7:986-990.
97. Pearl FMG, Lee D, Bray JE et al. Assigning genomic sequences to CATH. Nucl Acids Res 2000; 28:277-282.
98. Holm L, Sander C. An evolutionary treasure: Unification of a broad set of amidohydrolases related to urease. Proteins 1997; 28:72-82.

99. Reeck GR, de Haen C, Teller DC et al. "Homology" in proteins and nucleic acids: A terminology muddle and a way out of it. Cell 1987; 50:667.
100. Fitch WM. Distinguishing homologous from analogous proteins. Syst Zool 1970; 19:99-113.
101. Goldstein RA, Lutheyschulten ZA, Wolynes PG. Optimal protein-folding codes from spin-glass theory. Proc Natl Acad Sci USA 1992; 89:4918-4922.
102. Shakhnovich EI, Gutin AM. Engineering of stable and fast-folding sequences of model proteins. Proc Natl Acad Sci USA 1993; 90:7195-7199.
103. Abkevich VI, Gutin AM, Shakhnovich EI. Improved design of stable and fast-folding model proteins. Fold Des 1996; 1:221-230.
104. Shakhnovich EI. Theoretical studies of protein-folding thermodynamics and kinetics. Curr Opin Struct Biol 1997; 7:29-40.
105. Bryngelson JD, Wolynes PG. Spin-glasses and the statistical-mechanics of protein folding. Proc Natl Acad Sci USA 1987; 84:7524-7528.
106. Abkevich VI, Gutin AM, Shakhnovich EI. Theory of kinetic partitioning in protein folding with possible applications to prions. Proteins 1998; 31:335-344.
107. Shakhnovich EI. Protein design: A perspective from simple tractable models. Fold Des 1998; 3:R45-R58.
108. Altschuh D, Vernet T, Berti P et al. Coordinated amino-acid changes in homologous protein families. Protein Eng 1988; 2:193-199.
109. Thomas DJ, Casari G, Sander C. The prediction of protein contacts from multiple sequence alignments. Protein Eng 1996; 9:941-948.
110. Mirny LA, Shakhnovich EI. Universally conserved positions in protein folds: Reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 1999; 291:177-196.
111. Pazos F, Helmer-Citterich M, Ausiello G et al. Correlated mutations contain information about protein-protein interaction. J Mol Biol 1997; 271:511-523.
112. Axe DD, Foster NW, Fersht AR. Active barnase variants with completely random hydrophobic cores. Proc Natl Acad Sci USA 1996; 93:5590-5594.
113. Rowsell S, Pauptit RA, Tucker AD et al. Crystal structure of carboxypeptidase G(2), a bacterial enzyme with applications in cancer therapy. Structure 1997; 5:337-347.
114. Chevrier B, Schalk C, Dorchymont H et al. Crystal-structure of aeromonas-proteolytica aminopeptidase - A prototypical member of the cocatalytic zinc enzyme Family. Structure 1994; 2:283-291.
115. Dietmann S, Holm L. Identification of homology in protein structure classification. Nat Struct Biol 2001; 8:953-957.
116. Sedgewick R. Algorithms in C. MA: Addison-Wesley, Reading, 1990.
117. Havlin S, Benavraham D. Diffusion in disordered media. Adv Phys 1987; 36:695-798.
118. Stauffer D, Aharony A. Introduction to percolation theory. Philadelphia, 1994.
119. Bollobas B. Random graphs. London: Academic Press, 1985.
120. Albert R, Barabasi AL. Statistical mechanics of complex networks. Reviews of Modern Physics 2002; 74:47-97.
121. Finkelstein AV, Ptitsyn OB. Why do globular-proteins fit the limited set of folding patterns. Prog Biophys Mol Biol 1987; 50:171-190.
122. Vendruscolo M, Dokholyan NV, Paci E et al. Small-world view of the amino acids that play a key role in protein folding. Phys Rev E Stat Nonlin Soft Matter Phys 2002; 65:061910.
123. Deeds EJ, Dokholyan NV, Shakhnovich EI. Protein evolution within a structural space. Biophys J 2003; 85:2962-2972.
124. Deeds EJ, Shakhnovich B, Shakhnovich EI. Proteomic traces of speciation. J Mol Biol 2004; 336:695-706.
125. Aravind L, Koonin EV. Gleaning nontrivial structural, functional and evolutionary information about proteins by iterative database searches. J Mol Biol 1999; 287:1023-1040.
126. Jordan IK, Kondrashov F, Rogozin I et al. Constant relative rate of protein evolution and detection of functional diversification among bacterial, archaeal and eukaryotic proteins. Genome Biol 2001; 2:research0053.
127. Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001; 294:93-96.

128. Li H, Tang C, Wingreen NS. Are protein folds atypical? Proc Natl Acad Sci USA 1998; 95:4987-4990.
129. Csete ME, Doyle JC. Reverse engineering of biological complexity. Science 2002; 295:1664-1669.
130. Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 2001; 307:1113-1143.
131. Brenner SA. Natural progression. Nature 2001; 409:459.
132. Ponting CP, Russell RB. Identification of distant homologues of fibroblast growth factors suggests a common ancestor for all beta-trefoil proteins. J Mol Biol 2000; 302:1041-1047.
133. Cooper VS, Schneider D, Blot M et al. Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of Escherichia coli B. J Bacteriol 2001; 183:2834-2841.
134. Ashburner M, Ball CA, Blake JA et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000; 25:25-29.
135. Shakhnovich BE, Dokholyan NV, Delisi C et al. Functional fingerprints of folds: Evidence for correlated structurefunction evolution. J Mol Biol 2003; 326:1-9.
136. Schug J, Diskin S, Mazzarelli J et al. Predicting gene ontology functions from ProDom and CDD protein domains. Genome Res 2002; 12:648-655.