Letter

Protein structure and evolutionary history determine sequence space topology

Boris E. Shakhnovich,[1] Eric Deeds,[2] Charles Delisi,[1] and Eugene Shakhnovich[3,4]

[1]Bioinformatics Program, Boston University, Boston, Massachusetts 02215, USA; [2]Department of Molecular and Cellular Biology and [3]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA

Understanding the observed variability in the number of homologs of a gene is a very important unsolved problem that has broad implications for research into coevolution of structure and function, gene duplication, pseudogene formation, and possibly for emerging diseases. Here, we attempt to define and elucidate some possible causes behind the observed irregularity in sequence space. We present evidence that sequence variability and functional diversity of a gene or fold family is influenced by quantifiable characteristics of the protein structure. These characteristics reflect the structural potential for sequence plasticity, i.e., the ability to accept mutation without losing thermodynamic stability. We identify a structural feature of a protein domain—contact density—that serves as a determinant of entropy in sequence space, i.e., the ability of a protein to accept mutations without destroying the fold (also known as fold designability). We show that (log) of average gene family size exhibits statistical correlation ($R^2 > 0.9$.) with contact density of its three-dimensional structure. We present evidence that the size of individual gene families are influenced not only by the designability of the structure, but also by evolutionary history, e.g., the amount of time the gene family was in existence. We further show that our observed statistical correlation between gene family size and contact density of the structure is valid on many levels of evolutionary divergence, i.e., not only for closely related sequence, but also for less-related fold and superfamily levels of homology.

Gene family and domain-fold family sizes are known to vary widely (Finkelstein and Ptitsyn 1987; Finkelstein et al. 1995; Orengo et al. 1999; Teichmann et al. 1999; Yanai et al. 2000; Vitkup et al. 2001; Koonin et al. 2002)—from orphans (families that have only a single member) to considerably populated sets of far-diverged homologs. The observed variability in the number and divergence of gene family members raises many questions, e.g., which genetic mechanisms and evolutionary dynamics could have led to the observed unevenness? Evolutionary biologists have proposed models designed to explain these size distributions (which often follow power laws) (Yanai et al. 2000; Dokholyan et al. 2002; Koonin et al. 2002; Deeds et al. 2003), while assuming no inherent physical differences between gene families from the outset (Huynen and van Nimwegen 1998; Qian et al. 2001; Dokholyan et al. 2002; Koonin et al. 2002). However, many of these models are overly abstract to adequately explain family size distributions in a constructive manner that relate specific features of gene families with their reported size. Neither do these models provide explicit insights into the mechanistic details that might explain observed differences. On the other hand, some researchers have hypothesized that the heterogeneity in family size is due to an underlying distribution of biological or physical properties (Finkelstein et al. 1995; Govindarajan and Goldstein 1996; Li et al. 1996; Taverna and Goldstein 2000; Koehl and Levitt 2002; Miller et al. 2002) of proteins encoded by gene sequences, but until now, such properties could only be hypothetically characterized for a limited class of simplified two-dimensional and three-dimensional lattice models.

In particular, in a recent study, Taverna and Goldstein (2000) analyzed the contribution from various factors, such as evolutionary history and fold designability, to the development of uneven protein family sizes in simplified two-dimensional lattice models. These authors modeled several scenarios of evolution and demonstrated that more "designable" structures indeed feature more populated (or overpopulated) sequence families. Interestingly, they find that the relationship between designability of a structure (defined in their model as a number of sequences that can have nondegenerate ground state in that structure) and the size of the family exhibits a noticeable scatter indicative of the influence of evolutionary history on the observable outcome (Taverna and Goldstein 2000).

Recent successes in structural genomics and bioinformatics provide a wealth of data for statistical analysis of the distributions of gene family sizes of real proteins with known structures. On the other hand, recent research in our lab and others has increased our understanding of the structural determinants of protein designability (Wolynes 1996; Shakhnovich 1998; England and Shakhnovich 2003), and has made it possible to analyze the structural features of real protein domains that might be responsible for the observed inequality of gene family sizes. Obtaining new insights into the relative roles of physical and biological factors that contribute to the genesis of modern gene families may bring us closer to a greater understanding of the natural history of protein domains.

From a biological perspective, we may hypothesize that gene family size is at least in part influenced by functional constraints related to the number of different, but perhaps related functions needed by the cell (Lespinet et al. 2002). For example, some functions such as kinase activity have varied specificities within a relatively small number of sequence mutations (Manning et al. 2002), while others such as globins have much less-functional flexibility despite, in some cases, substantial sequence divergence (Bashford et al. 1987). From a physical perspective, the potential of a gene to obtain new function upon duplication may depend on its ability to accept mutations without destroying the three-dimensional structure of a protein domain that it

[4]Corresponding author.
E-mail eugene@belok.harvard.edu; fax (617) 384-9228.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.3133605.

encodes. In this work, we will focus mostly on the effect of the physical constraints imposed on the structure encoded by sequences of the gene family. We will show that variability in these constraints represents difference in potential for sequence diversity of gene families. This effect can be observed for real families both on average and in the case of specific families, taking into account their differential time of evolution.

Results

Building PDUG

In order to consider sequence, structure, and function information in a unified, systematic way, we define both gene families and fold families quantitatively using the Protein Domain Universe Graph (PDUG) (Dokholyan et al. 2002). The PDUG is a graph where nodes are sets of closely related sequences folding into structurally characterized domains (Dietmann et al. 2001; Orengo et al. 2003) and edges are connections between the nodes that are based on structure comparison. To separate the gene families, we define the sequences of the domain structures inside each node as having <25% identity to other representative sequences in PDUG. Thus, nodes in PDUG are a set of representative structures. For calculation of sequence family size, we incorporate both SWISS-PROT (Boeckmann et al. 2003) and NRDB90 (Holm and Sander 1998) databases. This enables us to calculate the size of the gene family by using all available sequence data and, at the same time, discounting database bias. We use NRDB to calculate gene family size and SWISS-PROT in combination with Inter-Pro (Apweiler et al. 2000) and GO (Ashburner et al. 2000) to calculate functional divergence for every domain (see Methods). We obtain structures from the Dali Domain Dictionary (Dietmann et al. 2001) and use BLAST (Altschul et al. 1997) and DALI (Holm and Sander 1993) sequence and structure comparison tools (see Methods). Thus, the size of the gene family as represented on PDUG is the number of nonredundant sequences from NRDB that are highly homologous to the representative structure of the domain. Consequently, the size of the structural neighborhood is defined as the number of sequences homologous to a set of structures equivalent to a Fold. (Fig. 1).

Using this PDUG formalism, we can define a gene family based on micro-evolutionary considerations; the PDUG represents the variability accessible to a given gene upon mutation, whether that variability occurs in sequence, function, or structure space. Unlike other definitions of gene families (Sonnhammer et al. 1998; Orengo et al. 2003), we make our definition entirely local, i.e., with respect to a particular gene. The gene family of a gene is therefore all the immediate sequence neighbors of that gene that fold into the same (broadly defined) structure (Baker and Sali 2001; Hegyi and Gerstein 2001). By our construction of the PDUG, a gene family is represented by sequences within a single PDUG node. Analogously, the fold family of a structure is defined as all of the structural neighbors of that domain on PDUG (Fig. 1). By defining the cutoff value for sequence or structure comparison (see Methods), we can control the variance of that attribute, thus implicitly controlling the time scale of evolutionary divergence over which we calculate structure-function determinants.

The role of designability

Our first task is to determine what, if any, physical factors may be responsible for the variability in gene family size. To this end, we define an inherent structural characteristic related to the number of sequences that a structure can accommodate without loss of thermodynamic stability, i.e., we use a structural determinant of designability (Li et al. 1996). This feature has been previously hypothesized (Finkelstein et al. 1995; Li et al. 1996; Miller et al. 2002) to be one of the key influences responsible for overrepresentation of some folds over others. Recent analysis (England and Shakhnovich 2003) suggested that structures with greater values of traces of powers of their contact matrices (CM) (i.e., $Tr[CM]^2$, $Tr[CM]^4$, etc.) are predicted to be more designable (England and Shakhnovich 2003; see Methods). Sequence-space Monte Carlo (England and Shakhnovich 2003) calculations for simple lattice models show that this characteristic of a structure does indeed correlate strongly with its designability, which we define as logarithm of the number of sequences that are stable in the structure.

The physical explanation for the correlation between traces of powers of the CM (a structural feature) and sequence entropy (i.e., designability) follows from the fact that these traces of powers of the CM reflect topological characteristics of the network of contacts within the structure. For example, the trace of CM^2 simply gives the total number of contacts (or equivalently the total number of two-step, self-returning walks) and the trace of CM^4 reflects the number of length-4 closed loops in the system, and so on. One may also note that certain closed sets of contacts allow for optimal placement of amino acids that interact very favorably. For example, if four amino acids that strongly attract each other are folded into an architecture where they all interact favorably (e.g., on four corners of a square, see Fig. 2), this formation represents a greater contribution to the stability of the overall structure than configurations in which the same four amino acids are arranged linearly, or in cases where the last of the contacts is out of the contact range (Fig. 2). Such optimal placement of several strongly interacting amino acids allows more sequences to be folded into the structure by relaxing energy constraints for the rest of the sequence. Thus, structures that provide certain features, such as availability of long closed loops of interactions and higher density of contacts per residue, are expected to be able to accommodate a wider variety of different sequences. This qualitative argument is similar in spirit to derivation of Boltzmann distribution in Statistical Mechanics (Landau et al. 1978) and similar to the justification for the "Boltzmann device" used in the derivation of knowledge-based potentials (Finkelstein et al. 1995; Grzybowski et al. 2002) for the study of protein folding and prediction of ligand-binding energies.

For this study, we use the trace of the second order of the contact matrix normalized by chain length as a simplest approximation for designability. This quantity, known as the contact density (CD), is proportional to the number of contacts per amino acid residue (see Methods); it corresponds to the lowest second-order term in the expansion of equation 1. A designability criterion at this level of approximation has been considered earlier by several authors (Wolynes 1996; Shakhnovich 1998), and these studies predicted that the number of contacts, along with other factors such as dispersion of interaction energies, as well as the proportion of long- and short-range contacts in a structure, may play an important role in determining the designability of a structure.

Correlation between CD and sequence family size on average

First, we calculate the CD for every representative domain structure in PDUG as a measure of the designability of that node. We

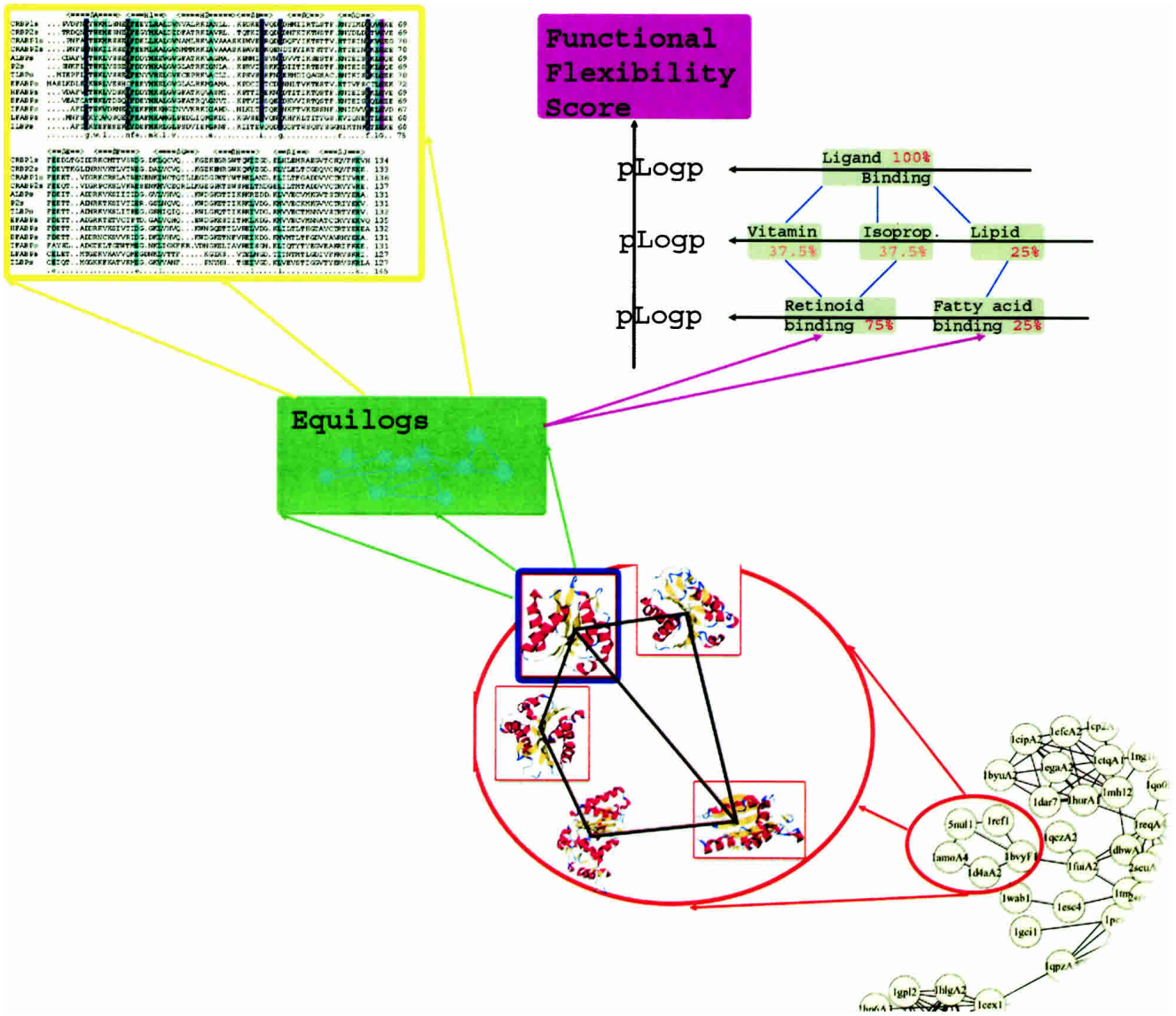

Figure 1. A schematic picture of the scaled organization and intrinsic properties of the protein domain universe graph. The PDUG is built hierarchically, so that each level of evolutionary divergence can be considered independently. The domain structures are compared with each other using DALI (see Methods), and from this information, the structural graph is created (Dokholyan et al. 2002). All of the sequences from NRDB with >25% identity to the original sequence of each domain on PDUG are collected into a gene family. All of the equilogs (sequences with the same function) (Apweiler et al. 2000) matching the gene family are collected and used to create a probabilistic GO tree, from which the FFS is calculated using equation 2. As an example of how to build a structural neighborhood, consider the domain inside the blue rectangle, then all of the domains with red rectangles are its structural neighbors.

define a gene family as the set of sequences with >25% identity to the sequence of the crystallized structure of the domain, excluding close sequence homologs using NRDB90 (Holm and Sander 1998). Clearly, this calculation is predicated on the assumption that SWISS-PROT and NRDB represents a fair estimate of the variability inside each gene family. Remarkably, we observe that there is a marked positive correlation between a domain's designability calculated via CD and the average gene family size (Fig. 3A). However, we note that the observed correlation, while very pronounced, is nonetheless statistical in nature; each point in Figure 3 is a bin in (log) family size that contains 100–250 domains with a distribution of CD values, and the distributions in different bins overlap. Regardless of this caveat, we find that, on average, gene families that encode more designable protein structures are statistically the ones that perform more varied functions (Apweiler et al. 2000), encode more sequences, and therefore, constitute larger families.

Next, we want to assess the robustness of the average correlation, as well as estimate the area of sequence space affected by designability. Structural determinants may influence small areas of sequence space, such as those evaluated in Figure 3A, or larger ones, defined by fold-level structure comparison. In part, the area influenced will depend on how CD changes with respect to divergence in structure. Thus, we perform analysis on distantly related gene families as defined through structural comparison between nodes on PDUG. To this end, we take the structural

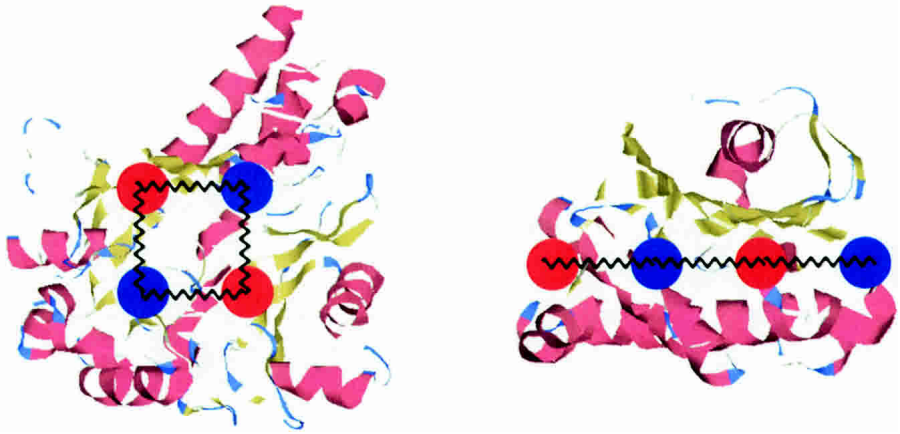

Figure 2. An illustration of physical reasons for differences in designability between two structures. The balls schematically represent amino acids. Suppose that the interaction between the "red" amino acid and the "blue" amino acid is favorable and gives E = −1. The configuration on the *left* yields lower energy −4, compared with *right* structures, where contribution from interactions between these amino acids is only −3. Thus, the 4-loop in the *left* structure contributes more to the stability of the structure overall, allowing more freedom to select the remaining part of the sequence to obtain overall stabilization of the structure. Similar considerations apply to 3-loops, 5-loops, etc.

neighborhood of a given domain to be all nodes that are connected by an edge on the PDUG (Dokholyan et al. 2002; Fig. 1) (see Methods). We then look at the correlation between the number of sequences that fold into all structures belonging to the same structural neighborhood and the average CD for those structures. This approach evaluates a larger, fold-level area of sequence space and correlates sequence variability with designability.

Figure 3B shows that average CD, which serves as a proxy for average designability of a structural neighborhood (Fold), itself correlates with the (log) of the gene-sequence family size of that neighborhood. Together, Figure 3, A and B, show that gene family size and designability (as approximated by CD) correlate across various scales of evolutionary distance. This could indicate that designability affects large sequence and structure spaces spanning not only close sequence homology, but extending into sets of sequences with identifiable homology only through structure comparison. From an evolutionary standpoint, this may indicate that domains with higher CD diverge to produce other high-designability domain structures.

Since these observations of correlations between designability and gene family size are statistical in nature, we want to comment on the robustness of the reported results. There are two issues to consider, the variability of contact density (CD) for structures within gene families and the robustness in the calculation of the mean number of sequences for all gene families in each bin. To address these concerns, we first calculate the intrafamily deviation in CD for each gene family on PDUG (see Methods). While the points in Figure 3, A and B, show mean values of the CD for the representative domains (nodes on PDUG), we also include estimates of the deviation in CD, taking into account sequences inside gene families with solved structures (i.e., domains that have sequence homology to the representative domain). In order to calculate this deviation, we take all solved structures for domains with sequence homology to the representative domain and calculate the standard deviation of CD inside each gene family. We then calculate the average standard deviation in every bin of Figure 3, A and B. The deviation is shown as CD-axis (X) error bars in Figure 3, A and B. It is apparent from the size of the error bars that the deviation in CD within each gene family is relatively small, on the order of 0.05 or less. Indeed, as expected, the intrafamily dispersion deviation of CD gets smaller as average contact density increases. The CD deviation ranges from 0.01 at CD = 4.8 to 0.06 at CD = 3.8 in Figure 3A. The deviation is much smaller when considering domains inside fold-level structural neighborhoods, i.e., the deviation falls to be on the order of 0.001. This calculation is primarily meant to show that the choice of the representative structure for each gene family size is not expected to significantly affect the results.

Next, we calculate the possible error in the calculation of the mean in the size of the gene family for each bin. This quantity is proportional to the square root of the number of observations in the bin according to Central Limit Theorem. We include this as the gene family size (Y) axis error bars in Figure 3. It is worth noting that this measures the deviation of the mean over all gene families belonging to a given bin only, and does not reflect the scatter of the distribution inside the bin. That quantity is considered separately in detail, later. Clearly, the consideration of both of these errors is small enough so that it does not affect the conclusions drawn from Figure 3, A and B. It is also worth noting that, as the size of the error bars suggests, changing the binning does not appreciably affect these results. However, it is important to point out that even considering all the possible caveats mentioned above, the correlation between CD and average sequence variability on both the domain and fold levels is striking, and the error bars show the surprising level of robustness of these results.

For a more biological perspective, we determine how gene family size is related to the diversity of functions which that family performs. We define the functional determinant of a gene family as entropy in function space. When we calculate this measure in the context of PDUG, we utilize Gene Ontology (GO) (Ashburner et al. 2000) to define the functional variability (functional flexibility score or FFS) of a set of genes (see Methods). FFS is a measure of the total amount of information needed to describe all of the functionality of a gene family. Perhaps not surprisingly, FFS statistically correlates with CD (Fig. 4). This is not surprising because FFS statistically correlates with the total number of sequences in a gene family (data not shown). However, this analysis serves two purposes. First, the correlation of FFS and CD shows that designability directly affects the underlying biology of the domain. Domains with low CD have a much lower chance of performing many different functions. Secondly, this serves as a corroboration of the previous result using a different database, annotation method, and a completely different measure of sequence variability. Finally, the correlation of FFS instead of just simply calculations of gene family size ensures that we measure entropy on sequences that are sufficiently diverged to yield different functions, thus minimizing the effect of database bias.

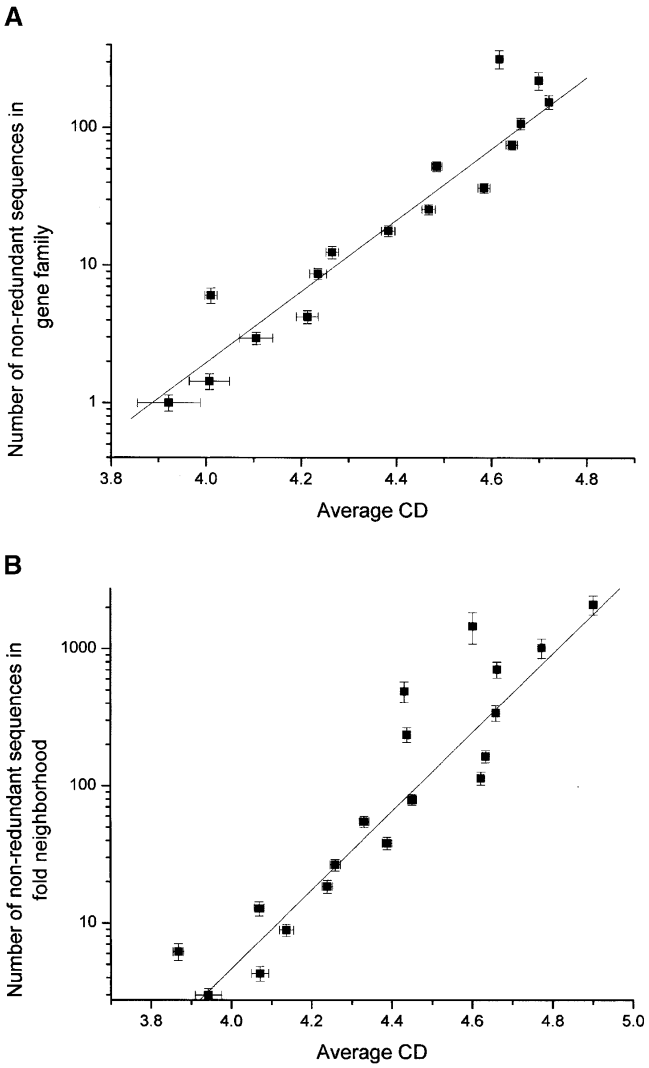

Figure 3. (*A*) The plot of the logarithm of average gene family size vs. the structural contact density parameter calculated for the structures encoded by these sequences (as explained in Methods). Each point represents a bin in log (gene family size), with a step size of ~0.35. Each bin contains 100–250 families. Binning in (log) of gene family sizes provides the advantage of having an approximately equal number of gene families in each bin. The statistical correlation of the linear fit is R = 0.95 with $P < 0.001$. The error bars on the CD axis represent the average deviation of CD inside each gene family averaged for all families belonging to the bin (see Methods). The error bars on the vertical axis correspond to the deviation of the mean number of members for each gene family inside the bin. (*B*) The correlation between the average CD of the structural neighborhood as defined on the PDUG (Fig. 1) and the log of the family sizes of all the sequences inside that neighborhood. Here, R = 0.95 with $P < 0.001$. The error bars are calculated as described for *A*.

The role of evolution

While average statistical correlations of gene family size and FFS with CD are highly significant, how predictive are they when it comes to calculations of gene family size for a particular domain? To answer this question, we present a scatter plot of gene family size versus CD that shows all domains in the PDUG (Fig. 5A). Though the scatter reveals significance, it is clear that CD is not a reliable predictor of gene family size for every domain. This is perhaps not surprising, given that other factors may have influenced gene family sizes. A natural possibility that has also been observed in lattice simulations (Taverna and Goldstein 2000) is that evolutionary history may influence gene family size. Longer evolutionary time of divergence means a higher chance of finding a suitable sequence mutation, thus increasing the gene family size.

Understanding the evolutionary history of all of the protein domains on the PDUG requires construction of the most parsimonious scenario for protein structure evolution, a complex proposition (Mirkin et al. 2003) that is beyond the scope of this work. The simplest construction that still yields useful information is the delineation of the very old domains. Any domain shared by the three kingdoms of life can be placed in the last universal common ancestor (LUCA) (Mirkin et al. 2003). If any such domain were not placed in the LUCA, multiple independent discovery (or horizontal transfer) events would be required to explain the occurrence of this domain in all kingdoms. The "extra" evolution involved in this case would result in a less parsimonious scenario. Inclusion of other domains is more probabilistic and depends on the exact form and method of parsimony construction used (Mirkin et al. 2003).

We thus define the structural content of the LUCA to be all domains that have homologs in at least one prokaryotic and at least one eukaryotic species. This yields approximately a third of the structural content of PDUG. We present the LUCA domains on a separate scatter plot in Figure 5B. Two observations are immediately apparent. First, LUCA domains clearly feature greater CD, suggesting that "first" domains were more designable (difference of means 0.27, *t*-test *P*-value < 1e-8). Secondly, even at equal CD (designability) with their younger counterparts, LUCA domains feature greater family sizes, on average 37 more members (Fig. 5B, scatter plot is markedly shifted toward higher gene family sizes *P*-value < 1e-10). This observation provides evidence that, as simulations on simple lattice models suggest (Taverna and Goldstein 2000; Deeds et al. 2003), designability is only the potential for larger family size that has to be coupled with other mitigating factors for a full understanding of the evo-

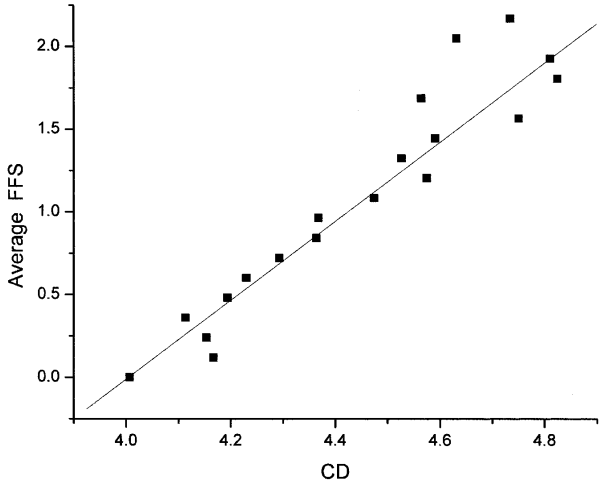

Figure 4. The correlation between CD and functional flexibility score (FFS) of the gene family calculated via equilogs using equation 1. This is evidence that structural determinant of designability, CD, serves as a direct influence on the number of functions that a gene family does, with linear fit correlation R = 0.97. Each datapoint represents a bin in FFS, with step 0.1 containing 50–200 families. The datapoints represent the average CD over all gene families represented in an FFS bin.

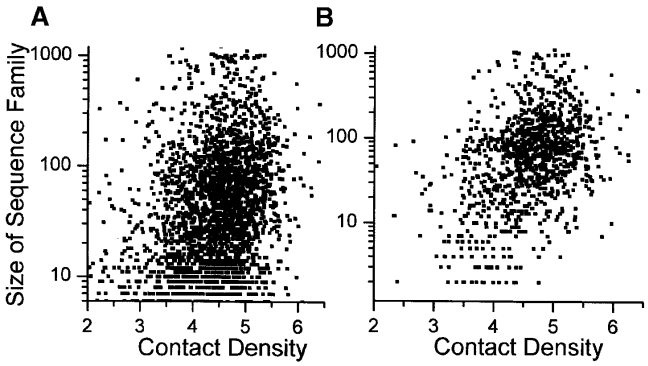

Figure 5. (*A*) Scatter plot of gene family size vs. CD for all PDUG domains. (*B*) The Scatter plot of only the domains that are shared between eukaryotes and prokaryotes (1117 domains). Note that these domains are statistically more designable (higher CD, difference of means 0.27, *P*-value < 1e-8) and that at the same CD, their families are more populated, on average 37 more members in each family. (*P*-value < 1e-10). Linear fit to all domains vs. gene family size returns R = 0.30 with $P < 0.001$, while linear fit to only LUCA domains vs. gene family size returns R = 0.40 with $P < 1e\text{-}4$. Random resampling of LUCA domains yields mean R = 0.30 and standard deviation 0.025. The LUCA domains are not statistically overrepresented in genomes, so the sampling bias is not expected to account for the difference in family sizes.

lutionary history of that domain. For two domains with the same CD, but differing times of divergence, the domain with the longer divergence time will most likely have more sequence members.

To avoid circularity in the calculation of gene family size difference, we calculate the average number of genomes where LUCA domains are present and compare that to the background distribution of all domains. We find that LUCA domains are not present in a significantly larger number of genomes (data not shown); however, they do exhibit a statistically significant increase in gene family size as outlined above. Furthermore, we see the importance of designability even within LUCA domains by noting that higher CD domains exhibit higher gene family size within LUCA. To underline this observation further, we calculate the linear fit for all domains (R = 0.30) and compare that with the LUCA domains. We observe that the goodness of fit (R) increases to 0.40. We tested the statistical significance of this increase by modeling a random assignment of LUCA domains. We randomly picked the same number of domains from PDUG and calculated the linear fit (R value). We then repeated the sampling 1000 times. Predictably, we find that the mean R value from random simulation is 0.30, as the background distribution and standard deviation is 0.025. From this simple experiment, we can conclude that LUCA domains represent a biased sample, where the linear fit of the correlation between CD and sequence family size is four standard deviations away from random. The increase in the goodness of the linear fit for the LUCA domains is consistent with our theory that given the same amount of time for divergence, higher CD domains will have larger sequence families. However, the result mainly outlines the importance of evolutionary history in fulfilling the potential for sequence family size defined by the structural designability of that family. The increase in the linear fit (R) also underlines the independence of this result from bias stemming from uneven genome distribution of LUCA domains.

Discussion

In this study, we presented evidence that across widely varying evolutionary distances, there are significant statistical correlations between structural designability, functional flexibility, and gene family size. The statistical nature of these observations is obvious from the scatter plot presented in Figure 5. We have found that this scatter may be explained, at least in part, by variations in the evolutionary history (Ponting and Russell 2002) of protein domains. Because of this, neither CD nor any other proxy calculation of designability can be used as a predictor of gene family size. As shown by simulation (Taverna and Goldstein 2000) and our own analysis presented in this study, designability represents only the "potential" for sequence entropy allowed by a structure. The actual size depends not only on the potential, but also on the amount of time that evolution had to explore the sequence space around that structure. This, in part, reconciles the very strong correlation of the means observed in Figure 3 and the significant scatter of the specific observations in Figure 5A, while decreasing in scatter observed in Figure 5B lends evidence to the importance of evolutionary history in determining sequence family size.

While we believe that these results are illuminating, we must mention several caveats. Using CD as a proxy for entropy in sequence space is an approximation that assumes, among other things, that protein energetics may be correctly represented in contact form and that the second-order approximation of equation 2 is sufficient to capture the designability of a structure. An additional and perhaps more interesting caveat to consider is that the "designability principle" in its canonical form assumes equilibrium in sequence space, in which all structures take full advantage of their designability potential and that this fact is reflected in the data. Consideration of phylogeny clearly shows that this is not an entirely valid assumption. On the other extreme, several dynamic divergent evolution models predict uneven fold populations without assuming any structural preferences due to designability (Dokholyan et al. 2002), positing that gene family sizes may be due to pure chance in the complex natural history of protein domains. Our observations are not inconsistent with divergent evolution. In fact, we have done simulations that indicate that a combination of divergent evolution models and designability yield a stunning correspondence with observed phenomena (Tiana et al. 2004).

In this work, we clearly see that domains with low CD are most likely to represent smaller size families, while more designable, higher CD domains may exhibit both large and small family sizes. This is exactly what one would expect from the interplay of historical and physical factors; while physical constraints impose upper bounds on sizes of families of low-CD domains, more designable domains may exhibit greater family sizes if they are "old," and smaller sizes if they are "young." Higher designability thus reflects the potential for higher family size, but does not necessarily imply it.

Another interesting observation is that older domains seem more designable. One may speculate that early protein evolution could have imposed more stringent constraints on domain designability, either due to more challenging conditions (e.g., higher temperature) (England et al. 2003) or due to insufficient time to effectively search sequence space to make it possible to select viable sequences for less-designable structures. We show in further studies on lattice models that evolution progresses to-

ward lower designability (Tiana et al. 2004) in accordance with empirical results presented in this work.

The findings presented here may have broad implications for our understanding of structural genomics as well as structure-function relationships and coevolution. However, more quantitative evolutionary models are required to fully rationalize our findings. Further research along these lines may provide new insights into the genetic mechanisms underlying both neofunctionalization and the potential development of resistance to emerging diseases. These results provide an example of how fundamental physical principles can be statistically predictive in the biological Universe of protein folds and gene sequences.

Methods

PDUG

In order to build the PDUG, we use sequences from NRDB90 (Holm and Sander 1998) and all structural domains from HSSP (Dietmann et al. 2001). We use BLAST (Altschul et al. 1997) sequence homology to find all sequences in NRDB90 with >25% sequence identity to each HSSP domain. We combine that set of sequences into a single gene family. We then use cross-indexing between SWISS-PROT (Boeckmann et al. 2003) and InterPro to find the set of all equilogs (sequences with the same function) (Apweiler et al. 2000) belonging to every gene family. We use those equilogs to reconstruct the FFS using equation 2. (see Fig. 1) We use DALI (Holm and Sander 1993) to make all pairwise structural comparisons, and we build structural neighborhoods as described in the text and in Figure 1. For this study, we use Dali $Z_c = 9$ as the cutoff value at which we consider two domains to be structural neighbors, although we believe that changing this value will not drastically alter the results, as evidenced by the correlation between domains and FFS (Fig. 3A). We choose $Z_c = 9$ because this level of structural divergence corresponds roughly to the superfold level of SCOP. Further justification of this threshold selection is given in Dokholyan et. al. (2002).

An important issue in this study is one of sequence weighting. The use of NRDB to exclude close sequence homologs ensures that we calculate sequence entropy by including far diverged sequences. The calculations of FFS provide another corroboration with the same result, but a different weighting of sequences. Inclusion of all sequences from SWISS-PROT will introduce noise due to oversequencing of some genes versus others, and will not yield a sufficient approximation of entropy in sequence space.

Designability

England and Shakhnovich (2003) showed recently that for a large class of amino acid interaction potentials B, the free energy per monomer f in sequence space for a protein structure defined by its contact matrix, (CM) C is given by

$$f = -\frac{1}{N}\sum_{n=2}^{\infty}(\mathrm{Tr}\ C^n)a_n \qquad (1)$$

where the weights a_i are all positive functions that depend on the interaction energies B. The contact matrix C is defined as $C_{ij} = 1$ if amino acids i and j are in contact, and 0 otherwise. Definitions of contact may vary, but in this study, we use the standard cutoff of 7.5 angstroms between C_β atoms (C_α for Gly).

Calculation of variability in CD of intrafamily members

To calculate the variability of designability in Figure 3 (error bars on the x-axis), we considered all solved structures where the sequences are homologous to a representative domain on PDUG. We calculated the CD for all domains inside each sequence family where the number of homologous domains with resolved structures was larger than two. For this calculation, we used the domain boundaries that were delineated for the whole PDB (Westbrook et al. 2003) by Dietmann and Holm (Dietmann et al. 2001). This resulted in consideration of over 34,000 domains in ~3400 nonredundant representative homologous gene families. For each homologous sequence family, we calculated the standard deviation of CD for the structures belonging to that family. We then averaged all calculated standard deviations for gene families falling inside the gene family bin in Figure 3, and represented that quantity as the error bars on the CD axis.

FFS

In order to calculate functional entropy, we begin by combining all sequences into a set. We then match these sequences to InterPro (Apweiler et al. 2000) equilogs. We reconstruct the whole GO tree from the annotations of equilogs and calculate the number of equilogs of the family that are assigned a particular functional annotation, normalized by total number of annotations at each level (see Fig. 1). We may thus calculate the average amount of information per annotation level needed to fully describe the function of each gene family using the following equation:

$$FFS = -\frac{1}{Max(L)}\sum_{l}\sum_{i\in\{\text{nodes on Level } l\}} p_i Log(p_i). \qquad (2)$$

Here, $Max(L)$ 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 percentage of the family that is annotated with function i (see Fig. 1).

Acknowledgments

We are grateful to Jeremy England and Hooman Hennessey for their help, as well as to Nikolay Dokholyan, Andrew Murray, and Nick Grishin for fruitful discussions and critical readings of the manuscript, and to NIH for support.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., et al. 2000. InterPro–An integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **16:** 1145–1150.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25:** 25–29.

Baker, D. and Sali, A. 2001. Protein structure prediction and structural genomics. *Science* **294:** 93–96.

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

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., et al. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31:** 365–370.

Deeds, E.J., Dokholyan, N.V., and Shakhnovich, E.I. 2003. Protein evolution within a structural space. *Biophys. J.* **85:** 2962–2972.

Dietmann, S., Park, J., Notredame, C., Heger, A., Lappe, M., and Holm, L. 2001. A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. *Nucleic Acids Res.* **29:** 55–57.

Dokholyan, N.V., Shakhnovich, B., and Shakhnovich, E.I. 2002. Expanding protein universe and its origin from the biological Big Bang. *Proc. Natl. Acad. Sci.* **99:** 14132–14136.
England, J.L. and Shakhnovich, E.I. 2003. Structural determinant of protein designability. *Phys. Rev. Lett.* **90:** 218101.
England, J.L., Shakhnovich, B.E., and Shakhnovich, E.I. 2003. Natural selection of more designable folds: A mechanism for thermophilic adaptation. *Proc. Natl. Acad. Sci.* **100:** 8727–8731.
Finkelstein, A.V. and Ptitsyn, O.B. 1987. Why do globular proteins fit the limited set of folding patterns? *Prog. Biophys. Mol. Biol.* **50:** 171–190.
Finkelstein, A.V., Gutin, A.M., and Badretdinov, A. 1995. Boltzmann-like statistics of protein architectures. Origins and consequences. *Subcell. Biochem.* **24:** 1–26.
Govindarajan, S. and Goldstein, R.A. 1996. Why are some proteins structures so common? *Proc. Natl. Acad. Sci.* **93:** 3341–3345.
Grzybowski, B.A., Ishchenko, A.V., Shimada, J., and Shakhnovich, E.I. 2002. From knowledge-based potentials to combinatorial lead design in silico. *Acc. Chem. Res.* **35:** 261–269.
Hegyi, H. and Gerstein, M. 2001. Annotation transfer for genomics: Measuring functional divergence in multi-domain proteins. *Genome Res.* **11:** 1632–1640.
Holm, L. and Sander, C. 1993. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233:** 123–138.
———. 1998. Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics* **14:** 423–429.
Huynen, M.A. and van Nimwegen, E. 1998. The frequency distribution of gene family sizes in complete genomes. *Mol. Biol. Evol.* **15:** 583–589.
Koehl, P. and Levitt, M. 2002. Protein topology and stability define the space of allowed sequences. *Proc. Natl. Acad. Sci.* **99:** 1280–1285.
Koonin, E.V., Wolf, Y.I., and Karev, G.P. 2002. The structure of the protein universe and genome evolution. *Nature* **420:** 218–223.
Landau, L.D., Lifshitz, E.M., and Pitaevskii, L.P. 1978. *Statistical physics*. Pergamon Press, Oxford, New York.
Lespinet, O., Wolf, Y.I., Koonin, E.V., and Aravind, L. 2002. The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.* **12:** 1048–1059.
Li, H., Helling, R., Tang, C., and Wingreen, N. 1996. Emergence of preferred structures in a simple model of protein folding. *Science* **273:** 666–669.
Manning, G., Plowman, G.D., Hunter, T., and Sudarsanam, S. 2002. Evolution of protein kinase signaling from yeast to man. *Trends Biochem. Sci.* **27:** 514–520.
Miller, J., Zeng, C., Wingreen, N.S., and Tang, C. 2002. Emergence of highly designable protein-backbone conformations in an off-lattice model. *Proteins* **47:** 506–512.
Mirkin, B.G., Fenner, T.I., Galperin, M.Y., and Koonin, E.V. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.* **3:** 2.
Orengo, C.A., Todd, A.E., and Thornton, J.M. 1999. From protein structure to function. *Curr. Opin. Struct. Biol.* **9:** 374–382.
Orengo, C.A., Pearl, F.M., and Thornton, J.M. 2003. The CATH domain structure database. *Methods Biochem. Anal.* **44:** 249–271.
Ponting, C.P. and Russell, R.R. 2002. The natural history of protein domains. *Annu. Rev. Biophys. Biomol. Struct.* **31:** 45–71.
Qian, J., Luscombe, N.M., and Gerstein, M. 2001. Protein family and fold occurrence in genomes: Power-law behaviour and evolutionary model. *J. Mol. Biol.* **313:** 673–681.
Shakhnovich, E.I. 1998. Protein design: A perspective from simple tractable models. *Fold Des.* **3:** R45–R58.
Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A., and Durbin, R. 1998. Pfam: Multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.* **26:** 320–322.
Taverna, D.M. and Goldstein, R.A. 2000. The distribution of structures in evolving protein populations. *Biopolymers* **53:** 1–8.
Teichmann, S.A., Chothia, C., and Gerstein, M. 1999. Advances in structural genomics. *Curr. Opin. Struct. Biol.* **9:** 390–399.
Tiana, G., Shakhnovich, B.E., Dokholyan, N.V., and Shakhnovich, E.I. 2004. Imprint of evolution on protein structures. *Proc. Natl. Acad. Sci.* **101:** 2846–2851.
Vitkup, D., Melamud, E., Moult, J., and Sander, C. 2001. Completeness in structural genomics. *Nat. Struct. Biol.* **8:** 559–566.
Westbrook, J., Feng, Z., Chen, L., Yang, H., and Berman, H.M. 2003. The Protein Data Bank and structural genomics. *Nucleic Acids Res.* **31:** 489–491.
Wolynes, P.G. 1996. Symmetry and the energy landscapes of biomolecules. *Proc. Natl. Acad. Sci.* **93:** 14249–14255.
Yanai, I., Camacho, C.J., and DeLisi, C. 2000. Predictions of gene family distributions in microbial genomes: Evolution by gene duplication and modification. *Phys. Rev. Lett.* **85:** 2641–2644.

Received August 10, 2004; accepted in revised form November 23, 2004.