PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2008) **275**, 1595–1602
doi:10.1098/rspb.2007.1617
Published online 22 April 2008

Protein robustness promotes evolutionary innovations on large evolutionary time-scales

Evandro Ferrada[1,3,*] and Andreas Wagner[1,2,3]

[1]*Department of Biochemistry, University of Zurich, Building Y27, Winterthurerstrasse 190, 8057 Zurich, Switzerland*
[2]*The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA*
[3]*The Swiss Institute of Bioinformatics*

Recent laboratory experiments suggest that a molecule's ability to evolve neutrally is important for its ability to generate evolutionary innovations. In contrast to laboratory experiments, life unfolds on time-scales of billions of years. Here, we ask whether a molecule's ability to evolve neutrally—a measure of its robustness—facilitates evolutionary innovation also on these large time-scales. To this end, we use protein designability, the number of sequences that can adopt a given protein structure, as an estimate of the structure's ability to evolve neutrally. Based on two complementary measures of functional diversity—catalytic diversity and molecular functional diversity in gene ontology—we show that more robust proteins have a greater capacity to produce functional innovations. Significant associations among structural designability, folding rate and intrinsic disorder also exist, underlining the complex relationship of the structural factors that affect protein evolution.

Keywords: robustness; evolutionary innovations; protein designability; functional diversity

1. INTRODUCTION

What makes a biological system able to produce evolutionary innovations (Müller & Wagner 1991), new adaptations that may aid in survival and reproduction? Do some systems have a greater ability to innovate than others? A rigorous answer to these questions requires a systematic comparison of many different systems and the innovations they have produced. Whole organisms are not readily amenable to such systematic comparison. By contrast, molecular innovations can be more easily studied. This is because we know millions of protein sequences, as well as thousands of structures, and their associated functions. For this reason, here we address the opening questions with protein molecules and their functional diversity, which is a record of past evolutionary innovations.

Recent experimental work suggests that a molecule's ability to evolve neutrally is important for its ability to evolve new functions. Such neutral evolution leaves a primary function of the molecule unchanged, while paving the way for new functions to emerge. Cases in point are the enzymes serum paraoxonase and cytochrome P450. These enzymes have a primary catalytic function, but they can also metabolize other secondary substrates at greatly reduced rates (Amitai *et al.* 2007; Bloom *et al.* 2007). Laboratory evolution experiments show that neutral mutations that do not change the primary function of these enzymes can cause substantial fluctuations in their secondary activities. Natural selection can then rapidly increase these 'promiscuous' activities (Aharoni *et al.* 2005). A different kind of experiment with two catalytic RNA molecules makes a similar point. In this experiment, Schultes & Bartel (2000) mutagenized two ribozymes unrelated in sequence, structure and catalytic activity. These authors created a path of single mutations through sequence space that connected the two ribozymes. After most of the steps in this path, the catalytic activity of the mutated molecules did not change much, except for a small transition region approximately halfway between the two starting molecules. In this region, the activity of one molecule switched to the activity of the other molecule. Here again, neutral mutations paved the way for a molecule with a new function. In both cases, the ability to evolve neutrally facilitated a molecule's ability to acquire functional innovations.

If these observations hold more generally, the following prediction arises for two different molecules A and B: if A can undergo more neutral mutations than B—it has greater mutational robustness than B—then A should also show a greater propensity to evolve new functions. This prediction has been confirmed for cytochrome P450 in another recent experiment (Bloom *et al.* 2006*b*), which showed that thermostable or mutationally robust variants of this enzyme more readily evolve new catalytic activities. A theoretical work on RNA structures provides a larger context and intuitive explanation for this observation (Wagner 2007). Populations of mutationally robust structures can explore a set of all possible genotypes rapidly through neutral mutations. They are thus genotypically diverse and can produce large amounts of structural variation by single point mutations. This increased access to structural diversity promotes evolutionary innovations, even though only a small fraction of structural variants may lead to new functions.

Laboratory experiments can explore evolutionary innovations on laboratory time-scales. However, life unfolded on time-scales of billions of years. Does the connection between robustness and evolutionary innovation hold on these vastly larger time-scales? This is the question we address here. To do so, we need to analyse a protein

*Author and address for correspondence: Department of Biochemistry, University of Zurich, Building Y27, Winterthurerstrasse 190, 8057 Zurich, Switzerland (e.ferrada@bioc.uzh.ch).

Received 12 March 2008
Accepted 2 April 2008

structure's ability to evolve neutrally—its mutational robustness—for many different structures. This ability is directly related to the number of sequences in a genotype space that can fold into a given structure, also known as the designability of the structure. The concept of designability was first coined by Li *et al.* (1996). Using a simple lattice model, these authors showed that the number of sequences that can adopt a given structure is related to the structure's regularity and to its robustness to mutations. Further studies have shown that designability is also related to evolutionary rate (Bloom *et al.* 2006*a*). The sequences folding into a structure are typically connected in large neutral networks (Babajide *et al.* 1997; Bastolla *et al.* 2003).

Here we show that more robust proteins show greater propensity to evolve new functions on vast evolutionary time-scales. To this end, we use quantitative estimates of protein designability that can be determined from a protein's contact density matrix (England & Shakhnovich 2003), or from the diversity of sequences adopting a protein structure (Shakhnovich *et al.* 2005). As a record of past evolutionary innovations, we use the functional diversity of protein domains, as encapsulated in their diversity of enzymatic functions (Pegg *et al.* 2006) and in their gene ontology annotations (Ashburner *et al.* 2000) of molecular functions.

2. MATERIAL AND METHODS

Our main source of data is the class, architecture, topology and homologous superfamily (CATH) protein structure classification database v. 3.1.0 (Greene *et al.* 2007). Here we focus on the 1924 representative protein domains in CATH, which exceed a minimal length of 50 residues. The number of different functions known for a domain depends on the time since a domain originated in evolution: for two domains—one young and another old—with equal designability (robustness), the young domain had less time to accumulate sequence and functional diversity. We exclude this confounding factor by focusing some of our analyses on a subset of ancient domains that are present in all sequenced bacterial, archaeal and eukaryotic genomes (Ranea *et al.* 2006), and that were thus present in the last universal common ancestor of extant life. Since this dataset was derived from a previous CATH release, we filter these domains to obtain 112 ancient domains that occur in the current release.

(a) *Measures of designability*

In our analysis, we use two complementary estimates of a protein fold's designability. We refer to these estimates as structural designability (D_S) and diversity designability (D_D). Structural designability was introduced by England & Shakhnovich (England & Shakhnovich 2003; Shakhnovich *et al.* 2005). These authors showed that the number of sequences that can adopt a given structure is approximated by the length-normalized maximum eigenvalue of the contact density matrix at a defined distance cut-off, based on a coarse-grained structural description (using only C_α and C_β atoms). The contact density matrix $A=(a_{ij})$ is a binary (0-1) matrix, where $a_{ij}=1$ if two residues i and j that are not neighbours ($|i-j|>1$) are in contact. For our purpose, we consider two non-neighbouring residues in contact, if any of their C_α and C_β atoms occur within a 6.0 Å radius of each other. An alternative measure of structural designability is the average number of atomic contacts per residue (England & Shakhnovich 2003; Bloom *et al.* 2006*a*). However, this measure is so closely correlated with D_S (Spearman's $r=0.989$; $p<10^{-100}$) that it yields virtually identical results. We thus focus exclusively on the length-normalized structural designability, D_S.

We obtain our second estimate of designability (D_D) from diversity data of protein sequences, in an approach similar to that of Shakhnovich *et al.* (2005). Specifically, we analyse sequences in the non-redundant dataset NRDB90 (Holm & Sander 1998). We examine each sequence in this set and assign it to an ancient representative CATH domain, if the sequence has 25% or more identity to the CATH representative, as suggested by the analysis of Chothia & Lesk (1986). We use BLAST (Altschul *et al.* 1997) to determine the extent of sequence identity. Since the number of similar sequences observed per representative domain is dependent on its length, we also normalized D_D by the sequence length.

Because designability may be related to the complexity and amount of disorder of a protein fold, we also explored their relationship with functional diversity. As a measure of fold complexity, we used the absolute contact order (ACO) as introduced by Plaxco *et al.* (1998). ACO is the average distance on the amino acid sequence of two residues that contact each other in the structure. Proteins with high ACO fold slowly. We calculate ACO as in Ivankov *et al.* (2003), where we consider two residues to be in contact if any of their C_α or C_β are inside a sphere of 6.0 Å.

To explore intrinsic disorder (ID) in the sequence domain dataset described above, we use the tool IUPred (Dosztányi *et al.* 2005*a*,*b*). Briefly, IUPred estimates for each residue in a sequence an index that indicates the amount of disorder this residue is subject to. We calculate the disorder average for each sequence in the NRDB90 dataset and assign this value to a CATH representative domain if the BLAST comparison shows a per cent identity of the sequence that is greater than 25%. Finally, we simply calculate the average over the whole set of disorder scores assigned to a representative domain.

(b) *Functional annotation*

We estimate the capacity to evolve functional innovations using information from two sources. The first is the structure–function linkage database (SFLD) that associates sequence, structure and functional annotation for a diverse spectrum of enzyme superfamilies. This functional annotation is based on structural similarities of enzyme active sites (Pegg *et al.* 2006). In September 2007, the SFLD contained 6280 protein sequences grouped in 138 families and six superfamilies. We determined the diversity of functions on the family level for all sequences that shared more than 25% identity with any of the CATH representative domains.

We express functional diversity of a domain in two ways. The first (FE_1) is simply the number of different SFLD families assigned per domain and normalized by the domain length. The length-normalization is needed to correct for the fact that the longer the sequence, the higher the chance to find a second sequence that shares 25% of identity. The second (FE_2) is a measure akin to an entropy that takes into account the frequency of different enzymatic functions observed per domain. If a set of sequences associated with a domain has k different associated enzymatic functions (some of which may occur multiple times), and if p_i is the frequency with which each function i occurs in the set of sequences, then $FE_2 = -\sum_{i=1}^{k} p_i \log p_i$.

Table 1. Spearman's rank correlation coefficients. (D_S, structural designability; D_D, diversity designability; FE_1, enzymatic functional diversity; FG_1, diversity of molecular functions (based on gene ontology); FE_2, entropic measure of enzymatic functional diversity; FG_2, entropic measure of molecular functional diversity (GO); ACO, absolute contact order; ID, intrinsic disorder. The upper right triangle shows Spearman's rank correlation coefficients (r). The lower left triangle shows the corresponding p-values. Diversity designability as well as functional diversity measures are reported for the set of highly conserved evolutionary domains.)

	D_S	D_D	FE_1	FG_1	FE_2	FG_2	ACO	ID
D_S	—	0.882	0.702	0.938	0.877	0.973	−0.698	0.923
D_D	7.25×10^{-53}	—	0.801	0.961	0.877	0.868	−0.662	0.897
FE_1	1.09×10^{-30}	2.31×10^{-40}	—	0.818	0.872	0.700	−0.625	0.705
FG_1	2.99×10^{-68}	2.50×10^{-79}	1.69×10^{-42}	—	0.916	0.938	−0.765	0.931
FE_2	7.13×10^{-52}	7.13×10^{-52}	6.40×10^{-51}	5.42×10^{-61}	—	0.886	−0.604	0.889
FG_2	4.08×10^{-88}	3.48×10^{-50}	1.58×10^{-30}	2.99×10^{-68}	1.09×10^{-53}	—	−0.638	0.952
ACO	9.11×10^{-91}	1.14×10^{-27}	3.56×10^{-25}	2.22×10^{-36}	7.25×10^{-24}	5.07×10^{-26}	—	−0.607
ID	$<10^{-100}$	4.07×10^{-56}	6.23×10^{-31}	1.08×10^{-65}	2.50×10^{-54}	2.29×10^{-74}	$<10^{-100}$	—

The second source of functional information used in this study is the GOA database that maps UniProt (The UniProt 2007) entries to gene ontology (GO) terms (Camon *et al.* 2004). We obtained the GOA database from the EMBL-EBI FTP site (ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT), and filtered the complete database to obtain only those UniProt entries that were annotated with molecular functions. We then created a non-redundant database of sequences using the NRDB90 tool (Holm & Sander 1998). Subsequently, we examined each sequence in this database and mapped the associated GO terms to a CATH representative domain, if the sequence shared more than 25% identity with the CATH domain. Analogous to enzymatic diversity, we express molecular functional diversity in two ways. The first (FG_1) is simply the number of different GO molecular functions per domain, normalized by the domain length. The second (FG_2) is the entropy measure described above, but now for the frequency distribution of GO terms observed per representative domain.

(c) *Statistics*

All statistical analyses were carried out with the statistics software R v. 2.1.1 (R Development Core Team 2005; http://www.r-project.org/). For the principal component regression (PCR) analysis, we used the R package 'pls'.

3. RESULTS

(a) *More designable proteins show a greater capacity to produce enzymatic diversity*

Here we use two complementary measures of protein designability. The first of them is structural designability (D_S), as estimated by the length-normalized principal eigenvalue of a protein's contact density matrix (England & Shakhnovich 2003). The contact density matrix $A=(a_{ij})$ is a binary (0-1) matrix, where $a_{ij}=1$ if two non-neighbouring residues i and j ($|i-j|>1$) are in contact. The principal eigenvalue of the contact density matrix tends to be larger for proteins with more amino acid contacts per residue, adopting a value between the average number of contacts per residue and the maximal number of contacts of any given residue (Porto *et al.* 2004). The measure D_S reflects the number of groups of interacting amino acids. A large number of such groups allow more sequences to adopt a structure by relaxing energy constraints for the rest of the sequence (Shakhnovich *et al.* 2005).

Our second measure is diversity designability (D_D), which is the number of sequences from a non-redundant database (see §2) that fold into a structure, normalized by the sequence length. This second measure is vulnerable to a confounding factor, the different age of proteins. Old proteins may have more sequences associated with them than younger proteins, just because they originated early in life's evolution. To exclude this factor, we restricted our analysis of diversity designability (D_D) to a set of 112 ancient protein domains in the CATH database, which were probably present in the most recent common ancestor of all extant life (Ranea *et al.* 2006). Both measures of designability are highly correlated for this age-corrected set of domains (Spearman's $r=0.88$; $p<7.25\times10^{-53}$; table 1) and for the complete set of more than 1924 CATH domains (Spearman's $r=0.89$; $p<10^{-100}$; figure 2*a*). Similar associations have been reported for different domain datasets (Shakhnovich *et al.* 2005). They suggest that D_S is reflective of the number of sequences that adopt a structure.

We used two complementary measures of protein functional diversity. The first is a measure of diversity of enzymatic functions, based on structural similarities of enzyme active sites. The relevant information is curated in a recently developed database, which classifies enzymes into three hierarchical levels of function, of which we use the lowest (familial) level here (Pegg *et al.* 2006). We use two quantitative indicators of enzymatic functional diversity. These are FE_1, the number of enzyme families associated with a protein domain, and FE_2, which takes into account that different enzymatic functions occur at different frequencies in a set of sequences associated with a domain (see §2). We explored the association between protein designability and functional diversity for these two different notions of functional diversity.

Figure 1*a* shows an example of two structures with very different designabilities (figure 2*a*). The colour spectrum in the tertiary structure ranges from blue to red, corresponding to positions with low and high sequence diversity (D_D), respectively. The structure in figure 1*a*(i) has lower designability and lower functional diversity, as indicated by the number of associated enzymatic functions, than the structure in figure 1*a*(ii). The less designable domain is associated with two enzyme superfamilies and three families, whereas the more designable domain is associated with four enzyme superfamilies and 11 families. Figure 1*b*

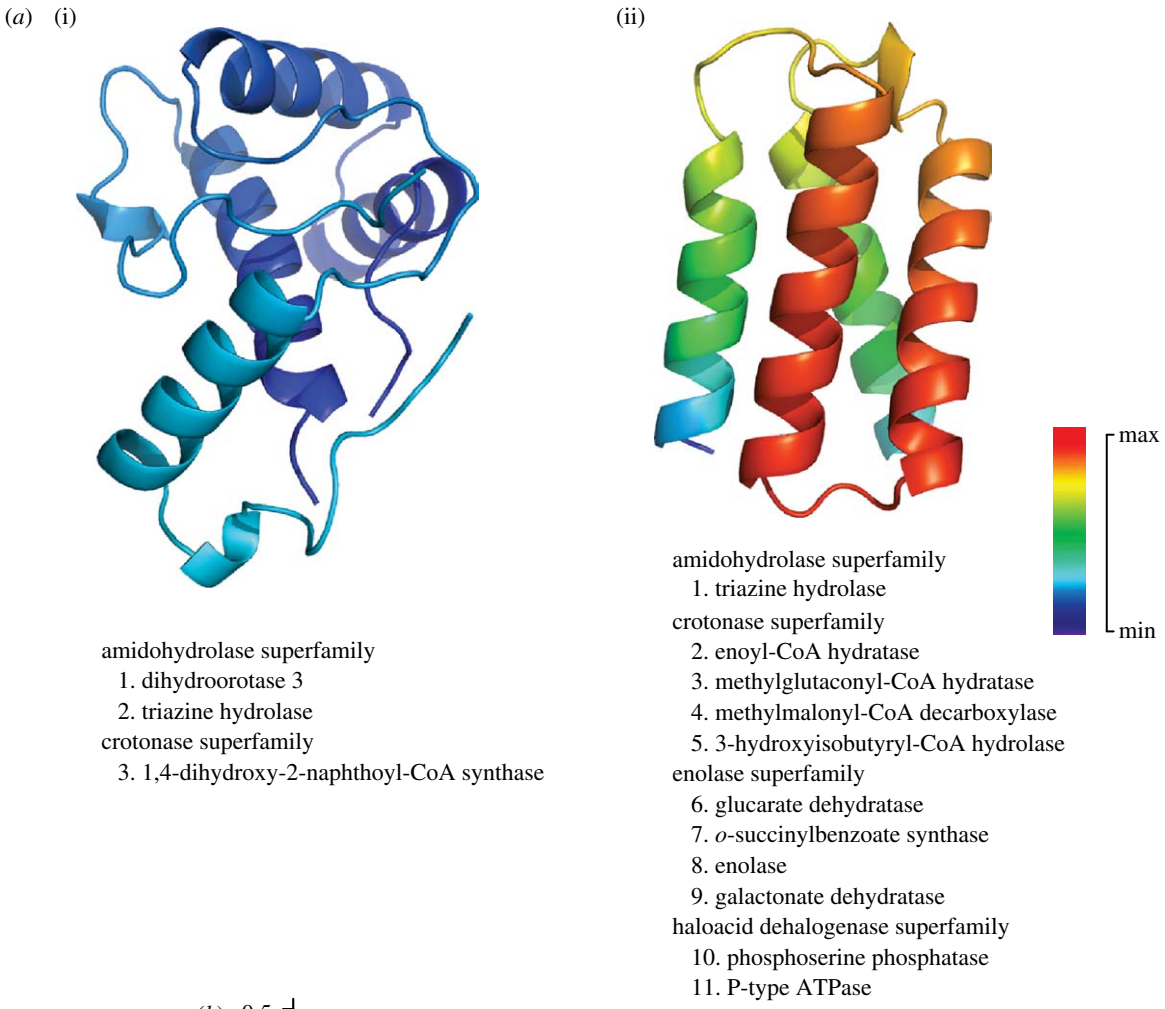

Figure 1. (*Caption opposite.*)

shows a scatterplot of D_D and enzymatic functional diversity (FE_1) for eight arbitrarily chosen ancient structures that are colour-coded in the same way. It suggests that the difference evident from figure 1*a* is not just a peculiarity of the two sequences chosen.

For the complete dataset of ancient domains, we observe a statistically significant and highly positive association between enzymatic functional diversity and D_D (Spearman's $r=0.80$; $p<2.31\times10^{-40}$; figure 3*a*). A structure with more associated sequences might be expected to have more associated functions, but this association persists if we normalize the number of functions by the total number of sequences associated with each fold (Spearman's $r=0.44$; $p<1.95\times10^{-15}$).

Figure 1. (*Opposite*.) (*a*) An example of protein domains with different designabilities and different functional diversities. For the purpose of illustration, the minimum and maximum number of sequences has been scaled linearly. Thus, the colour spectrum indicates a measure of sequence diversity, where blue (red) corresponds to minimum (maximum) sequence diversity estimated per residue. The diversity designability of a domain D_D is a domain-wide average over this sequence diversity. The enzyme families associated with each domain are listed. (i) A domain with low designability (CATH identifier: 1mw9X04: topoisomerase 1, domain 4). It has a complex fold and is associated with three enzyme families that fall into two superfamilies (Pegg *et al.* 2006). (ii) A domain with high designability (1ls1A01: the A subunit of the four-helix bundle hemerythrin domain). It has a simpler fold and is associated with 11 enzyme families and four superfamilies. Superfamilies and families are listed. (*b*) Enzymatic functional diversity (FE_1) increases with protein designability. Enzymatic functional diversity (FE_1) is expressed as the number of different enzyme families per representative CATH domain (Pegg *et al.* 2006). Eight highly conserved CATH domains (1n55A00, 1qz5A01, 1q6zA03, 1rl6A02, 1k7wA03, 1ls1A01, 1vq8V00 and 2bm0A03) have been arbitrarily chosen to illustrate the association between enzymatic functional diversity (FE_1) and designability (D_D, D_S). The Spearman rank correlation coefficient between D_S and FE_1 for these eight domains is 0.92.

We also examined the association between structural designability D_S and enzymatic functional diversity. This association is also positive, regardless of whether we normalize for the number of sequences associated with a fold (Spearman's $r=0.55$; $p<1.24\times10^{-20}$) or not (Spearman's $r=0.70$; $p<1.09\times10^{-30}$; figure 3*b*). An even higher positive association exists if we use the frequency-weighted measure of enzymatic functional diversity, FE_2 (D_D: Spearman's $r=0.88$; $p<7.13\times10^{-52}$; D_S: Spearman's $r=0.88$; $p<7.13\times10^{-52}$).

(b) *More designable proteins show greater overall diversity of molecular functions*

Our second measure of functional diversity encompasses GOAs of molecular functions. The GO database includes the most comprehensive information about functional diversity of proteins. It is not restricted to enzymes. The GO project has developed a dynamic controlled vocabulary based on three aspects of function (molecular function, process and location) that encompass complementary notions of gene functions in living cells (Ashburner *et al.* 2000). For our purpose, the appropriate aspect of function is molecular function. We used two measures of molecular functional diversity. The first (FG_1) is simply the number of molecular function annotations associated with a protein domain and the second (FG_2) weights different functions by their frequency in a set of proteins (see §2).

We observe a statistically significant and highly positive association between functional diversity (FG_1) and D_D, regardless of whether we normalize for the number of sequences per domain (Spearman's $r=0.62$; $p<1.53\times10^{-24}$) or whether we do not normalize (Spearman's $r=0.96$; $p<2.5\times10^{-79}$; figure 3*c*). We also examined the association between D_S and FG_1, which is positive independent of whether the values are normalized (Spearman's $r=0.86$; $p<1.94\times10^{-48}$) or whether they do not normalize (Spearman's $r=0.94$; $p<2.99\times10^{-68}$; figure 3*d*). An even higher positive association exists if we use the frequency-weighted measure of functional diversity, FG_2 (D_D: Spearman's $r=0.87$; $p<3.48\times10^{-50}$; D_S: Spearman's $r=0.97$; $p<4.08\times10^{-88}$).

(c) *Fold complexity and ID influence designability and diversity*

Protein designability may be correlated with a number of other protein properties. Although such properties are not the main focus of our analysis, we wanted to examine how some of them relate to functional diversity. The first of these properties is the complexity of a protein fold. Among various available measures (Arteca 1995; Enright & Leitner 2005), we use the ACO as a measure of fold complexity. ACO is the average distance on the amino acid

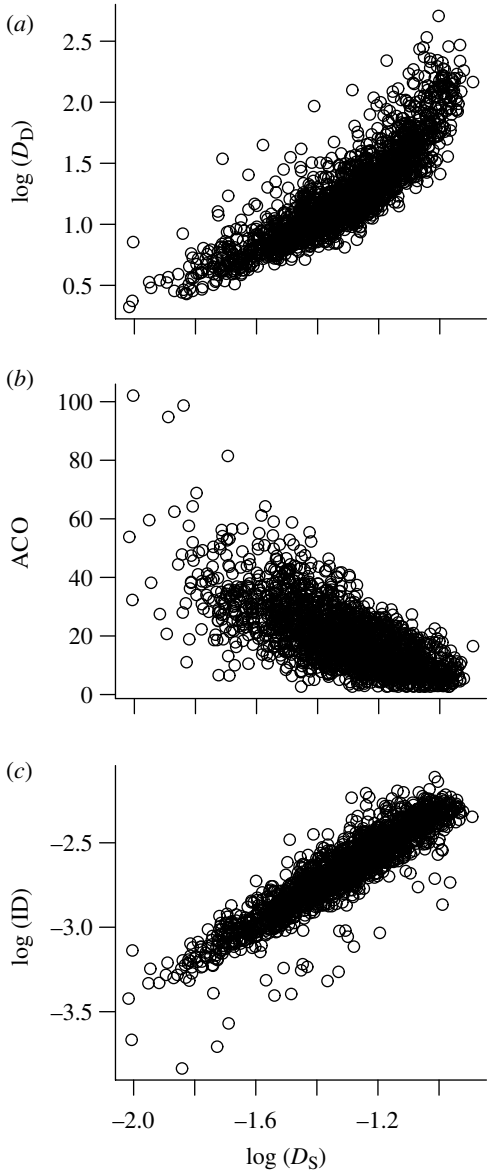

Figure 2. Designability, fold complexity and disorder are associated properties. (*a*) Diversity designability (D_D) versus structural designability (D_S). (*b*) Fold complexity (ACO) versus structural designability (D_S). (*c*) Intrinsic disorder (ID) versus structural designability (D_S). D_D corresponds to the total number of sequences per residue per representative domain. ID is calculated as a length-normalized average per representative domain. Decadic logarithm is applied.

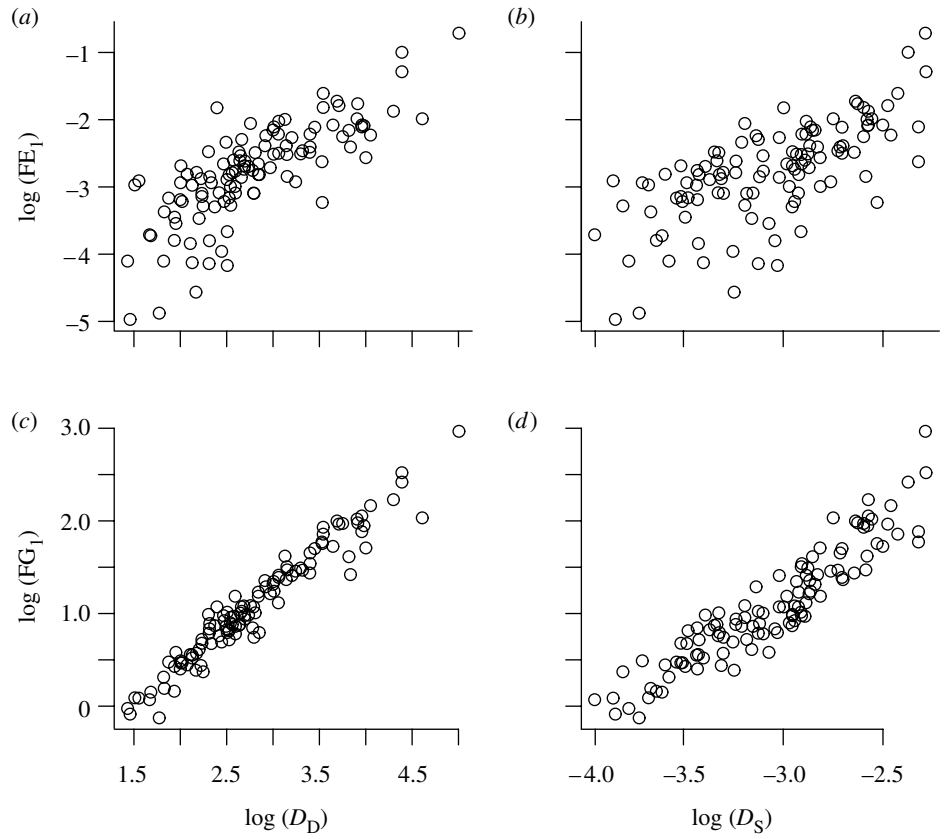

Figure 3. Functionally diverse proteins are highly designable. (*a*) Enzymatic functional diversity (FE_1) as a function of diversity designability (D_D). (*b*) Enzymatic functional diversity (FE_1) as a function of structural designability (D_S). (*c*) Molecular (gene ontology) functional diversity (FG_1) as a function of diversity designability (D_D). (*d*) Molecular (gene ontology) functional diversity (FG_1) as a function of structural designability (D_S). Functional diversity measures shown are normalized by the total number of sequences associated with each representative domain. D_D corresponds to the length-normalized number of sequences per representative domain.

sequence of two residues that contact each other in the structure. It can be thought of as a measure of how 'entangled' a structure is. It is a good predictor of a protein's folding rate, regardless of whether the folding kinetics is dominated by one or several steps (Ivankov *et al.* 2003). Highly designable proteins have low fold complexity. (D_S: Spearman's $r=-0.70$; $p<9.11\times10^{-91}$; D_D: Spearman's $r=-0.66$; $p<1.14\times10^{-27}$; figure 2*b*).

Second, we also explore the relationship between designability and a measure for the amount of conformational disorder a protein can tolerate. Highly disordered proteins are more flexible than others. The measure we use is the 'intrinsic disorder' of a protein, as defined in Dosztányi *et al.* (2005*b*). Specifically, here we use the average ID of the set of sequences associated with each CATH representative domain (see §2). We would predict that proteins with high intrinsic disorder can tolerate more sequence change, and that they might thus also be more designable. This is the case (D_S: Spearman's $r=0.92$; $p<10^{-100}$; D_D: Spearman's $r=0.90$; $p<4.07\times10^{-56}$; figure 2*c*). Not surprisingly, these properties are also associated with each other (table 1).

Because protein fold complexity and disorder are associated with designability, they might also be associated with functional diversity. This is indeed the case (table 1). The diversity of enzymatic and general molecular functions increases for short proteins (FE_1: Spearman's $r=-0.685$; $p<2.33\times10^{-29}$; FG_1: Spearman's $r=-0.94$; $p<1.22\times10^{-68}$), for proteins with low fold complexity (FE_1: Spearman's $r=-0.63$; $p<3.6\times10^{-25}$; FG_1: Spearman's $r=-0.77$; $p<2.22\times10^{-36}$) and for proteins with high intrinsic disorder (FE_1: Spearman's $r=0.71$; $p<6.22\times10^{-31}$; FG_1: Spearman's $r=0.93$; $p<1.1\times10^{-65}$).

The pairwise associations we have discussed so far may conceal subtle interactions among the multiple variables we consider here. To better disentangle their relationship, we thus performed a PCR analysis. This analysis allows us to understand how the three critical variables—designability, fold complexity and disorder—contribute to functional diversity. The results of this analysis reveal no unforeseen new relationships (figure 4). One dominant principal component accounts for more than 80% of the variance in functional diversity. This component is dominated by the positive role of designability and ID for functional diversity and by the negative role of fold complexity (figure 4). The second and third principal components contribute only 15 and 4% of the variance, respectively. Similar results (not shown) hold if diversity designability or enzyme functional diversity is used in the analysis.

4. DISCUSSION

In summary, our observations show that highly designable proteins evolve more functional innovations on large timescales. Our measures of designability estimate a given domain's ability to explore sequence space and access a

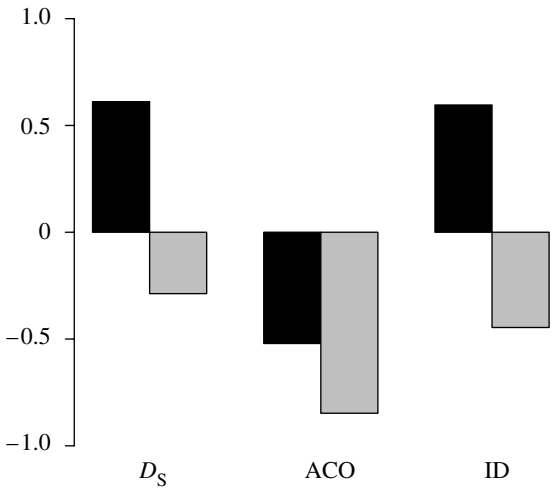

Figure 4. PCR analysis of molecular functional diversity (gene ontology) against structural designability (D_S), fold complexity (ACO) and intrinsic disorder (ID) of folds. Shown are the two principal components that together account for 96.6% of the variance observed for functional diversity. Component 1 (black bars) and component 2 (grey bars) accounts for 80.8 and 15.8% of the total variance, respectively.

diverse spectrum of functions. Because functional diversity is a record of past evolutionary innovations, this means that more designable proteins may have a greater facility to evolve new functions. In addition, because proteins of similar structure are connected in genotype space (Babajide *et al.* 1997, 2001; Bornberg-Bauer 1997; Bastolla *et al.* 1999; Wroe *et al.* 2007), more robust proteins may show greater propensity to evolve functional innovations. This association holds for two complementary measures of functional diversity: diversity of enzymatic functions and GO-based diversity of molecular functions. It also holds for two different measures of designability: one based purely on structural information and the other based on the number of sequences associated with each protein fold. The associations persist if we correct for the different numbers of sequences associated with a fold. For gene ontology annotations, these associations are also corroborated by an analysis based on a different domain dataset (Shakhnovich *et al.* 2005), whose main focus was to explain different sequence family sizes associated with different folds.

A number of other protein properties are associated with designability, and thus, not surprisingly, with functional diversity. Specifically, long proteins, proteins with complex folds (and thus proteins with slower folding rates; Ivankov *et al.* 2003) and proteins with low amounts of disorder in their tertiary structure show low functional diversity. Most of these associations have intuitive explanations. For example, it is easy to see how a high complexity of a fold may lead to smaller numbers of sequences being able to adopt a fold.

With respect to disorder in protein structures, conflicting interpretations can be brought to bear on its relationship to designability. On the one hand, a more disordered structure may be more flexible, and thus tolerate more amino acid changes, implying greater robustness and designability. On the other hand, a disordered structure may be less thermodynamically stable (Dosztányi *et al.* 2005*b*) and greater thermodynamic stability has been associated with robustness (Bastolla & Demetrius 2005; Bloom *et al.* 2006*b*). Although explanations that could resolve this conflict have been put forward (Bastolla & Demetrius 2005), such resolution is not within the scope of this contribution.

A caveat to our—and any other—comparative study is that statistical association is not equivalent to causation. Other known features (expression level, domain architecture, etc.) and unknown features of proteins may show hidden associations with functional diversity that may explain some of its variation. To identify such features would be a worthwhile subject of future studies, as would be the reduction of biases in the data, as well as the elimination of errors contained in some measures of structural differences among proteins. For example, the ID estimate we use (IUPred) has a true positive rate of 85% (Dosztányi *et al.* 2005*b*), which could be improved.

Complex relationships with other variables notwithstanding, it is clear that designable and robust proteins have evolved many novel functions. This shows that a pattern derived from recent experimental findings, and applicable only to laboratory time-scales, also holds on vastly greater geological time-scales (Aharoni *et al.* 2005; Bloom *et al.* 2006*b*). The possible explanation has its root in how populations explore vast sequence spaces: populations of highly robust folds can explore sequence space rapidly, and thus access large amounts of structural diversity in their neighbourhood (Wagner 2007). A small fraction of this diversity can subsequently give rise to proteins with new functions.

A.W. acknowledges support through grant 315200-116814 from the Swiss National Foundation, as well as support from the Santa Fe Institute.

REFERENCES

Aharoni, A., Gaidukov, L., Khersonsky, O., McQ Gould, S., Roodveldt, C. & Tawfik, D. S. 2005 The evolvability of promiscuous protein functions. *Nat. Genet.* **37**, 73–76. (doi:10.1038/ng1482)

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402. (doi:10.1093/nar/25.17.3389)

Amitai, G., Gupta, R. D. & Tawfik, D. S. 2007 Latent evolutionary potentials under the neutral mutational drift of an enzyme. *HFSP J.* **1**, 67–78. (doi:10.2976/1.2739115)

Arteca, G. 1995 Scaling regimes of molecular size and self-entanglements in very compact proteins. *Phys. Rev. Lett.* **51**, 2600–2610. (doi:10.1103/PhysRevE.51.2600)

Ashburner, M. *et al.* 2000 Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* **25**, 25–29. (doi:10.1038/75556)

Babajide, A., Hofacker, I. L., Sippl, M. J. & Stadler, P. F. 1997 Neutral networks in protein space: a computational study based on knowledge-based potentials of mean force. *Fold Des.* **2**, 261–269. (doi:10.1016/S1359-0278(97)00037-0)

Babajide, A., Farber, R., Hofacker, I. L., Inman, J., Lapedes, A. S. & Stadler, P. F. 2001 Exploring protein sequence space using knowledge-based potentials. *J. Theor. Biol.* **212**, 35–46. (doi:10.1006/jtbi.2001.2343)

Bastolla, U. & Demetrius, L. 2005 Stability constraints and protein evolution: the role of chain length, composition and disulfide bonds. *Protein Eng. Des. Sel.* **18**, 405–415. (doi:10.1093/protein/gzi045)

Bastolla, U., Roman, H. E. & Vendruscolo, M. 1999 Neutral evolution of model proteins: diffusion in sequence space and overdispersion. *J. Theor. Biol.* **200**, 49–64. (doi:10.1006/jtbi.1999.0975)

Bastolla, U., Porto, M., Eduardo Roman, M. H. & Vendruscolo, M. H. 2003 Connectivity of neutral networks, overdispersion, and structural conservation in protein evolution. *J. Mol. Evol.* **56**, 243–254. (doi:10.1007/s00239-002-2350-0)

Bloom, J. D., Drummond, D. A., Arnold, F. H. & Wilke, C. O. 2006*a* Structural determinants of the rate of protein evolution in yeast. *Mol. Biol. Evol.* **23**, 1751–1761. (doi:10.1093/molbev/msl040)

Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. 2006*b* Protein stability promotes evolvability. *Proc. Natl Acad. Sci. USA* **103**, 5869–5874. (doi:10.1073/pnas.0510098103)

Bloom, J. D., Romero, P. A., Lu, Z. & Arnold, F. H. 2007 Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. *Biol. Dir.* **2**, 17. (doi:10.1186/1745-6150-2-17)

Bornberg-Bauer, E. 1997 How are model protein structures distributed in sequence space? *Biophys. J.* **73**, 2393–2403.

Camon, E. *et al.* 2004 The gene ontology annotation (GOA) database: sharing knowledge in Uniprot with gene ontology. *Nucleic Acids Res.* **32**, D262–D266. (doi:10.1093/nar/gkh021)

Chothia, C. & Lesk, A. M. 1986 The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.

Dosztányi, Z., Csizmok, V., Tompa, P. & Simon, I. 2005*a* IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 3433–3434. (doi:10.1093/bioinformatics/bti541)

Dosztányi, Z., Csizmok, V., Tompa, P. & Simon, I. 2005*b* The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.* **347**, 827–839. (doi:10.1016/j.jmb.2005.01.071)

England, J. L. & Shakhnovich, E. I. 2003 Structural determinant of protein designability. *Phys. Rev. Lett.* **90**, 218101. (doi:10.1103/PhysRevLett.90.218101)

Enright, M. B. & Leitner, D. M. 2005 Mass fractal dimension and the compactness of proteins. *Phys. Rev. E* **71**, 011912. (doi:10.1103/PhysRevE.71.011912)

Greene, L. H. *et al.* 2007 The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res.* **35**, D291–D297. (doi:10.1093/nar/gkl959)

Holm, L. & Sander, C. 1998 Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics* **14**, 423–429. (doi:10.1093/bioinformatics/14.5.423)

Ivankov, D. N., Garbuzynskiy, S. O., Alm, E., Plaxco, K. W., Baker, D. & Finkelstein, A. V. 2003 Contact order revisited: influence of protein size on the folding rate. *Protein Sci.* **12**, 2057–2062. (doi:10.1110/ps.0302503)

Li, H., Helling, R., Tang, C. & Wingreen, N. 1996 Emergence of preferred structures in a simple model of protein folding. *Science* **273**, 666–669. (doi:10.1126/science.273.5275.666)

Müller, G. B. & Wagner, G. P. 1991 Novelty in evolution: restructuring the concept. *Annu. Rev. Ecol. Syst.* **22**, 229–256. (doi:10.1146/annurev.es.22.110191.001305)

Pegg, S. C. *et al.* 2006 Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. *Biochemistry* **45**, 2545–2555. (doi:10.1021/bi052101l)

Plaxco, K. W., Simons, K. T. & Baker, D. 1998 Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–994. (doi:10.1006/jmbi.1998.1645)

Porto, M., Bastolla, U., Roman, H. E. & Vendruscolo, M. 2004 Reconstruction of protein structures from a vectorial representation. *Phys. Rev. Lett.* **92**, 218101. (doi:10.1103/PhysRevLett.92.218101)

Ranea, J. A., Sillero, A., Thornton, J. M. & Orengo, C. A. 2006 Protein superfamily evolution and the last universal common ancestor (LUCA). *J. Mol. Evol.* **63**, 513–525. (doi:10.1007/s00239-005-0289-7)

Schultes, E. A. & Bartel, D. P. 2000 One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science* **289**, 448–452. (doi:10.1126/science.289.5478.448)

Shakhnovich, B. E., Deeds, E., Delisi, C. & Shakhnovich, E. 2005 Protein structure and evolutionary history determine sequence space topology. *Genome Res.* **15**, 385–392. (doi:10.1101/gr.3133605)

The UniProt, C. 2007 The universal protein resource (UniProt). *Nucleic Acids Res.* **35**, D193–D197. (doi:10.1093/nar/gkl929)

Wagner, A. 2007 Robustness and evolvability: a paradox resolved. *Proc. R. Soc. B* **275**, 91–100. (doi:10.1098/rspb.2007.1137)

Wroe, R., Chan, H. S. & Bornberg-Bauer, E. 2007 A structural model of latent evolutionary potentials underlying neutral networks in proteins. *HFSP J.* **1**, 79–87. (doi:10.2976/1.2739116)