39

PROTEIN DESIGNABILITY AND ENGINEERING

Nikolay V. Dokholyan

INTRODUCTION

Rapid scientific and technological advances in biological sciences that started in the second half of the twentieth century have led to exploration and engineering at the scale of single biological molecules. DNA manipulation techniques, such as polymerase chain reaction (PCR), allowed almost arbitrary control of proteins that are expressed in living cells. These developments prompted scientists to ask increasingly ambitious questions pertaining to cell life, evolution, and the molecular origins of diseases. Rational manipulation of protein function through alteration in sequence and structure became an important aspect in the quest to answer these questions.

A protein's sequence uniquely defines its three-dimensional structure. This relationship is perhaps the most central paradigm of protein biophysics. Interestingly, however, a protein structure does not uniquely define a single sequence: multiple sequences often correspond to similar protein structures. This aspect turned out to be central for the field of structural bioinformatics. Similarly, a protein's structure defines its function, although a given structure can have several biological functions. Hence, one approach to manipulate protein function is through its sequence. Nature diversifies protein function through sequence over the course of evolution. This process is often referred to as *protein evolution*. Understanding protein evolution has become an important subject because of the temptation to manipulate protein function through rational alternation in protein structure via its sequence. This process is referred to as *protein design* or *protein engineering*, although the latter term has a grander-scale connotation, best applicable to more complex systems such as protein–protein or protein–DNA complexes.

How can the manipulation of protein function help in understanding cellular life and the molecular origins of diseases? The most obvious applications are direct examinations of how

Structural Bioinformatics, Second Edition Edited by Jenny Gu and Philip E. Bourne

protein function alterations are coupled to phenotypic or symptomatic outcomes. More complex applications include various biosensors engineered to “sense” various target proteins’ states, protein–protein interactions, protein cellular localization, and expression levels. Potential therapeutic applications of protein engineering also include design of novel antibodies and biomarkers to proteins that are elevated during a disease state with respect to the healthy state.

The appeal of protein structure-function manipulation led to a number of studies pertaining to questions about the possibilities and limitations of protein engineering. Protein design is a search procedure, either computational or experimental, that aims to determine amino acid sequences that correspond to a specific structure. Perhaps, the most central question of protein design is how “designable” is a specific structure, that is, how many sequences correspond to a specific protein structure. This question, first addressed from a purely theoretical perspective, has become one of the most intriguing questions in protein evolution and design, and is the subject of this chapter.

Before turning to the discussion of designability, we first review the intrinsic properties of naturally occurring proteins, organization of the protein structural universe, key determinants of protein structure, and protein evolution. We then discuss essential requirements for and other questions pertaining to protein design. Finally, we talk about challenges in protein design.

PROTEIN STRUCTURAL UNIVERSE

An understanding of the range of protein structures that can be adopted by Nature must first be appreciated before undertaking protein engineering efforts. Protein structures are organized hierarchically as discussed in more detail in Chapter 2. Several secondary structure elements—α-helices, β-strands, hairpins, and loops—form domains, quanta of protein structure. The domains can be defined unambiguously from the thermodynamic and folding kinetics points of view. Protein domains must be able to exist on their own, that is, isolated protein domain sequences must be able to reach their native states and be thermodynamically stable. Proteins can be further built by linking multiple domains. One must appreciate the wisdom of Nature in creating such modularity in the protein universe, because modularity permits the combinatorial multiplicity of possible proteins.

The consequence of structural modularity is that protein functions are modular as well. Indeed, let us imagine a need for a protein that would respond to a specific stimulus, cross the nucleus, associate with DNA, and initiate transcriptional response to this stimulus. Such a protein can be built by incorporating the stimulus-sensing domain (e.g., binding a specific small molecule), a nuclear transport domain, a DNA binding domain (e.g., containing a Zn-finger motif), and potentially, a catalytic domain to instigate a. The beauty of this modular design, a strategy used by Nature, is that the need to extensively explore protein sequence and structural space is reduced. Instead, the selection and combination of a few domain architectures often produce a synergistic solution to constructing a protein with a desired biological function.

Further protein assembly into functional complexes with these modular units adds new dimensions to protein function. Thus, protein function is also hierarchical (Anantharaman, Aravind, and Koonin, 2003; Shakhnovich et al., 2003) and can be defined at multiple levels of protein structure: from the domain level to the level of molecular complexes. Perhaps the functional quantum of protein structure is a domain. Hence, it is plausible that evolution of proteins is dictated by the evolution of functional domains.

The question "how many protein folds exist?" has been an extremely hot topic of discussion. A number of answers have appeared, but all of them report the number of folds (N_F) to range between 1000 and 10,000 (Chothia, 1992; Orengo, Jones, and Thornton, 1994; Wolf, Grishin, and Koonin, 2000). How many should we expect if Nature were to exhaustively explore all possibilities? Let us count. A typical $N = 100$ residue protein has $N_{ss} = 15$ secondary structure elements, including strands, helices, loops, and turns. Then theoretically, $N_F = N_{ss}! \approx 10^{12}$ possibilities to arrange these structural elements with respect to each other (Table 39.1). This number is significantly larger than what is observed and even projected. Hence, it is clear that either evolution has not fully explored protein fold space or some folds are not designable, or both.

Importantly, the number of structures N_S that a 100-residue protein can explore is much larger than the number of folds (N_S ? N_F). To estimate N_S, let us position a simplified protein model on a lattice, where each residue is represented as vertices, with the number of adjacent lattice vertices (the coordination number) z (for cubic lattice $z = 6$). The coordination number defines the number of possible interactions between residues. With this model neighboring residues can be positioned in $(z - 1)$ vertices, adjacent to the current residue. Then, there are $(z-1)^{(N-1)} = 10^{99 \log 5} \approx 10^{69}$ possibilities to build a $N = 100$ residue protein structure. We ignore self-avoidance of polymer chains (Grosberg and Khokhlov, 1997), since it is potentially compensated by our lattice considerations. However, only a fraction of 10^{69} possible protein structures will be compact. To estimate the fraction of compact proteins, we refer to a recent work by Chen, Ding, and Dokholyan (2007) who demonstrated that on average $\alpha \approx 1.7$ proximity constrains per residue define a protein structure within 3.5 Å root-mean-square distance from its native structure. Then, the total number of possible structures that a 100-residue proteins can adopt is $N_S \approx 10^{69}/\alpha^{99} \approx 10^{47}$. This number corresponds to the total number of compact 100-residue protein structures, which is also the number of sequences N_Σ that correspond to compact protein structures, that is, the number of protein sequences. Importantly, the total number of possible sequences for a 100-residue

TABLE 39.1. Estimations and Observed Numbers for the Protein Universe for $N = 100$ Residue Long Proteins

	$\log_{10}(N_X)$	
	Estimated	Naturally Observed or Projected
Estimated number of folds, N_F	12	2–3 (Holm and Sander, 1997; Martin et al., 1998;Wolf, Grishin, and Koonin, 2000; Dietmann et al., 2001)
The number of possible structures, N_S	47	5 (Berman et al., 2000)
The number of sequences corresponding to compact structures, N_Σ	47	NA
The total number of possible sequences, N_{Seq}	130	130
The average number of structures per fold, N_Σ/N_F	35	0–1 (Qian, Luscombe, and Gerstein, 2001; Dokholyan, Shakhnovich, and Shakhnovich, 2002)

protein is $20^{100} \approx 10^{130}$, much larger than the number of sequences that correspond to compact proteins.

These estimations suggest that if all folds are equivalent, that is, there are equal numbers of structures that correspond to each fold, then, there are $N_S/N_F = N_\Sigma/N_F \approx 10^{35}$ protein structures per fold. This assumption does not likely hold: Dokholyan, Shakhnovich, and Shaknovich (2002) reported that the number of naturally occurring protein domains in fold families are distributed unevenly with respect to the structures that contain them and follow a power law distribution $P(n_S) \propto n_S^{-2.5}$. While such an uneven distribution may suggest that some domains are more designable than others, it may also be the result of diverging evolutionary processes. We return to the issue of designability later in this chapter.

DETERMINANTS OF PROTEIN DOMAIN EVOLUTION

There are several different evolutionary factors that impact the designability of proteins. At the level of organisms, evolution optimizes adaptation of organisms to new or changing environments. This adaptation is accommodated through modifications of protein function that are achieved through changes of protein sequences and structures, or through changes in the expression levels of the proteins. Protein evolution is determined by the former modifications of protein sequence and consequent structural rearrangements. What are the factors that drive protein evolution? By examining mutational strategies and the effects of mutations in Nature, the examples will help guide our understanding of the underlying design rules that needs to be considered when addressing protein designability and engineering experiments.

Protein evolution proceeds through amino acid substitutions and fragmental rearrangements (deletions, insertions, truncations). These mutations can affect three aspects of proteins: (1) structure, (2) function, and (3) folding kinetics. If a mutation disrupts the folding kinetics of a protein, it may lead to misfolding and aggregation. There are a large number of misfolding-related disorders. Perhaps the most illustrative example is cystic fibrosis (Boucher, 2004) since 90% of all cases are attributed to a single amino acid deletion ΔPhe508 in the cystic fibrosis transmembrane conductance regulator (CFTR; Riordan et al., 1989; Rommens et al., 1989). The mutant protein's thermodynamic stability is not perturbed and it is not functional because it never matures as a transmembrane protein, suggesting that deletion of phenylalanine affects the folding kinetics of this protein. Clearly, this example demonstrates that preservation of protein folding kinetics may be a key factor in protein evolution.

Mutations leading to a loss of protein function can often be lethal. There are numerous examples of mutations that perturb functional sites, thereby affecting organism viability (see, for example, Akiyama et al., 2007; He et al., 2007). A number of loss-of-function mutations are embryonically lethal, especially those that involve proteins that play central roles during development. These facts suggest that the preservation of protein functional sites may also be a key factor in protein evolution.

Mutations that affect the protein stability are also associated with numerous diseases. Examples include familial amyotrophic lateral sclerosis (FALS) caused by over 100 mutations in Cu, Zn superoxide dismutase 1 (SOD1; Cleveland, 1999; Khare, Caplow, and Dokholyan, 2006; Figure 39.1) and familial amyloid polyneuropathy caused by mutations in transthyretin (Benson and Kincaid, 2007). Interestingly, in the case of FALS, many

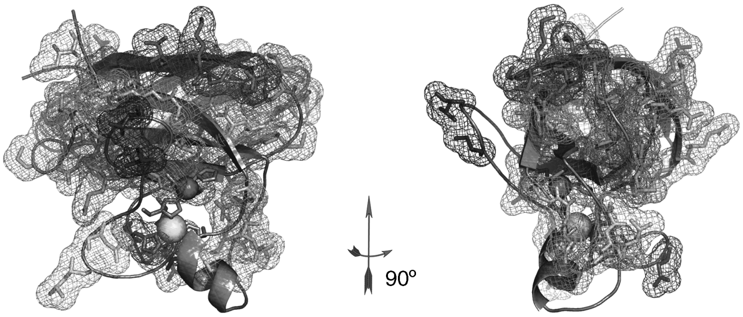

Figure 39.1. Single nucleotide polymorphisms in the human Cu, Zn superoxide dismutase (SOD1) associated with the amyotrophic lateral sclerosis. Over 70 mutations are scattered throughout the tertiary structure affecting multiple structurally distinct sites, marked by colored meshes. For simplicity only one monomer of the homodimeric SOD1 is shown in two projections. Figure also appears in Color Figure section.

SOD1 mutants are as active as wild type, but they potentially gain toxic function due to association into pathogenic oligomeric states due to destabilization. This toxic gain-of-function may be under cellular repair machinery control until this control deteriorates due to aging. If the organisms are at or past the reproductive age when disease manifests itself, they can produce offspring, therefore organism evolution is not affected significantly. Hence, it may seem that mutations that affect protein stability are not as important as those that affect protein folding kinetics or function.

Surprisingly, this is not the case. Dokholyan and Shakhnovich (2001) showed that if one attempts to mimic protein evolution through sequence alterations that preserve protein thermodynamic stability, then amino acid positions that contribute the most to stability appear the most conserved during such alterations. This observation is further supported by examining the conservation pattern obtained in synthetic evolution, which is found to closely match that is observed naturally (Dokholyan and Shakhnovich, 2001). The correlation between the naïve synthetic and observed amino acid conservations is as high as 70%, suggesting that proteins are under a dominating evolutionary pressure to preserve their structural stability. Of course, the correlation is not perfect, suggesting that the preservation of folding kinetics and function are under evolutionary pressure as well.

Why would evolution preserve residues important for structural stability in proteins? There could be several reasons. First, the number of residues that constitute either the folding nucleus or the active site of the protein is usually much smaller than those that constitute the protein core. Hence the probability that a random substitution will affect either the folding nucleus or the active site is much smaller than the probability that it will affect the protein core. Second, function itself is not conserved in the course of evolution. Although proteins belonging to a specific fold tend to share a set of functions (Shakhnovich et al., 2003), these functions can be diverse enough that the active sites are fully altered. The folding of homologous proteins can be distinct as well, as for example in the case of two homologues Im7 and Im9 (Friel, Capaldi, and Radford, 2003), suggesting that the folding nucleus may not be under strong evolutionary pressure to be preserved. Thus, among three factors—protein structure, folding kinetics, and function—the protein structure is the most invariant. Hence, residues that maintain structural stability are under evolutionary pressure.

Mutations that impact protein stability also impact protein function. Proteins that are severely destabilized upon mutation are either eliminated due to a loss of function or retained through a compensating mutation that reverses the effect by recovering their thermodynamic stabilities. Recently Bloom et al. (2006) demonstrated that improving thermodynamic stability of marginally stable cytochrome P450 increases the chances of the enzyme to exhibit new or improved function. Mildly destabilizing mutations are perhaps the most dangerous as the proteins may still be functional, but not to the extent necessary to maintain the health of the organism. A number of recent studies support the hypothesis that mildly destabilizing mutations in the human genome are the underlying genetic origin of complex diseases (Sunyaev et al., 2001; Yampolsky et al., 2005; Eyre-Walker, Woofit, and Phelps, 2006; Kryukov, Pennacchio, and Sunyaev, 2007).

Evolutionary pressure to preserve protein structure has important implications on protein design. It is plausible that Nature has found "working" folds and exploits them by grafting necessary functions on to them. If stability is preserved and the scaffold can accommodate the new active site or a binding surface, a new protein with protein function can appear. This idea has been utilized in proof-of-principle studies: Hellinga and coworkers successfully transplanted triose phosphate isomerase activity (Dwyer et al., 2004) from one ribose-binding protein to another homologous enzyme (Allert, Dwyer, and Hellinga, 2007). However, while evolution appear to select for structural preservation of folds, Nature needs new folds to explore new functional space. Consequently, evolution is a balance of forces that conserve and diversify the protein domain universe. The mechanisms of protein domain evolution are discussed next.

MECHANISMS OF PROTEIN DOMAIN EVOLUTION

Clues to protein evolution can be observed by understanding conservation patterns imprinted in the protein sequence-structure-function space (Dokholyan and Shakhnovich, 2005; Dokholyan and Shakhnovich, 2007). Protein sequences and structures (Levitt and Chothia, 1976), as well as functions, are not distributed uniformly with respect to each other when corresponding measures of similarity between sequences, structures, or functions are used. Instead, proteins tend to fall into clusters of families with sequences sharing more than 20% sequence similarity often adopting similar structure and often function. These families of proteins are typically assumed to have originated from a common ancestor. However, it is also possible for proteins with less than 10% sequence similarity to share similar structures and functions. For these instances, convergent evolution is suspected to be involved because the sequence similarity is equal to what would be expected of two randomly selected proteins (Holm and Sander, 1993; Rost, 1997).

Convergent evolution is the result of structural exploration from two distinct origins reaching an equilibrium where the structural solution for a given function is similar to each other. The argument for a convergent evolution scenario can also be presented without the assumption of equilibrium in structural exploration. In which case, a specific force *X* of unknown origin that selects for specific folds to perform a given function needs to be specified, rather than exploring a multiplicity of other folds for function optimization. Moreover, this force *X* should be universal since it helps guide the selection of many distinct families of analogues.

There are several important arguments that are at odds with the convergent evolution theory. First, the assumption of equilibrium in structural exploration is inconsistent with the estimated number of protein folds since the number of theoretically possible folds still exceeds what is observed (see discussion in Protein Structural Universe section). Second, the origin of the mysterious force *X* is unknown and it is not clear what would make two presumably unrelated protein structures adopt the same structure and function. An alternative model suggests that protein designability is the guiding force of evolution and can be used to explain the uneven distribution of the number of members of protein fold families. An extension of this model, and simpler explanation for this distribution, was proposed by Dokholyan, Shakhnovich, and Shakhnovich (2002) using graph theoretical analysis that does not rely on any assumption (described in more detail later). Therefore, these models argue against the missing force *X* in the convergent evolution theory based on the *lex parsimoniae* principle. Third, it is arguable that protein domains presumed to be unrelated may in fact be related, but the connection between them is lost (Pei et al., 2003). The premise for establishing evolutionary relationship is often based in the amount of sequence conservation that may not necessarily hold. Indeed, sequences have been observed to diverge rapidly below the 15% similarity threshold during the course of simulated evolution while maintaining the original protein fold.

In the *divergent* protein evolution theory (Chothia and Gerstein, 1997; Grishin, 1997; Murzin, 1998; Zeldovich et al., 2007; Figure 39.2a) all proteins originate from a few protoproteins. The collective mutations accumulated across time create a repertoire of different proteins while maintaining their structural integrity. The current protein universe is a snapshot of the exploding protein universe. To reconstruct relationships between proteins, a comparison of sequences may not be sufficient. Structures are more conserved than sequence across a long evolutionary time and is useful to establish connections between seemingly unrelated protein domains. For some families, it may be possible to uncover connections between distinct fold families that are related by identifying proteins that can convert from one fold to another with few mutations (Cordes et al., 1999, 2000). Perhaps the most prominent example of such chameleon proteins is the dimeric $\alpha + \beta$ Cro protein from bacteriophage lambda, which evolved from an ancestral all-α monomeric protein (Figure 39.2b; LeFevre and Cordes, 2003). By retroevolving lambda Cro, Le Fevre et al. suggested possible mutations that are responsible for the evolution of the ancestral monomeric protein to the modern dimeric homologue.

The protein domain universe graph (PDUG) is constructed to map the relationships between proteins (Dokholyan, Shakhnovich, and Shakhnovich, 2002) and helps to highlight the mutational strategies used by Nature to evolve different folds. In PDUG, each node represents a protein domain and is connected by a bond that carries a weight corresponding to the degree of their structural similarity. PDUG is organized by a large number of subclusters (Figure 39.3) that serves to reflect the imprint of the evolutionary mechanisms. Examining the connectivities within the PDUG subclusters offers important insights into the appearance of new folds. Although chameleon proteins may be exceptions rather than the rule for the appearance of new folds, it is evident that there are certain proteins that play a role of a gatekeeper for specific folds. Elimination of these proteins during evolution separates connected families of protein folds, thereby giving birth to a new fold family (Dokholyan, 2005). One such example is a hydrolytic enzyme cutinase, which degrades cutin (Dokholyan, 2005). Cutin is a polyester composed of hydroxy and epoxy fatty acids, which serves as a protective shield of aerial plants against pathogen entry (Purdy and Kolattukudy, 1975). Cutin degradation is the first step for plant infection and is exploited by fungi expressing cutinase to invade plants. This unique mechanism of plant protection and infection may trace back to the very origin of plants. Coevolution of plants and pathogens may have spurred an observed spread of domains in the largest cluster of the PDUG (Figure 39.3).

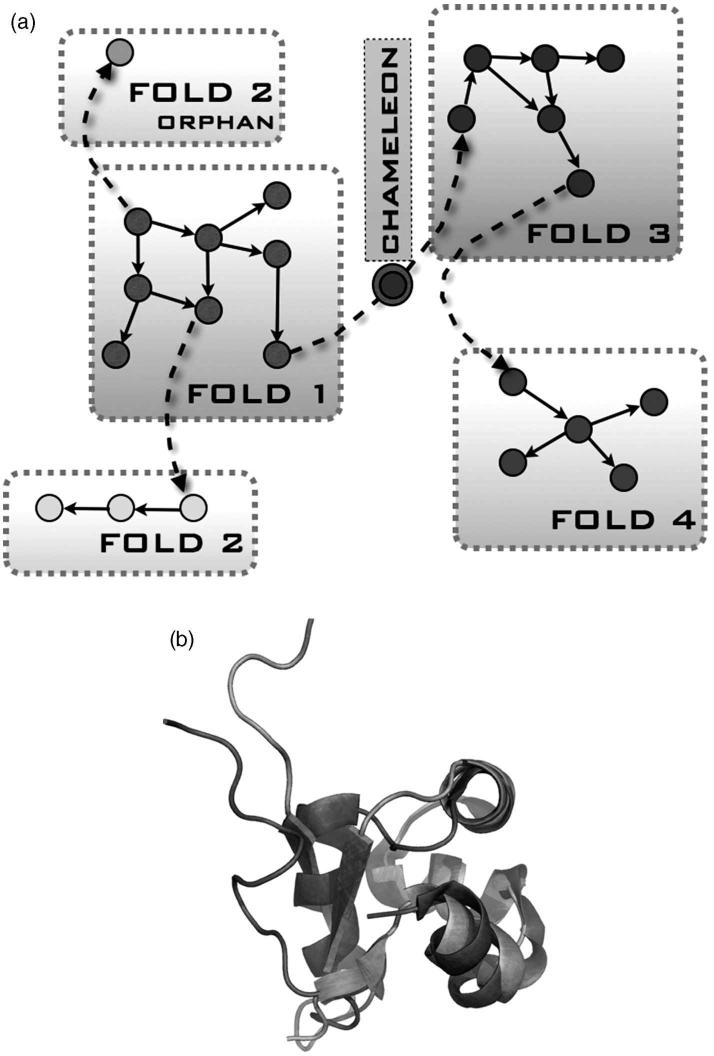

Figure 39.2. Mechanism for protein structure evolution. (a) A schematic representation of the proposed mechanisms for protein evolution (Dokholyan and Shakhnovich, 2005; Dokholyan and Shakhnovich, 2007). Minor structural changes associated with point mutations in proteins diversify the fold families. Major structural changes due to accumulated mutations result in the formation of new fold families. These fold families are populated unequally, some families are represented by just a single protein (orphan family; Dokholyan, Shakhnovich, and Shakhnovich, 2002). In some cases, the proteins may adopt structures that would "bridge" two fold families. These "chameleon" proteins are sensitive to a small number of mutations that make these proteins part of one family or another. (b) An example of the chameleon proteins λ Cro (Protein DataBank access number: 5CRO) and P22 Cro (Protein DataBank access number: 1RZS) (Newlove et al., 2006): structural alignment of these two proteins demonstrates high structural similarity despite dramatic differences in their secondary structures.

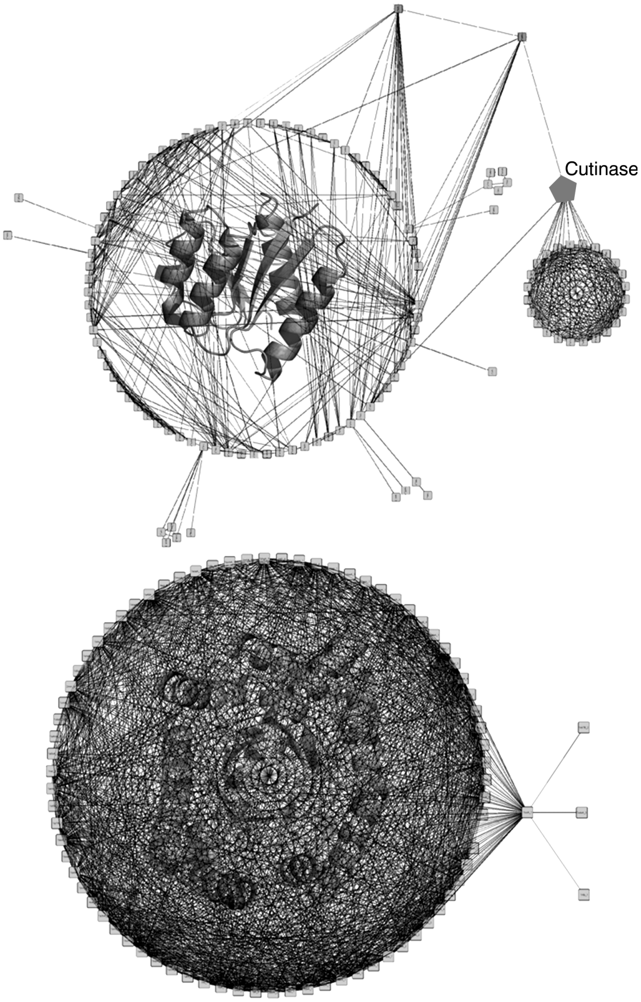

Figure 39.3. The first and the second largest PDUG subclusters. Representative structures for the first (upper graph) and the second largest (lower graph) subclusters of the PDUG are overlayed with the graphs. Figure also appears in Color Figure section.

Examining the PDUG can give some support regarding whether convergent or divergent evolution is being utilized to create the current set of observed protein domains. The population of each of PDUG subcluster described above is observed to follow the power law distribution. However, the significance of this distribution is nullified when compared to a random graph (null hypothesis) that has been constructed to contain the identical number of nodes and edges found in PDUG. The control model showed the same distribution of members in fold families and therefore suggests that this distribution is a property of a random graph and do not have any further deep underlying meaning. On the contrary, the distribution of the connections between protein domains is very particular to PDUG when compare to a random graph. Using a simple model that accounts for divergent protein evolution with only two natural processes—gene duplications and point mutations—the observed distributions of domain connectivity can be reproduced (Dokholyan, Shakhnovich, and Shakhnovich, 2002) thus providing strong support for the divergent evolution theory (Figure 39.2).

In the protein domain evolution model presented by Dokholyan, Shakhnovich, and Shakhnovich (2002), proteins undergo gene duplications and new domains (paralogs and orthologs) accumulate point mutations over time (Figure 39.4). If these mutations do not significantly alter protein folding stability, that is, new proteins are stable enough to perform

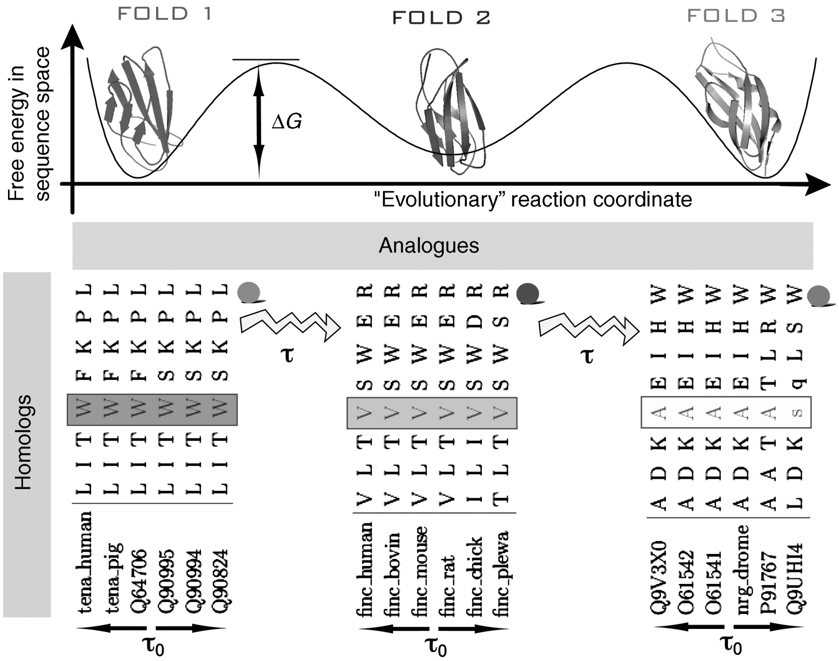

Figure 39.4. Schematic representation of the evolutionary processes that result in conservation patterns of amino acids. For a given family of folds, for example, immunoglobulin (Ig) fold in this schema, there are several minima (3) in the hypothetical free energy landscape in the sequence space as a function of the "evolutionary" reaction coordinate (such as time) (Dokholyan and Shakhnovich, 2005). Each of these minima are formed by mutations in protein sequences at typical time scales (τ_0), that do not alter the protein's thermodynamically and kinetically important sites, forming families of homologous proteins. Transitions from one minimum to another occur at longer 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 alter several amino acids at the important sites of the proteins so that the protein properties are not compromised, but the fold is changed. At time scale τ the family of analogues is formed. Three minima in this schema represent three families of homologues (1TEN, 1FNF, and 1CFB) each comprised of six homologous proteins. Only eight positions in the aligned proteins are shown: 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 homologues, but vary between these families. This position corresponds to position 21 in Ig fold alignment (to 1TEN) and is highly conserved (Dokholyan and Shakhnovich, 2001). Modified from Dokholyan NV, Shakhnovich EI. Scale-Free Evolution: From Proteins to Organisms. In: Koonin EV, Wolf YI, Karev GP, eds. Power Laws, Scale-Free Networks and Genome Biology. Austin: Landes Bioscience, 2005.

their functions, then they constitute the same fold. When a critical number of mutations accumulate, the core is potentially at risk to be disrupted or rearranged forcing significant structural changes, thereby giving birth to a new fold family. Within a fold family, sequences also diverge unequally: sequences defining homologues share 20% sequence similarity, while a fold family unifies numerous families of homologues with distinct numbers of homologue family members. Since the domain core preservation is an essential constraint in protein evolution, sequences with $\sim$20% similarity are "guaranteed" to have structural similarity if the fraction of the shared residues constitute the critical component of the core. However, the conservation of the core may not be the only contributing factor effecting the distribution of fold families. Another contributing factor may be the physicochemical properties of the amino acids—the protein alphabet that defines protein sequence and structure. Next, we discuss the properties of the protein alphabet.

PROTEIN ALPHABET

The protein structure–sequence relationship is determined by the amino acid composition and amino acid order (see Chapter 2 for full details). The 20 amino acids have overlapping physicochemical properties resulting in a redundancy of this set of building blocks. Hypothetically the redundancy implies that proteins can be encoded with a smaller subset of the current amino acids. How would be the minimum set of amino acid residues be? To make a crude approximation of the ideal alphabet size, M_I, we compare the number of available sequences with this alphabet M_I^N for a protein of size N with the number of available protein conformations $(z-1)^{N-1}$ (estimated earlier in Protein Structural Universe section), where z is the coordination number. Equating these two numbers leads to $M_I \approx z-1 = 5$. This result is very surprising because it is fourfold smaller than the natural alphabet.

Although this estimation is extremely naïve, a number of studies have supported similar estimates. First, using information theory and bioinformatics, Strait and Dewey showed that a typical protein sequence features Shannon entropy values of approximately 2.5 bits. This number means that one needs $2^{2.5} \approx 5.7$ letters to encode protein structure. Of course, these letters are not amino acids, but rather a juxtaposition of various amino acids. Second, a study using lattice protein models and pairwise amino acid interaction potentials, demonstrated that one could design stable proteins using 20 but not 2 amino acid alphabet (Shakhnovich, 1994). The folding properties (foldability; Klimov and Thirumalai, 1996) of lattice protein models are also not significantly altered using a five-letter alphabet (Wang and Wang, 2000) but are altered using a two-letter alphabet. Third, by performing simulations of protein evolution, the amino acid conservation in protein families can be recapitulated using a six-letter amino acid alphabet (Dokholyan and Shakhnovich, 2001). Fourth, the SH3 domain can be experimentally redesigned to a sequence that consisted of only five amino acid types (Riddle et al., 1997); the expected probability of five-letter alphabet usage for a 56-residue protein is $p \sim 10^{-29}$ (Dokholyan, 2004). Last but not least, the usage of amino acid types in natural proteins is less than expected if amino acids were chosen randomly with naturally occurring frequencies (Dokholyan, 2004). Taken together, it seems that Nature has utilized a larger alphabet than it could have for the protein universe.

Why does Nature exploit a larger alphabet than necessary for creating diversity in protein structure and function? Naively, it seems that a larger alphabet size would be inefficient because of the additional metabolic machinery that is needed by each additional amino acid. However, there are other important properties of natural proteins that are only possible with the current, and larger, set of amino acids. For examples, it is important to retain robustness in the protein structure. By creating diversity within the repertoire of amino acids, to the impact of potentially damaging mutations is reduced. If Nature utilized the most efficient "ideal" set of amino acids, then proteins are not safe guarded by mutations that would be damaging to protein folds.

Another benefit for having a larger alphabet size is that the evolutionary barriers (free energy barriers with respect to some "evolutionary" reaction coordinate—Figure 39.4) that separate fold families are lowered. This property allows Nature to optimally explore fold families in the course of protein evolution. Extremely large barriers, which would be a consequence if a reduced set of amino acids were used, would "freeze" the protein universe in one state and evolution would not be able to proceed. Hence, it is expected that as evolution proceeds, the alphabet size increases (Jordan et al., 2005), although not all amino acids may be utilized at any given point during evolution.

It is plausible that Nature chooses a golden mean between robustness (tolerance to mutations) and metabolic efficiency in selecting the protein alphabet size (Figure 39.5). The redundancy within the amino acid set also has important implications on protein design: it offers a margin for error associated with the inaccuracies inherent in the current force fields used by computer-based protein design. It also allows for the diversification of sequences without altering protein fold. Next, we describe the effects of alphabet size and other physical factors on the population of a given fold.

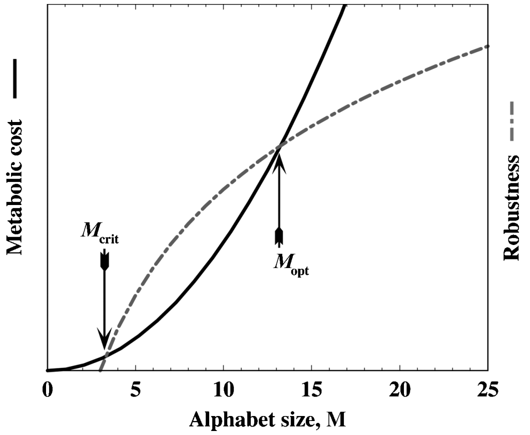

Figure 39.5. Hypothetical diagram of the benefits and drawbacks of protein alphabet size. The robustness associated with protein folds increases as the alphabet size M increases (dot–dashed line). At the same time, metabolic cost for producing M number of amino acids is also an increasing function of M (solid line). We posit that the latter function is a more rapidly growing one than that for the former. Two intercepts of these lines correspond to the critical (M_{crit}) and the optimal (M_{opt}) alphabet sizes. The critical alphabet size signifies the point at which it becomes beneficial to produce proteins tolerant to mutations. The optimal alphabet size signifies the point at which the metabolic cost and robustness are at an optimal state, above which the metabolic cost outweighs the benefits of robustness of folds to mutations.

LESSONS FOR ENGINEERING STABLE PROTEINS

Protein stability is perhaps the most important determinant of protein evolution. In which case, what lessons for rational protein design could this observation teach us? The fold does provide constraints for the number of sequences that can adopt a given structure, but how would the number of available amino acids for use affect this number? What other physical factors contribute to protein designability? Next we attempt to shed light on these questions from a simplified physical consideration.

For a protein to be stable in its native state, the potential energy $H(\Gamma)$ of its native state Γ must be smaller than that of any other alternative (*decoy*) state Γ_D, for all decoys D:

$$H(\Gamma) < H(\Gamma_D), \quad \forall D. \tag{39.1}$$

Using the random energy model (Derrida, 1980), the thermodynamic stability of proteins and the potential energy difference (*energy gap*) ΔH between the native state and the lowest potential energy decoy Γ_C can be related (Bryngelson and Wolynes, 1989; Gutin and Shakhnovich, 1993), which is then normalized by the root-mean-square deviation of the distribution of potential energies $\sigma(H)$ of decoy conformations (Dokholyan, 2004):

$$Z \equiv \frac{\Delta H}{\sigma(H)} = -1. \tag{39.2}$$

Equation 39.2 provides a measure that helps to find a sequence S that minimizes the Z-score during the process of rational design.

This Z-score has been minimized to find an *ideal* force field for a M-letter alphabet (Goldstein, Luthey-Schulten, and Wolneys, 1992), and the actual value of the minimal Z-score value has recently been derived, $Z_{\min}^{(M)}$ (Dokholyan, 2004):

$$Z_{\min}^{(M)} = -\left[\sum_{\sigma_a,\sigma_b}^{M} K^2(\sigma_a,\sigma_b)\right]^{1/2} + \sqrt{2N\log\gamma}, \tag{39.3}$$

where N is a protein length, γ is an effective number of degrees of freedom per atom—a parameter that reflects protein flexibility (Dokholyan, 2004). The kernel function $K(\sigma_a,\sigma_b)$ is

$$K(\sigma_a,\sigma_b) = \frac{\sum\limits_{i\neq j=1}^{M} s_i(\sigma_a)s_j(\sigma_b)(\Delta_{ij}-f_{ij})}{\left[\sum\limits_{i\neq j=1}^{M} s_i(\sigma_a)s_j(\sigma_b)f_{ij}(1-f_{ij})\right]^{1/2}}, \tag{39.4}$$

where f_{ij} are the frequencies of contacts between atoms i and j in decoy conformations, Δ_{ij} is the contact matrix element, which is equal to 1 or 0 depending on the presence of the contact between atoms i and j in the native state. $s_i(\sigma_a)$ is the amino acid at the position i of type σ_a.

Furthermore, Eq. 39.3 takes on a simpler form when using the Gō, model (Gō, 1983), in which atomic interactions are determined by the native protein structure (Dokholyan, 2004)

$$Z_{\min}^{(\mathrm{Go})} = -\left[\sum_{i\neq j=1}^{N} \frac{(\Delta_{ij}-f_{ij})^2}{f_{ij}(1-f_{ij})}\right]^{1/2} + \sqrt{2N\log\gamma}. \tag{39.5}$$

Comparing Eqs 39.3 and 39.5, it is straightforward to prove that for all M

$$|Z_{\min}^{(M)}| \leq |Z_{\min}^{(M+1)}| \leq \cdots \leq |Z_{\min}^{(\mathrm{Go})}|, \quad \forall M. \tag{39.6}$$

Equation 39.6 has profound implications on protein design: it suggests that with an increase of the size of available protein alphabet, more stable proteins can be designed. The limiting case is the Gō model, in which the alphabet size is equal to the number of possible contacts in a protein.

Another important implication of Eqs 39.3 and 39.4 is that they can answer the question regarding the smallest alphabet size M_c required to encode a protein. Interestingly, this value

$$M_c = 2\sqrt{\ln \gamma / \varepsilon} \tag{39.7}$$

Q1 is independent of the protein length N (Dokholyan, 2004; ε is a factor independent of N and M). Thus, with an alphabet of size $M \geq M_c$ one can design a protein of arbitrary length.

The values of the $Z_{\min}$ suggest three principal factors that govern protein designability: (i) the alphabet size M, (ii) protein flexibility γ, and (iii) properties of the unfolded or misfolded states f_{ij}. The alphabet size determines the maximum possible stability of designed proteins and, therefore, how many sequences a given fold can adopt. The protein flexibility and unfolded states are important factors to account for during rational protein design as they differentiate the designed proteins from the reference states (decoys). Hence, proper sampling of protein conformations during computational protein design is essential.

An important consequence of Eqs 39.3–39.5, is that they suggest the concept of designability (Li et al., 1996), that is, that some folds may have more sequences that make them stable than others. Indeed, since the properties of unfolded or *molten globule* (unfolded with some compact units; Ptitsyn, 1995; Finkelstein and Ptitsyn, 2002) states are determined by the native state (Pappu, Srinivasan, and Rose, 2000; Fitzkee and Rose, 2004; Kohn et al., 2004; Ding, Jha, and Dokholyan, 2005), the frequencies of contacts f_{ij} impact $Z_{\min}$. Hence, the lower the $Z_{\min}$ for a given fold, the more variation in stability due to mutations a given fold can tolerate, and the larger fold family we expect. England and Shakhnovich (2003) further uncovered role of the contact matrices on protein designability by deriving an explicit dependence of the number of designable sequence on the largest eigenvalue of the contact map.

The thermodynamic stability requirement has no implication for how stable natural proteins are (Taverna and Goldstein, 2002; Xia and Levitt, 2002). In fact, it has been established that natural proteins can be "redesigned" to be more stable than wild type (Chen et al., 2000; Dantas et al., 2003). Extremely stable proteins are not necessarily beneficial to Nature. As we show below, it is hard to evolve proteins to be extremely stable (whose Z-values are proximal to $Z_{\min}$). Additionally, extremely stable proteins are maximally sensitive to random mutations. Hence, a majority of natural proteins are "marginally" stable (Taverna and Goldstein, 2002). As long as a protein optimally performs its function, there is no evolutionary pressure to make it more stable.

Theoretically, one can always improve the stability of a protein up to $Z_{\min}$ value for a corresponding fold. However, the thermodynamic stability of the original fold and the fold limiting value $Z_{\min}$ significantly affect the ability to find such sequences. To demonstrate this, let us note that the distribution of the Z-values of amino acid sequences that stabilize a given fold is normal. Indeed, it is an assumption of the Random Energy Model that the potential energy is a sum of “random” pairwise energy terms between atoms and, therefore, the values of potential energies of protein conformations follow a Gaussian distribution, which after normalization (Eq. 39.2) becomes normal. Hence, the number of sequences Ξ that are more stable than a given one with the Z-value equals to Z_0:

$$\Xi(Z<Z_0) \propto \int_{Z_0}^{Z_{\min}} \exp\left(-\tfrac{Z^2}{2}\right)\mathrm{d}Z \propto \operatorname{erf}\left(-\tfrac{Z_{\min}}{\sqrt{2}}\right)-\operatorname{erf}\left(-\tfrac{Z_0}{\sqrt{2}}\right). \tag{39.8}$$

If the thermodynamic stability of a protein that we wish to redesign is proximal to a fold’s minimal value $Z_{\min}$, the number of sequences becomes extremely hard to find as their number significantly decreases

$$\Xi(Z<Z_0) \propto |Z_{\min}-Z_0|e^{-\frac{1}{2}Z_{\min}^2}\left\{1-\frac{1}{2}Z_{\min}|Z_{\min}-Z_0|\right\}+O(|Z_{\min}-Z_0|^3), \quad Z_0 \rightarrow Z_{\min}. \tag{39.9}$$

Thus, the number of sequences more stable than a given one (with the Z-value equals to Z_0) $\Xi(Z<Z_0)$ decreases dramatically as Z_0 approaches its limiting value. This observation makes computational or experimental searches for sequences that make a protein even more stable an extremely difficult task.

An important implication of Eq. 39.9 is that most random mutations are destabilizing. This consequence is especially important for many human diseases, suggesting that single nucleotide polymorphisms (SNPs) associated with diseases may decrease the stability of proteins produced from the corresponding genes. From this discussion, we can appreciate the limitation state structural stability has on shaping the allowable sequence space for a given fold.

PROTEIN ENGINEERING: US VERSUS NATURE

Learning how Nature operates has obvious advantages, but it is always important to remember the underlying motives for evolving or designing a specific protein. We often pursue goals that are distinct from Nature’s goals when designing proteins. For Nature it is important to design a functional protein. Proteins are not under evolutionary pressure to be extremely stable, more active than that is required for existing cell environments, or extremely fast folders. Most existing computational algorithms for protein design, on the other hand, search specifically for the sequence that minimizes the potential energy of a given scaffold. Although the ultimate goal of the protein design field is to rationally manipulate protein function and cell life, Nature evolves proteins as a part of the complex cellular system, thereby taking into consideration a whole spectrum of factors and functions. In this perspective, the future of the protein evolution field is projected into the realm of *systems evolution*.

There are many other constraints that Nature honors in the course of evolution. For example, allostery is an important factor to account for protein function. It is currently extremely difficult to account for allosteric effects in proteins during rational design. Some success has been achieved with designing molecular switches (Ambroggio and Kuhlman, 2006). However, nobody has so far succeeded with the design of rational changes in protein structure upon ligand binding. This problem is one of the most complex in protein design since it requires design of networks of amino acid interactions throughout protein structure. While the communication of amino acids within protein structures has been established using graph theoretical analysis in a number of studies (Lockless and Ranganathan, 1999; Dokholyan et al., 2002; Vendruscolo et al., 2002), it is not clear whether it is purposely evolved or just a generic property of protein structure. For example, while there is a clear relation between topologically and kinetically important residues in proteins (Dokholyan et al., 2002; Vendruscolo et al., 2002), this relation may be a consequence of a similar one between protein structure and its folding kinetics. This observation is supported by several studies that identified significantly contributing residues to protein folding kinetics (Clementi, Nymeyer, and Onuchic, 2000; Ding et al., 2002a) using a structure-based model of protein energetics the Gō model (Gō, 1983). Thus, the success in designing allostery in proteins strongly depends on the significant progress in our understanding of amino acid interaction networks in protein structure and their evolution.

The function of some proteins is structurally achieved by disordered regions that lack secondary or any well-defined unique structure. In fact, there are a number of fully disordered proteins (Le Gall et al., 2007) and their potential impact on protein function, biology, and other functional roles are discussed in Chapter 38. Computational design of such proteins is a formidable challenge. Perhaps what makes rational design of disordered proteins even more complex is that despite the lack of well-defined structure of such proteins, they are distinct from random polymers. These proteins coexist in multiple states, which actually have unique structural properties at the ensemble level, that is, they share specific *structural signatures*. One remarkable example is α-synuclein, which lacks a unique native state, but is more compact than would be expected for a random polypeptide (Dedmon et al., 2005). Although the function of α-synuclein is unknown, disordered regions are often employed by other proteins, such as antibodies and cell signaling proteins, which have intrinsic plasticity to recognize and bind other molecules. In such proteins, disorder is only visible in free structure. Local or global structural reorganization occurs upon the binding of partner molecules. Such proteins are a significant complication for computational protein design because the stability of the complex with a partner protein, rather than the thermodynamic stability of free proteins, is the priority for Nature. Since many such proteins have a large number of partners, a further challenge for protein design is to engineer protein interactions with several of the partners, which may force a designed protein to adopt a number of alternative bound structures.

Another important focus of natural design is the active site of enzymes. Enzyme activity is extremely sensitive to the geometry and dynamic properties of the active site and its conformation in the transition states. This sensitivity is controlled at the sub-Angstrom level, making Nature the most elegant and scrupulous designer. The level of accuracy required for enzyme design makes it one of the most challenging tasks for protein designers. Although there has been significant progress in computational enzyme design, pioneered by Hellinga (Dwyer et al., 2004), we are still not in the realm of natural enzyme design.

Besides the differences in design philosophies between Nature and us, there are a number of other issues that we need to overcome to become efficient in designing functional proteins. In the next section, we discuss some of these issues.

CHALLENGES IN PROTEIN DESIGN

There are a number of experimental and computational approaches to protein design. While experimental approaches that attempt to mimic evolution *in vitro* offer a powerful search methodology for desired proteins, computational design has particular appeal since it provides rational explanation, and correspondingly a higher perspective for control, for the effect of changes on protein structure and function. Next, we discuss the main difficulties facing computational protein design.

Energetics

One of the most important components in protein design, as well as protein folding and structure prediction, is the force field. The force field is a set of parameters that contributes to the calculation of the potential energy of the protein. There are a number of strategies to derive a force field for a given molecular model: physical, empirical, knowledge based, constraint based, and their combinations. Physical force fields (e.g., CHARMM and Amber) require physical entities, that is, atoms, and therefore are most appropriate for detailed all-atom models. Empirical force fields rely on specific observations pertinent to a given molecule, and therefore are molecule type-specific. Among empirical force fields is the structure-based Gō model, which biases proteins to their native states. Knowledge-based force fields are constructed on current knowledge of molecular structure, physical/chemical properties, and evolution, and therefore are limited to molecules that resemble known molecules. Constraint-based force fields rely on experimental data, molecular structure, and dynamics.

The more details of interactions accounted by the force field (i.e., the more accurate it is), the slower the calculation of the potential energy of a given protein state. Slow computation of the potential energy, in turn, results in diminished sampling or search abilities, whether it is a search for a protein sequence during protein design, or conformational sampling during protein folding or structure prediction. Hence, there is a fine balance between the accuracy of the force field and the ability to search or sample. Finding such a balance between the potential energy calculations and the ability to search for a solution is one of the principal challenges in protein design.

Conformational sampling is the other side of the coin from estimating the free energy of a given protein state and its corresponding thermodynamic stability. The free energy estimation of a given state requires understanding the properties of the ensemble of all conformations in the free energy basin of a state of interest. During protein design, one needs to find a large number of conformations that a given sequence can adopt within such a basin. An amino acid substitution may result in atomic clashes between the newly substituted amino acid and other residues (Ding and Dokholyan, 2006; Yin et al., 2007). Such clashes can be resolved by "massaging" the protein structure via conformational sampling of the protein backbone. Surprisingly the perturbation of the backbone required for resolving a clash is often very minimal, less than 1 Å (Yin et al., 2007), although it is important to note that not all clashes can be resolved.

Estimation of protein thermodynamic stabilities is even more challenging as it requires the knowledge of the free energies of the reference (unfolded) states. The free energy of the unfolded states is predominantly determined by the accurate modeling of the ensembles of unfolded state structures. Due to difficulties in determining the ensemble of unfolded states, many design algorithms rely on approximations of the reference states or derive these parameters by training using various strategies. Rapid sampling techniques may potentially improve thermodynamic stability evaluations through explicit sampling (Ding, Jha, and Dokholyan, 2005).

Alternative Binding Sites

An important potential complication during design of protein–protein interactions is the presence of alternative binding sites. Even though interactions at the alternative binding sites are most likely to occur at lower affinity, these sites compete for the ligand and, therefore, decrease the effective binding to the target region. These complications could be alleviated by (1) improving sampling during docking of a target protein on a scaffold to identify the potential alternative binding sites, and (2) performing negative design to decrease binding affinities at the alternative binding sites.

Aggregation

Protein aggregation due to association of identical proteins is a common phenomenon and can result from multiple factors. Protein aggregation occurs when protein thermodynamic stability and concentration are such that identical species can form intermolecular hydrogen bonds between their backbones. Thermodynamic stability plays an essential role in protein aggregation, as proteins often need to unfold to expose their backbone to other proteins for hydrogen bond formation (Ding et al., 2002b).

There are two dominant reasons for protein aggregation in the course of protein design. First, the designed protein may not be stable enough. In such cases, further improvement of the protein's thermodynamic stability may alleviate the problem. Second, if the protein is rich in β-sheets, the amino acids constituting these β-strands are already in geometrical configurations consistent with those of protein aggregates. This is a serious complication during protein design as there is no clear understanding, despite a number of interesting considerations showing how Nature protects β-rich proteins from aggregation (Patki, Hausrath, and Cordes, 2006; Richardson and Richardson, 2002). It has been suggested that capping β-sheets with charged residues and creating bulges on the surface of β-sheets may protect them from associating with identical proteins. However, no successful design has been carried out to build a novel β-rich protein. Further understanding of Nature's defense against protein aggregation may offer new strategies for designing β-rich proteins.

CONCLUSIONS

Protein engineering enters a new era as rational manipulation of protein structure becomes more successful due to acquired knowledge about protein structure, energetics, and evolution. With the increase of the number of determined protein structures, we now better understand the organization of the protein structure universe. We are able to link structurally diverse protein families and understand evolutionary determinants of families of protein structures. By developing an understanding of the basic mechanisms of protein evolution, we learn to manipulate protein structure by mimicking the natural processes. From theoretical considerations, we have learned important lessons for engineering stable proteins, such as the protein alphabet size, protein flexibility, and the properties of the unfolded states.

Despite tremendous progress in protein engineering, we still face challenges. These challenges include design in the framework of cellular environments, for example when protein design must include potentially negative design against interactions with other proteins. Modeling unfolding proteins is essential to protein design and improvements in our models of unfolded states may allow for rational manipulations of naturally disordered proteins or proteins that feature functional unstructured regions. Further challenges are associated with sub-angstrom design that is necessary for building novel enzymes. Understanding protein aggregation will help us defy such natural phenomena as protein aggregation and nonspecific associations with other proteins.

Improvements may come from further development of the force fields and our ability to more exhaustively sample protein conformations. An understanding of the protein alphabet is important to computational protein design (Shakhnovich, 1998) because the underlying protein alphabet determines the parameter set for the energy function that is used for protein design. Therefore, further studies are needed to uncover the fundamental protein alphabet responsible for encoding protein three-dimensional structure. These studies may also shed light on the redundancy of the genetic code and identify the principal rules of molecular evolution.

Protein engineering is already influencing modern biology and offering help in understanding cellular life. Many of these considerations are applicable to other polymers, such as RNA and DNA molecules. Although these molecules have specific challenges associated with their design, most of the discussed aspects of protein design are also applicable to these molecules. Future engineering efforts offer remarkable promise for impacting medicine: from the engineering of antibodies and enzymes to the development of biomarkers for early diagnosis of disease. Protein engineering will also play an important role in elucidating the molecular etiologies of human diseases and the development of personalized medicine.

REFERENCES

Akiyama M, Sakai K, Ogawa M, McMillan JR, Sawamura D, Shimizu H (2007): Novel duplication mutation in the patatin domain of adipose triglyceride lipase (PNPLA2) in neutral lipid storage disease with severe myopathy. *Muscle Nerve* 36 (6):856–859.

Allert M, Dwyer MA, Hellinga HW (2007): Local encoding of computationally designed enzyme activity. *J Mol Biol* 366:945–953.

Ambroggio XI, Kuhlman B (2006): Design of protein conformational switches. *Curr Opin Struct Biol* 16:525–530.

Anantharaman V, Aravind L, Koonin EV (2003): Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins. *Curr Opin Chem Biol* 7:12–20.

Benson MD, Kincaid JC (2007): The molecular biology and clinical features of amyloid neuropathy. *Muscle Nerve* 36 (4):411–423.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000): The protein data bank. *Nucleic Acids Res* 28:235.

Bloom JD, Labthavikul ST, Otey CR, Arnold FH (2006): Protein stability promotes evolvability. *Proc Natl Acad Sci USA* 103:5869–5874.

Boucher RC (2004): New concepts of the pathogenesis of cystic fibrosis lung disease. *Eur Respir J* 23:146–158.

Bryngelson JD, Wolynes PG (1989): Intermediates and barrier crossing in a random energy-model (with applications to protein folding). *J Phys Chem* 93:6902.

Chen JM, Lu ZQ, Sakon J, Stites WE (2000): Increasing the thermostability of staphylococcal nuclease: Implications for the origin of protein thermostability. *J Mol Biol* 303:125.

Chen Y, Ding F, Dokholyan NV (2007): Fidelity of the protein structure reconstruction from inter-residue proximity constraints. *J Phys Chem B* 111:7432–7438.

Chothia C (1992): Proteins—1000 Families for the molecular biologist. *Nature* 357:543.

Chothia C, Gerstein M (1997): Protein evolution. How far can sequences diverge? *Nature* 385 (579):581.

Clementi C, Nymeyer H, Onuchic JN (2000): Topological and energetic factors: What determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J Mol Biol* 298:937.

Cleveland DW (1999): From Charcot to SOD1: Mechanisms of selective motor neuron death in ALS. *Neuron* 24:515.

Cordes MH, Burton RE, Walsh NP, McKnight CJ, Sauer RT (2000): An evolutionary bridge to a new protein fold. *Nat Struct Biol* 7:1129–1132.

Cordes MH, Walsh NP, McKnight CJ, Sauer RT (1999): Evolution of a protein fold in vitro. *Science* 284:325–328.

Dantas G, Kuhlman B, Callender D, Wong M, Baker D (2003): A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. *J Mol Biol* 332:449–460.

Dedmon MM, Lindorff-Larsen K, Christodoulou J, Vendruscolo M, Dobson CM (2005): Mapping long-range interactions in alpha-synuclein using spin-label NMR and ensemble molecular dynamics simulations. *J Am Chem Soc* 127:476–477.

Derrida B (1980): The random energy-model. *Phys Rep* 67:29.

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

Ding F, Dokholyan NV (2006): Emergence of protein fold families through rational design. *PLoS Comput Biol* 2:e85.

Ding F, Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI (2002a): Direct molecular dynamics observation of protein folding transition state ensemble. *Biophys J* 83:3525–3532.

Ding F, Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI (2002b): Molecular dynamics simulation of the SH3 domain aggregation suggests a generic amyloidogenesis mechanism. *J Mol Biol* 324:851–857.

Ding F, Jha RK, Dokholyan NV (2005): Scaling behavior and structure of denatured proteins. *Structure* 13:1047–1054.

Dokholyan NV (2004): What is the protein design alphabet? *Proteins* 54:622–628.

Dokholyan NV (2005): The architecture of the protein domain universe. *Gene* 347:199–206.

Dokholyan NV, Li L, Ding F, Shakhnovich EI (2002): Topological determinants of protein folding. *Proc Natl Acad Sci USA* 99:8637–8641.

Dokholyan NV, Shakhnovich B, Shakhnovich EI (2002): Expanding protein universe and its origin from the biological Big Bang. *Proc Natl Acad Sci USA* 99:14132–14136.

Dokholyan NV, Shakhnovich EI (2001): Understanding hierarchical protein evolution from first principles. *J Mol Biol* 312:289–307.

Dokholyan NV, Shakhnovich EI (2005): Scale-free evolution: from proteins to organisms. In: Koonin EV, Wolf YI, Karev GP, editors. *Power Laws, Scale-free Networks and Genome Biology*. Austin, TX: Landes Bioscience, Eurekah.com and Springer, pp. 86–105.

Dokholyan NV, Shakhnovich EI (2007): Towards unifying protein evolution theory. In: Bastolla U, Porto M, Roman HE, Vendruscolo M, editors. *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations*. Berlin: Springer, pp. 113–126.

Dwyer MA, Looger LL, Hellinga HW (2004): Computational design of a biologically active enzyme. *Science* 304:1967–1971.

England JL, Shakhnovich EI (2003): Structural determinant of protein designability. *Phys Rev Lett* 90:218101.

Eyre-Walker A, Woolfit M, Phelps T (2006): The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* 173:891–900.

Finkelstein AV, Ptitsyn O (2002): *Protein Physics: A Course of Lectures (Soft Condensed Matter, Complex Fluids and Biomaterials)*. Boston: Academic Press.

Fitzkee NC, Rose GD (2004): Reassessing random-coil statistics in unfolded proteins. *Proc Natl Acad Sci USA* 101:12497–12502.

Friel CT, Capaldi AP, Radford SE (2003): Structural analysis of the rate-limiting transition states in the folding of Im7 and Im9: similarities and differences in the folding of homologous proteins. *J Mol Biol* 326:293–305.

Gō N (1983): Theoretical studies of protein folding. *Annu Rev Biophys Bioeng* 12:183.

Goldstein RA, Luthey-Schulten ZA, Wolynes PG (1992): Protein tertiary structure recognition using optimized Hamiltonians with local interactions. *Proc Natl Acad Sci USA* 89:9029–9033.

Grishin NV (1997): Estimation of evolutionary distances from protein spatial structures. *J Mol Evol* 45:359–369.

Grosberg AY, Khokhlov AR (1997): *Giant Molecules*. Boston: Academic Press.

Gutin AM, Shakhnovich EI (1993): Ground-state of random copolymers and the discrete random energy-model. *J Chem Phys* 98:8174.

He X, van Waardenburg RC, Babaoglu K, Price AC, Nitiss KC, Nitiss JL, Bjornsti MA, White SW (2007): Mutation of a conserved active site residue converts tyrosyl-DNA phosphodiesterase I into a DNA topoisomerase I-dependent Poison. *J Mol Biol* 372 (4):1070–1081.

Holm L, Sander C (1993): Protein structure comparison by alignment of distance matrices. *J Mol Biol* 233:123–138.

Holm L, Sander C (1997): Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res* 25:231.

Jordan IK, Kondrashov FA, Adzhubei IA, Wolf YI, Koonin EV, Kondrashov AS, Sunyaev S (2005): A universal trend of amino acid gain and loss in protein evolution. *Nature* 433:633–638.

Khare SD, Caplow M, Dokholyan NV (2006): FALS mutations in Cu, Zn superoxide dismutase destabilize the dimer and increase dimer dissociation propensity: a large-scale thermodynamic analysis. *Amyloid* 13:226–235.

Klimov DK, Thirumalai D (1996): Criterion that determines the foldability of proteins. *Phys Rev Lett* 76:4070–4073.

Kohn JE, Millett IS, Jacob J, Zagrovic B, Dillon TM, Cingel N, Dothager RS, Seifert S, Thiyagarajan P, Sosnick TR, et al. (2004): Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc Natl Acad Sci USA* 101:12491–12496.

Kryukov GV, Pennacchio LA, Sunyaev SR (2007): Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet* 80:727–739.

Le Gall T, Romero PR, Cortese MS, Uversky VN, Dunker AK (2007): Intrinsic disorder in the protein data bank. *J Biomol Struct Dyn* 24:325–342.

LeFevre KR, Cordes MH (2003): Retroevolution of lambda Cro toward a stable monomer. *Proc Natl Acad Sci USA* 100:2345–2350.

Levitt M, Chothia C (1976): Structural patterns in globular proteins. *Nature* 261:552–558.

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

Lockless SW, Ranganathan R (1999): Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286:295.

Martin AC, Orengo CA, Hutchinson EG, Jones S, Karmirantzou M, Laskowski RA, Mitchell JB, Taroni C, Thornton JM (1998): Protein folds and functions. *Structure* 6:875.

Murzin AG (1998): How far divergent evolution goes in proteins. *Curr Opin Struct Biol* 8:380–387.

Newlove T, Atkinson KR, Van Dorn LO, Cordes MH (2006): A trade between similar but nonequivalent intrasubunit and intersubunit contacts in Cro dimer evolution. *Biochemistry* 45:6379–6391.

Orengo CA, Jones DT, Thornton JM (1994): Protein superfamilies and domain superfolds. *Nature* 372:631.

Pappu RV, Srinivasan R, Rose GD (2000): The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding. *Proc Natl Acad Sci USA* 97:12565–12570.

Patki AU, Hausrath AC, Cordes MH (2006): High polar content of long buried blocks of sequence in protein domains suggests selection against amyloidogenic non-polar sequences. *J Mol Biol* 362:800–809.

Pei J, Dokholyan NV, Shakhnovich EI, Grishin NV (2003): Using protein design for homology detection and active site searches. *Proc Natl Acad Sci USA* 100:11361–11366.

Ptitsyn OB (1995): Molten globule and protein folding. *Adv Protein Chem* 47:83.

Purdy RE, Kolattukudy PE (1975): Hydrolysis of plant cutin by plant pathogens. Purification, amino acids composition, and molecular weight of two isoenzymes of cutinase and a nonspecific esterase from *Fusarium solani* f. pisi. *Biochemistry* 14 (13):2824.

Qian J, Luscombe NM, Gerstein M (2001): Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J Mol Biol* 313:673.

Richardson JS, Richardson DC (2002): Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc Natl Acad Sci USA* 99:2754.

Riddle DS, Santiago JV, Bray-Hall ST, Doshi N, Grantcharova VP, Yi Q, Baker D (1997): Functional rapidly folding proteins from simplified amino acid sequences. *Nat Struct Biol* 4:805–809.

Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL, et al. (1989): Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* 245:1066–1073.

Rommens JM, Iannuzzi MC, Kerem B, Drumm ML, Melmer G, Dean M, Rozmahel R, Cole JL, Kennedy D, Hidaka N, et al. (1989): Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* 245:1059–1065.

Rost B (1997): Protein structures sustain evolutionary drift. *Fold Des* 2:S19–S24.

Shakhnovich BE, Dokholyan NV, DeLisi C, Shakhnovich EI (2003): Functional fingerprints of folds: evidence for correlated structure-function evolution. *J Mol Biol* 326:1–9.

Shakhnovich EI (1994): Proteins with selected sequences fold into unique native conformation. *Phys Rev Lett* 72:3907–3910.

Shakhnovich EI (1998): Protein design: a perspective from simple tractable models. *Fold Des* 3: R45.

Sunyaev S, Ramensky V, Koch I, Lathe W 3rd, Kondrashov AS, Bork P (2001): Prediction of deleterious human alleles. *Hum Mol Genet* 10:591–597.

Taverna DM, Goldstein RA (2002): Why are proteins marginally stable? *Proteins* 46:105.

Vendruscolo M, Dokholyan NV, Paci E, Karplus M (2002): Small-world view of the amino acids that play a key role in protein folding. *Phys Rev E Stat Nonlin Soft Matter Phys* 65:061910.

Wang J, Wang W (2000): Modeling study on the validity of a possibly simplified representation of proteins. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topic* 61:6981–6986.

Wolf YI, Grishin NV, Koonin EV (2000): Estimating the number of protein folds and families from complete genome data. *J Mol Biol* 299:897.

Xia Y, Levitt M (2002): Roles of mutation and recombination in the evolution of protein thermodynamics. *Proc Natl Acad Sci USA* 99:10382.

Yampolsky LY, Kondrashov FA, Kondrashov AS (2005): Distribution of the strength of selection against amino acid replacements in human proteins. *Hum Mol Genet* 14:3191–3201.

Yin S, Ding F, Dokholyan NV (2007): Eris: an automated estimator of protein stability. *Nat Methods* 4:466–467.

Zeldovich KB, Chen P, Shakhnovich BE, Shakhnovich EI (2007): A first-principles model of early evolution: emergence of gene families, species, and preferred protein folds. *PLoS Comput Biol* 3: e139.

Author Query

1. Please check the full the inserted symbol “=” for correctness.