PROTEINS: Structure, Function, and Bioinformatics 54:622–628 (2004)

SHORT COMMUNICATION

What Is the Protein Design Alphabet?

Nikolay V. Dokholyan
Department of Biochemistry and Biophysics, School of Medicine, University of North Carolina at Chapel Hill, School of Medicine, Chapel Hill, North Carolina

ABSTRACT **Selecting a protein sequence that corresponds to a specific three-dimensional protein structure is known as the *protein design problem*. One principal bottleneck in solving this problem is our lack of knowledge of precise atomic interactions. Using a simple model of amino acid interactions, we determine three crucial factors that are important for solving the protein design problem. Among these factors is the *protein alphabet*—a set of sequence elements that encodes protein structure. Our model predicts that alphabet size is independent of protein length, suggesting the possibility of designing a protein of arbitrary length with the natural protein alphabet. We also find that protein alphabet size is governed by protein structural properties and the energetic properties of the protein alphabet units. We discover that the usage of average types of amino acid in proteins is less than expected if amino acids were chosen randomly with naturally occurring frequencies. We propose three possible scenarios that account for amino acid underusage in proteins. These scenarios suggest the possibility that amino acids themselves might not constitute the alphabet of natural proteins. Proteins 2004;54:622–628.** © 2004 Wiley-Liss, Inc.

Key words: protein design problem; protein design alphabet; amino acid usage

INTRODUCTION

Significant advances have been made in the past decade in uncovering the mechanisms by which an amino acid protein sequence folds into its unique three-dimensional (3D) conformation (the *protein folding problem*[1])[2–7] The inverse problem—the *protein design problem*—is to uncover the principles that enable us to manipulate protein structures by selecting (designing) a sequence that folds into a target 3D structure.[8,9] The ability to manipulate protein structure has immediate implications for our capability to alter a protein's function or even tailor it to our needs. The principal obstacles to the solution of the protein design problem are our lack of precise knowledge of atomic interactions and protein folding time scales that are inaccessible to detailed molecular dynamics (MD) simulations. With unlimited computer power, the protein design problem is best approached from the first physical principles. Various approximations to complex many-particle systems allow simplified treatments of the protein design problem and even resolve protein folding in computer simulations. The trade-off for such approximations is simplified and less accurate models of atomic interactions (known also as a force-field or a scoring function). Despite intensive developments in the protein design field, success has been limited to the redesign of known folds and the design of small proteins.[10–13] Improvements will come from more appropriate approximations to atomic interactions. Conceptually, it is important to understand what the important ingredients of these approximations are. Here, using a simple model of amino acid interactions, we develop a theoretical framework of protein design and uncover these important ingredients.

One of the postulates in the protein folding field is that most proteins are *thermodynamically stable* in their native states.[14–17] Importantly, the thermodynamic stability requirement has no implication for how stable natural proteins are.[18,19] In fact, it has been established that natural proteins can be "redesigned" to be more stable than the wild-type.[20,21] The thermodynamic stability requirement has been widely used to develop atomic interaction potentials.[15,22–26] We apply the protein stability requirement to a simple model of atomic interactions when total potential energy of a protein is the sum of all pairwise interactions between atoms.

METHODS

We assume that the Hamiltonian $H(\Gamma,S)$ for a protein conformation Γ with a given sequence of amino acids S is the sum of pairwise interaction potentials $u(\sigma_a,\sigma_b)$ between atoms of types σ_a and σ_b:

Grant sponsor: University of North Carolina at Chapel Hill Research Council.

*Correspondence to: Nikolay V. Dokholyan, Department of Biochemistry and Biophysics, School of Medicine, University of North Carolina at Chapel Hill, Campus Box 7260, Chapel Hill, NC 27599. E-mail: dokh@med.unc.edu

Received 28 April 2003; Accepted 9 September 2003

$$H(\Gamma, S) = \sum_{\sigma_a,\sigma_b}^{M} \sum_{i\neq j=1}^{N} u(\sigma_a, \sigma_b) s_i(\sigma_a) s_j(\sigma_b) \Delta_{ij}, \qquad (1)$$

where $s_i(\sigma_a) \equiv \delta(\sigma_a - \hat{\sigma}_i)$, $\hat{\sigma}_i$ is the atom type at a position i along the sequence S, $\delta(x)$ is 1 if $x = 0$, or 0 otherwise; Δ_{ij} is a contact matrix element, defined to be 1 or 0 depending on whether atoms i and j are geometrically separated by a distance smaller or greater than some threshold distance d_c. M is the number of distinct atom types (the *alphabet size*). For example, when we choose a coarse representation of proteins so that each atom type is represented by a distinct amino acid, then $M = 20$. In the case of the HP-model, in which amino acids are hydrophobic (H) or polar (P), $M = 2$.

The thermodynamic stability requirement postulates that the potential energy of the protein native conformation Γ is significantly smaller than any alternative conformation (*decoy*) Γ_D [i.e., $H(\Gamma,S) \ll H(\Gamma_D,S)$] for all decoys. Bryngelson and Wolynes[27] and Guttin and Shakhnovich[28] adopted the Random Energy Model (REM)[29] to proteins and shed light on the nature of the above inequality. They related the actual thermodynamic stability of a protein to the potential energy difference (*energy gap*) ΔH between the native state Γ and the lowest energy decoy Γ_C: $\Delta H = H(\Gamma,S) - H(\Gamma_C,S) < 0$. In the REM, the potential energy of a decoy is the sum of a number of pairwise contributions. Since the decoy space is large, using the central limit theorem, one can approximate the distribution of decoy potential energies by a Gaussian distribution. From the distribution of potential energies of the decoys, one can relate the energy gap ΔH to the free energy difference between a protein ground state and the unfolded state. The stability requirement is then $-\Delta H \gg k_B T$, where T is the temperature and k_B is a Boltzmann constant. Since the number of protein conformations available at a given temperature is a function of $k_B T$, the stability requirement can be rewritten as

$$Z \equiv \frac{\Delta H}{\sigma(H)} \ll -1, \qquad (2)$$

where $\sigma(H)$ is the root-mean-square deviation (RMSD) of the potential energy values of the decoys. Thus, the goal of protein design from the perspective of REM is to find such a sequence that minimizes the value of the Z-score ($Z < 0$). Here we question what the set of atomic interaction parameters $u(\sigma_i,\sigma_j)$ is that minimizes the Z-score.

Goldstein et al.[30] minimized the Z-score and found the *ideal potential* of pairwise atomic interactions. We perform similar minimization for an M-letter alphabet and, in addition to Goldstein et al.,[30], obtain the *actual* minimal value of the Z-score:

$$Z_{\min}^{(M)} = -\left[\sum_{\sigma_a,\sigma_b}^{M} K^2(\sigma_a, \sigma_b)\right]^{1/2} + \sqrt{2N \log \gamma}, \qquad (3)$$

where N is a protein length, γ is the number of conformations per atom,[8] and

$$K(\sigma_a, \sigma_b) = \frac{\sum_{i\neq j=1}^{M} s_i(\sigma_a) s_j(\sigma_b)(\Delta_{ij} - f_{ij})}{\left[\sum_{i\neq j=1}^{M} s_i(\sigma_a) s_j(\sigma_b) f_{ij}(1 - f_{ij})\right]^{1/2}}, \qquad (4)$$

and f_{ij} is the frequency of contacts between atoms i and j in all decoy conformations. In the Gō model,[31] in which atomic interactions are defined based on the native protein structure, the alphabet size is equal to the number of possible contacts in a protein. For the Gō model, Eq. (3) takes a simple form:

$$Z_{\min}^{(Go)} = -\left[\sum_{i\neq j=1}^{N} \frac{(\Delta_{ij} - f_{ij})^2}{f_{ij}(1 - f_{ij})}\right]^{1/2} + \sqrt{2N \log \gamma}. \qquad (5)$$

RESULTS AND DISCUSSION

Lessons for Protein Design

Eqs. (3) and (5) shed light on *three principal factors* of the protein design problem: (i) the alphabet size M, (ii) the number of atomic conformations γ, and (iii) the frequency of contacts f_{ij} of decoy conformations. The number of atomic conformations γ directly corresponds to protein flexibility, so the accuracy of the designed interaction potential depends on a protein model. In fact, as we increase a protein model's flexibility, its entropy increases so that the enthalpy, necessary for protein stability, increases. Representation of the decoy space is also crucial for the accurate identification of the interaction potential. It is especially important for *training* interaction potentials to satisfy Eq. (2), because misrepresentation of the secondary structure elements in the decoy space leads to biases in the interaction potentials (Ding and Dokholyan, unpublished results). The three principal factors of protein design may not be mutually independent; protein model flexibility, as well as alphabet size, affects the decoy space. These factors mirror Shakhnovich's *Lessons 2–4* for protein design.[8]

An important property of Eqs. (3) and (5) is that $Z_{\min}^{(M)}$ is a decreasing function of M, that is,

$$|Z_{\min}^{(M)}| \leq |Z_{\min}^{(M+1)}| \leq \ldots \leq |Z_{\min}^{(Go)}|. \qquad (6)$$

Inequalities [Eq. (6)] can be proven by noting that as we increase M, we distribute $K(\sigma_a,\sigma_b)$ among a larger number of terms, so that the proof rests on the following inequality:

$$\frac{(x_1 + x_2)^2}{y_1 + y_2} \leq \frac{x_1^2}{y_1} + \frac{x_2^2}{y_2}, \quad \forall x_1 \geq 0, x_2 \geq 0, y_1 > 0, y_2 > 0. \qquad (7)$$

The set of inequalities [Eq. (6)] provides another valuable, although intuitive, lesson for protein design: The larger the size of the design alphabet, the higher the stability of the best-designed protein.

While a number of studies have indicated that, using the Gō model, it is possible to consistently reach the protein native state in folding simulations,[32–34] the question arises: What is the minimal number of atom types M_c that nature

needs to encode a protein? From Eq. (3) we estimate this number by noting that the value of $Z_{\min}^{(M)}$ becomes 0 (the protein becomes unstable) when $M = M_c$. The sum $\Sigma_{\sigma_a,\sigma_b}^{M} K^2(\sigma_a, \sigma_b)$ in Eq. (3) has approximately $M^2/2$ terms, and their values are proportional to the number of amino acids in the protein N; thus, it can be written as $\Sigma_{\sigma_a,\sigma_b}^{M} K^2(\sigma_a, \sigma_b) = M^2N\varepsilon/2$, where ϵ is some factor independent of M and N. Therefore, the minimal size of the design alphabet is given by $M_c = 2\sqrt{\ln \gamma/\varepsilon}$ and is independent of N. This fact is significant because it posits that one can encode a protein of an arbitrary length with a natural protein alphabet. The minimal size of the alphabet is a function of the structural (γ) and energetic (ϵ) properties of proteins and amino acids correspondingly.

Protein Alphabet

One way to shed light on the protein alphabet is to study the usage of various types of amino acids in actual proteins. From the DALI[35,36] classification of protein domains—the elementary structural units of proteins—we compute the number of distinct amino acid types used m for various domain lengths N [Fig. 1(a)]. Interestingly, the dependence $m(N)$ is "sigmoidal" [i.e., $m(N)$ saturates at $N \approx 120$–150 amino acids]. This number defines the length of a typical protein domain that utilizes maximally diverse amino acid types.

For control, we compare observed to expected amino acid usage in the case when we select amino acid types at random from a pool of M possible types, each type occurring in nature with frequency p_i, ($i = 1 \ldots M$). We estimate the probability $P_{\exp}(m|N;M)$ that the amino acid usage in a domain of length N is m:

$$P_{\exp}(m|N;M) = \frac{1}{\Lambda} \sum_{n_1,\ldots,n_M=0}^{M} \frac{N!}{n_1!\ldots n_M!} p_1^{n_1}\ldots p_M^{n_M} \delta(n_1 + \ldots + n_M - N) \sum_{\wp(i_1,\ldots,i_{M-m})} \delta(n_{i_1})\ldots\delta(n_{i_{M-m}}), \quad (8)$$

where $\Lambda \equiv \Sigma_{m=1}^{M} P_{\exp}(m|N;M)$ is the normalization factor, and $\wp(i_1,\ldots,i_{M-m})$ indicates all permutations of indices $i_1,\ldots,i_{M-m}$. The first half of Eq. (8) includes all possible combinations of decompositions of N amino acids into $n_1,\ldots,n_M$ groups, so that $n_1 + \ldots + n_M = N$. The second half of Eq. (8) sets all possible $(M-m)$ n_i terms to zero, $n_i = 0$, so that only m n_i terms are nonzero. Eq. (8) can be reduced to

$$P_{\exp}(m|N;M) = \frac{1}{\Lambda} \sum_{\wp(i_1,\ldots,i_{M-m})} (p_{i_1}+\ldots+p_{i_m})^N \approx \frac{1}{\Lambda} \frac{M!}{m!(M-m)!} \left(\frac{m}{M}\right)^N. \quad (9)$$

In Eq. (9) we approximate $p_{i_1} + \ldots + p_{i_m}$ by m/M.[a] We plot the expected amino acid usage $m_{\exp}(N)$ that corresponds to the maximal value of $P_{\exp}(m|N;M)$ derived for such a control in Figure 1(a). We find that $m_{\exp}(N)$ dependence on N saturates at much lower than observed values of $N_{\exp} \approx 60$ amino acids, suggesting that amino acid usage in proteins is by far nonrandom. In fact, the expected probabilities of various amino acid usages m for average domain lengths corresponding to m do not exceed 10^{-3} [Fig. 1(b)], except for the trivial case when $m = 20$.

A striking observation can be made from Figure 1: There are a number of large domains with amino acid usage significantly smaller than one would expect if amino acids were chosen randomly. For example, the expected probability of observing an amino acid usage of 17 for a domain of size 300 amino acids is approximately 10^{-17}. The fact that we observe a number of domains that have low expected probability of amino acid usage suggests three not mutually exclusive scenarios: (i) the protein alphabet size is smaller than 20; (ii) the usage of some amino acids is avoided due to their chemical properties (Kuhlman, private communication); and (iii) amino acids themselves do not form a protein alphabet (i.e., primary sequence in juxtaposition with the correlations between amino acids along the sequence determine the protein 3D structure).

The first scenario is supported by a number of experimental, computational, and phenomenological studies. Riddle et al.[37] experimentally redesigned a small β-sheet protein, the SH3 domain, so that redesigned SH3 domain consisted of only 5 amino acid types. The expected probability for 5 amino acid usage in a protein containing 56 amino acids (Src SH3 domain) is $P_{\exp}(m = 5|N = 56;M = 8) \approx 10^{-29}$. The folding rates of the redesigned SH3 domain were not significantly different from those of the wild-type. Shakhnovich[38] demonstrated that, using lattice proteins and a pairwise interaction potential, one can design a stable lattice protein with a unique native state with an alphabet of 20, but not 2. In a similar study, Wang and Wang[39] showed that a reduced alphabet of only 5, but not 2, types of amino acids does not significantly alter the folding properties ("foldability"[40,41]) of lattice proteins. Using information theory and bioinformatics, Strait and Dewey[42] found that the Shannon entropy of protein sequences is approximately 2.5 bits per amino acids, which corresponds to an alphabet size of 5.7 amino acids. From the evolutionary perspective, Dokholyan and Shakhnovich[43] reproduced amino acid conservation patterns in 5 protein fold families using an alphabet of only 6 amino acid types in a simplified model of protein evolution. Here, we also determine the most probable alphabet size M_{mp} for a given average domain length N with a given amino acid usage m [i.e., such a value $M = M_{mp}$ that maximizes $P_{\exp}(m|N;M)$ for given m and N]. We find that M_{mp} is always equal to m. Since the smallest amino acid usage observed is 8, we set a limit on the size of protein alphabet of less than 8.

The second scenario for amino acid underusage is also likely due to possible chemical and structural restrictions on various amino acids. To test this scenario, we compare the usage for each individual amino acid in protein domains of lengths between 400 and 500 amino acids versus that of lengths smaller than 100. We find no significant

[a]The approximation made in Eq. (9) is valid when the amino acid usage $m \gg 1$. In Figure 1(b), we estimate $P_{\exp}(m|N;M)$ for $m \geq 12$. The difference between frequences of the 12 least occurring amino acids [largest source of error in Eq. (9)] and 12/20 = 0.60 is approximately 0.16. The sum in Eq. (9) is dominated by terms that are much smaller than 0.16.

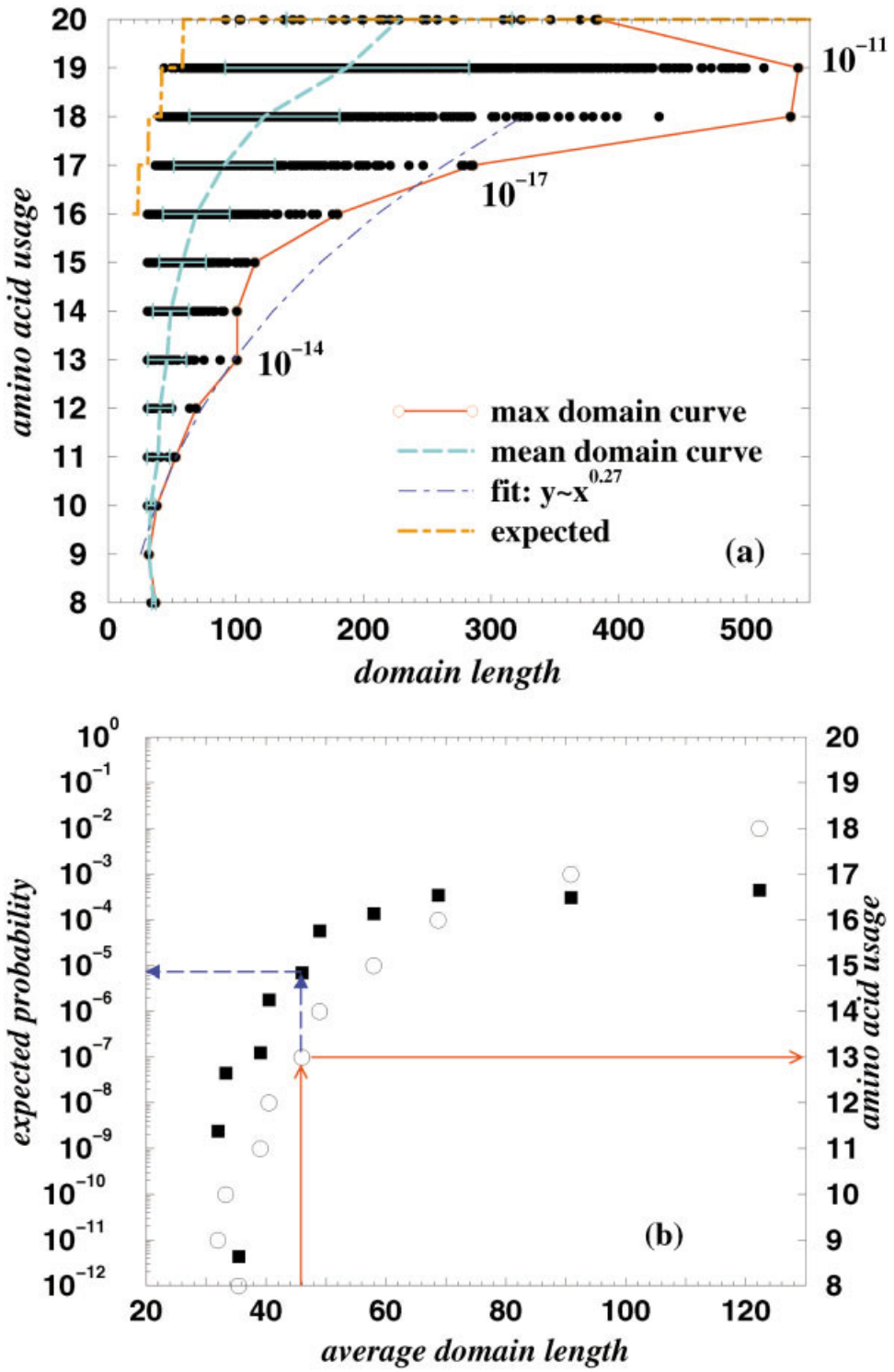

Fig. 1. (**a**) The number of distinct amino acid types used in DALI protein domains of various lengths N (●). The red solid curve traces the maximal domain length with a given amino acid usage, and the blue dot-dashed curve is its power–law fit. The blue dashed line traces the average domain length, along with the standard deviation. The dot-dashed yellow curve traces expected amino acid usage for a given domain length. For comparison, we put the values of probabilities ($p_{exp} = 10^{-14}, 10^{-17}, 10^{-11}$) to observe the amino acid usage of 13, 17, and 19, respectively, in protein domains of lengths 100, 290, and 540 correspondingly. (**b**) The expected probability $P_{exp}(m|N;M)$ (■; left y-axis) of the observed amino acid usage (○; right y-axis) for an average domain length corresponding to a given amino acid usage taken from Figure 1(a). For example, for an average domain length of 46 amino acids, the observed amino acid usage is 13 (red solid arrows); the corresponding expected probability $P_{exp}(m = 13|N = 46;M = 20) \approx 7 \cdot 10^{-6}$ (blue dashed arrows).

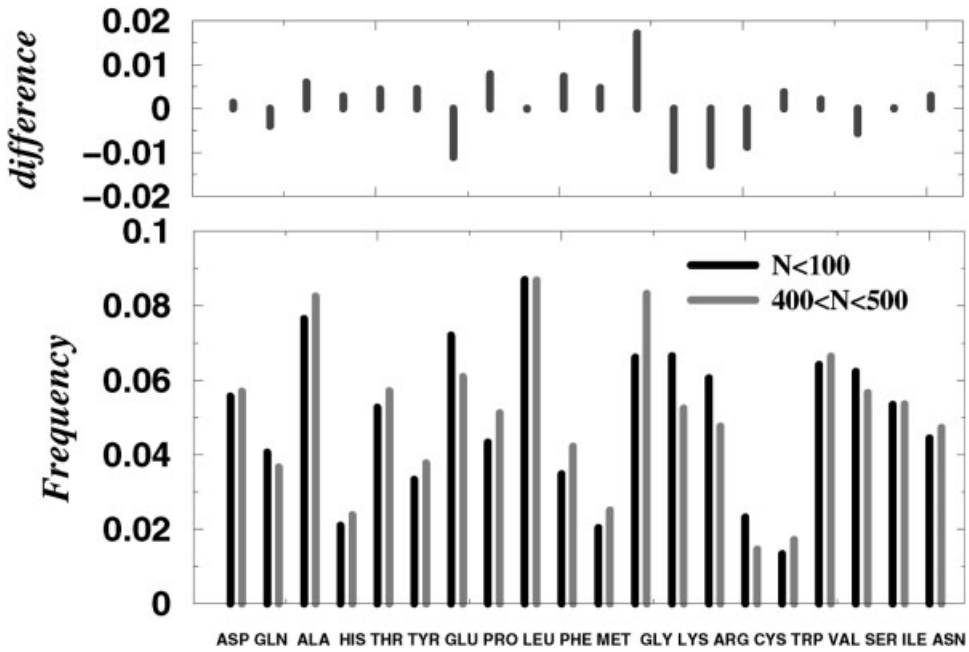

Fig. 2. The frequency of individual amino acid usage for protein domains of length no longer than 100 amino acids (black lines) compared to that for domains of length $400 < N < 500$ (gray lines). The differences between these frequencies for longer versus shorter domains are presented on the upper graph.

differences in amino acid usages except for the special amino acids Gly and Pro, charged amino acids Glu, Lys, Arg, and disulfide bridge-forming Cys. The number of charged amino acids is smaller in longer proteins, possibly because the surface area of globular proteins grows more slowly than volume as protein length increases, so that the number of hydrophobic and polar residues increases. However, since there is a large number of the latter, the differences between the usage of hydrophobic and polar residues in longer versus shorter protein domains are not significantly pronounced. Gly and Pro are used significantly more often in longer than in shorter domains, possibly because they play a crucial role in helix breaking and the formation of loops. Cys may be important in shorter proteins because formation of disulfide bonds stabilizes proteins. For longer proteins, formation of improper disulfide bonds may result in intermediates and slow the proteins' folding rates. Khare et al.[44] recently demonstrated in MD simulations of human Cu,Zn superoxide dismutase (SOD1) that SOD1 irreversible unfolding may be due to the formation of improper disulfide bonds between cysteins on a folding pathway. Abkevich and Shakhnovich[45] also observed an anticorrelation between the number of Cys residues and the aliphatic hydrophobic residues, in agreement with our findings (Fig. 2).

The third scenario of why the typical amino acid usage is less than 20 may be because amino acid usage does not represent the protein alphabet. It is a number of properties that include those of individual amino acids *and* correlations between physical properties of amino acids along protein sequences that determine the alphabet. So, in this case, the alphabet size is much larger than 20 and the pressure to use specific letters of the alphabet is diminished. There is also significant evidence for such a scenario. Chen and Stites[46] demonstrated the important role of three-body interactions in multiple-mutant studies of staphylococcal nuclease. Carter et al.[47] observed that frequencies of quadruples of amino acids that are in geometrical proximity to each other in proteins correlate with the changes in protein stabilities ($\Delta\Delta G$) upon mutations in these quadruplets. The question then is: What is the protein alphabet? The answer to this question may not lead to actual physical objects, but rather to some abstract objects that represent a superposition of a number of physical properties, such as partial charges, polarizations, the number of chemical bonds formed, and various quantum properties. An important goal for future studies then is to identify this alphabet and all possible physical properties that constitute it.

All three scenarios for why amino acid usage in proteins is smaller than 20 are not mutually exclusive. An understanding of the protein alphabet is essential to successful in silico protein design,[8,38,39] because the actual protein alphabet determines the parameter set for the energy function that is used for protein design.[b] Therefore, further studies are needed to uncover the true basis for encoding protein 3D structure. The knowledge of the protein alphabet will aid the development of the amino acid interaction models that capture protein native structure and folding mechanism and, thus, will allow the rational manipulation of protein structure and folding pathways. These studies may also shed light on the

[b]For example, the Hamiltonian in Eq. (1) is determined by $M(M-1)/2$ parameters.

redundancy of the genetic code and identify the principal rules of molecular evolution.

If the reason why amino acid usage in proteins is less than 20 is due to the smaller alphabet size, the question is then why has nature utilized 20 amino acids? It seems inefficient to develop and maintain an extensive machinery to produce/process "nonessential" amino acids. A possible answer to this question is that redundant amino acids preserve the stability of protein structures from a number of mutations that accumulate in protein sequences in the course of evolution. A physical explanation for protecting protein structures against mutations follows from Eqs. (1), (2), and (6): The larger the alphabet size, the larger the number of protein sequences that satisfy the stability requirement [Eq. (2)]. So a random substitution statistically preserves a protein's structure. Let us imagine a hypothetical situation when we add to the protein alphabet a new amino acid Ala′ that is very similar to Ala but encoded by a different triplet(s) of nucleotides. A mutation that substitutes Ala by Ala′ in a protein sequence does not alter the native protein stability, so that the structure and the function of the protein is preserved.

CONCLUSIONS

Using a simple model of atomic interactions, we demonstrate the three principal factors that must be addressed in the course of protein design: (i) protein alphabet, (ii) protein model flexibility, and (iii) unbiased sampling of the protein conformational (decoy) space. We also find that one can design a protein of an arbitrary length using an appropriate alphabet that is *consistent* with protein structure. In addition, we discover that, surprisingly, the number of amino acids used in proteins is much smaller than expected if they were chosen randomly with naturally occurring frequencies. We propose three possible scenarios to explain underusage of amino acids in proteins: (i) the actual protein alphabet size is smaller than 20; (ii) the usage of some amino acids is avoided due to their chemical properties; (iii) amino acids themselves do not form the protein alphabet (i.e., the primary sequence in juxtaposition with the correlations between amino acids along the sequence determine the protein 3D structure). These scenarios are not mutually exclusive and each is supported by a number of experimental, analytical, computational, and phenomenological studies.

ACKNOWLEDGMENTS

I am greatly indebted to S. V. Buldyrev, B. Kuhlman, and E. I. Shakhnovich for insightful discussions.

REFERENCES

1. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181:223–230.
2. Sali A, Shakhnovich E, Karplus M. How does a protein fold? Nature 1994;369:248–251.
3. Fersht AR. Nucleation mechanisms in protein folding. Curr Opin Struct Biol 1997;7:3–9.
4. Fersht AR. Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism. Proc Natl Acad Sci USA 2000;97:1525–1529.
5. Baker D. A surprising simplicity to protein folding. Nature 2000;405:39–42.
6. Bilsel O, Matthews CR. Barriers in protein folding reactions. Adv Protein Chem 2000;53:153–207.
7. Mirny L, Shakhnovich E. Protein folding theory; from lattice to all-atom models. Annu Rev Biophys Biomol Struct 2001;30:361–396.
8. Shakhnovich EI. Protein design: a perspective from simple tractable models. Fold Des 1998;3:R45–R48.
9. Bastolla U, Vendruscolo M, Knapp EW. A statistical mechanical method to optimize energy functions for protein folding. Proc Natl Acad Sci USA 2000:97:3977–3981.
10. Kamtekar S, Schiffer JM, Xiong HY, Babik JM, Hecht MH. Protein design by binary patterning of polar and nonpolar amino acids. Science 1993;262:1680–1685.
11. Dahiyat Bi, Mayo SL. De novo protein design: fully automated sequence selection Science 1997;278:82–87.
12. de la Paz ML, Lacroix E, Ramirez-Alvarado M, Serrano L. Computer-aided design of beta-sheet peptides. J Mol Biol 2001;312:229–246.
13. Kuhlman B, O'Neill JO, Kim DE, Zhang KYJ, Baker D. Accurate computer-based design of a new backbone conformation in the second turn of protein L. J Mol Biol 2002;315:471–477.
14. Shakhnovich EI, Gutin AM. Engineering of stable and fast-folding sequences of model proteins. Proc Natl Acad Sci USA 1993;90:7195–7199.
15. Goldstein RA, Luthey-Schulten ZA, Wolynes PG. Protein tertiary structure prediction using optimized Hamiltonians. Proc Natl Acad USA 1992;89:9029–9033.
16. Betancourt MR, Thirumalai D. Protein sequence design by energy landscaping. J Phys Chem B 2002;106:599–609.
17. Micheletti C, Seno F, Maritan A, Banavar JR. Protein design in the lattice model of hydrophobic and polar amino acids. Phys Rev Lett 1998;10:2237–2240.
18. Taverna DM, Goldstein RA. Why are proteins marginally stable? Proteins 2002;46:105–109.
19. Xia Y, Levitt M. Roles of mutation and recombination in the evolution of protein thermodynamics. Proc Natl Acad Sci USA 2002;99:10382–10387.
20. Chen J, Lu Z, Sakon J, Stites WE. Increase in the thermostability of staphylococcal nuclease: implications for the origin of protein thermostability. J Mol Biol 2000;303:125–130.
21. Dantas G, Kuhlman B, Callender D, Wong M, Baker D. A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. J Mol Biol 2003;332:449–460.
22. Maiorov VN, Crippen GM. Contact potential that recognizes the correct folding of globular proteins. J Mol Biol 1992;227:876–888.
23. Mirny LA, Shahnovich EI. How to derive a protein folding potential?: a new approach to an old problem. J Mol Biol 1996;264:1164–1179.
24. Seno F, Vendruscolo M, Maritan A, Banavar JR. Optimal protein design procedure. Phys Rev Lett 1996;77:1901–1904.
25. Vendruscolo M, Maritan A, Banavar JR. Stability threshold as a selection principle for protein design. Phys Rev Lett 1997;78:3967–3970.
26. Zhang, L, Skilnick J. How do potentials derived from structural databases relate to "true" potentials? Protein Sci 1998;7:112–122.
27. Bryngelson JD, Wolynes PG. Intermediates and barrier crossing in a random energy-model (with applications to protein folding). J Phys Chem 1989;93:6902–6915.
28. Guttin AM, Shakhnovich EI. Ground-state of random copolymers and the discrete random energy-model. J Chem Phys 1993;98:8174–8177.
29. Derrida B. The random energy-model. Physics Reports Review Section of Phys Lett 1980;67:29–35.
30. Goldstein RA, Lutheyschulten ZA, Wolynes PG. Protein tertiary structure recognition using optimized Hamiltonians with local interactions. Proc Natl Acad Sci USA 1992;89:9029–9033.
31. Go N. Theoretical study of protein folding. Annu Rev Biophys Bioeng 1983;12:183–210.
32. Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI. Identifying the protein folding nucleus using molecular dynamics. J Mol Biol 2000;296:1183–1188.
33. Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI. Discrete molecular dynamics studies of the folding of a protein-like model. Fold Des 1998;3:577–587.

34. Zhou YQ, Karplus M. folding thermodynamics of a model three-helix-bundle protein. Proc Natl Acad Sci USA 1997;94:14429–14432.
35. Dietmann S, Park J, Notredame C, Heger A, Lappe M, Holm L. A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. Nucleic Acids Res 2001;29:55–57.
36. Martin AC, Orengo CA, Hutchinson EG, Jones S, Karmirantzou M, Laskowski RA, Mitchell JB, Taroni C, Thornton JM. Protein folds and functions. Struct Fold Des 1998;6:875–884.
37. Riddle DS, Santiago JV, BrayHall ST, Doshi N, Grantcharova VP, Yi Q, Baker D. Functional rapidly folding proteins from simplified amino acid sequences. Nat Struct Biol 1997;4:805–809.
38. Shakhnovich EI. Proteins with selected sequences fold into unique native conformation. Phys Rev Lett 1994;72:3907–3910.
39. Wang J, Wang W. Modeling study on the validity of a possibly simplified representation of proteins. Phys Rev E 2000;61:6981–6986.
40. Klimov DK, Thirumalai D. Criterion that determines the foldability of proteins. Phys Rev Lett 1996;76:4070–4073.
41. Veitshans T, Klimov D, Thirumalai D. Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties. Fold Des 1997;2:1–22.
42. Strait BJ, Dewey TG. The Shannon information entropy of protein sequences. Biophys J 1996;71:148–155.
43. Dokholyan NV, Shakhnovich EI. Understanding hierarchical protein evolution from first principles. J Mol Biol 2001;312:289–307.
44. Khare S, Ding F, Dokholyan NV. Folding of Cu,Zn superoxide dismutase and familial amyotrophic lateral sclerosis. J Mol Biol 2003;334:515–525.
45. Abkevich VI, Shakhnovich EI. What can disulfide bonds tell us about protein energetics, function and folding?: simulations and bioinformatics analysis. J Mol Biol 2000;300:975–985.
46. Chen MM, Stites WE. Higher-order packing interactions in triple and quadruple mutants of staphylococcal nuclease. Biochemistry 2001;40:14012–14019.
47. Carter CW, LeFebvre BC, Cammer SA, Tropsha A, Edgell MH. Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. J Mol Biol 2001;311:625–638.