Hiking in the Energy Landscape in Sequence Space: A Bumpy Road to Good Folders

G. Tiana,1 R.A. Broglia,1,3,4 E.I. Shakhnovich*1,3,4
1The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark
2Department of Chemistry, Harvard University, Cambridge, Massachusetts
3Dipartimento di Fisica, Università di Milano, Milano, Italy
4INFN, Sez. di Milano, Milano, Italy

ABSTRACT With the help of a simple 20-letter lattice model of heteropolymers, we investigated the energy landscape in the space of designed good-folder sequences. Low-energy sequences form clusters, interconnected via neutral networks, in the space of sequences. Residues that play a key role in the foldability of the chain and in the stability of the native state are highly conserved, even among the chains belonging to different clusters. If, according to the interaction matrix, some strong attractive interactions are almost degenerate (i.e., they can be realized by more than one type of amino acid contacts), sequence clusters group into a few superclusters. Sequences belonging to different superclusters are dissimilar, displaying very small (<10%) similarity, and residues in key sites are, as a rule, not conserved. Similar behavior is observed in the analysis of real protein sequences. Proteins 2000;39:244–251.

INTRODUCTION

The degeneracy of protein folding code is well documented: many sequences exist that can fold to similar native conformations.1,2 Besides homologues, i.e., proteins that have a clear evolutionary connection, homologous sequences (differing by a few neutral mutations) and are often (but not always) functionally related, there exist analogues, i.e., structurally similar proteins that have nonhomologous sequences (with sequence identity less than 20%), unrelated functions, and no evident evolutionary relation.2,3 The analysis of the origin of analogues emphasizes the physical aspect of molecular evolution because they share common fold but not function.

An important question is whether present analogues emerged as a result of a long divergent evolution or originated from dissimilar sequences/structures and converged to structurally homologous folds. A physical approach to address this question is to study the topography of space of sequences that fold into the same target conformation. In particular, the connectivity of the space of sequences via neutral nets (i.e., single mutations that preserve the foldability into this structure4) may be a good evidence for divergent evolution as an origin of analogues,

whereas the presence of disconnected “isles” in sequence space would be an argument in favor of convergent evolution origin of analogs.

The aim of the present work is to address this question rigorously by analyzing the properties of the sequence space in an exactly tractable lattice model of a protein. For this purpose, we use a simple 20-letters, three-dimensional lattice model of heteropolymers5,6 and contact energies obtained from the statistical analysis of real proteins.7 Good-folder sequences are characterized by a large gap δ (compared with the standard deviation of the contact energies σ) between the energy of the sequence in the native conformation and the lowest energy (threshold Ec) of the conformations structurally dissimilar to the native one.6,8 In other words, good folders are associated with an “order parameter” ξ = δ/σ >> 1 (this quantity is closely related to the z-score9). Because the threshold energy Ec entering in the definition of the gap depends only on the composition of the chain,10 the knowledge of the energy landscape of the space of sequences with fixed amino acid composition enables us to select good folders and to investigate their properties. In particular, it is possible to study the space of folding sequences within the context of ground state properties by using only equilibrium thermodynamics.

The use of the lattice model to study properties of energy landscape in sequence space has two clear advantages. First, folding properties of each sequence can be examined directly from folding dynamics simulations that are quite feasible for such models. Second, all sequences designed to fold into a given conformation have the same length. This factor simplifies comparison between sequences, eliminating the need to introduce arbitrary gap/insertion penalties.

In particular, a single parameter, defined as the ratio of matching residues to the total number of residues (an overlap parameter) in a straightforward alignment of two sequences can serve as a natural measure of similarity between them. By using this parameter as a metric in the sequence space, we show that sequences folding to the
same native conformation cluster together in a rather articulated, hierarchical fashion, in the space of sequences. This clustering has direct evolutionary implications that we discuss in the end of this article.

THE MODEL

The model used in this calculation is a standard lattice model that presents amino acids as pointlike beads on the vertices of a cubic lattice, connected by rigid links (covalent bonds). Different monomers interact through the nearest-neighbor (in space) Hamiltonian

\[ H = \sum_{i<j} B_{\sigma_i\sigma_j} \Delta(r_i - r_j), \]

where \( \sigma_i \) and \( r_i \) are the kind and position of the \( i \)-th residue, and \( \Delta(r_i - r_j) \) is 1 if \( r_i \) and \( r_j \) are non-consecutive nearest neighbors and zero otherwise. The parameters \( B_{\sigma_i\sigma_j} \) are the contact energies between the amino acids.

The Sequence Space

As noted above, the similarity parameter \( q_{\alpha\beta}^a \) between sequence \( \alpha \) and sequence \( \beta \), defined as

\[ q_{\alpha\beta}^a = \frac{1}{N} \sum_{i=1}^{N} \delta(\sigma_i^\alpha, \sigma_i^\beta), \]

serves as a metric in sequence space. Here \( \delta(\sigma_i^\alpha, \sigma_i^\beta) \) is 1 if the \( i \)-th residue of sequence \( \alpha \) is equal to the \( i \)-th residue of sequence \( \beta \) and zero otherwise. A natural parameter to describe energy surface in sequence space is the distribution

\[ P(q_s) = <\delta(q_s - q_s^\beta)>, \]

where \( <> \) indicates a proper ("thermal") average over all possible pairs of sequences \( \alpha \) and \( \beta \).

To study the properties of the distribution \( P(q_s) \), we have chosen the 48-mer target structure (native conformation) shown in the inset to Figure 1. Using standard Monte-Carlo optimization in sequence space, we designed many sequences that have low energy in the target structure shown in Figure 1. The optimization in sequence space was conducted with constant amino acid composition corresponding to "average" composition in natural proteins. This condition implies optimization of the sequence fitness parameter \( \phi \) via optimization of the gap \( \delta \) keeping variance of interaction energies \( \sigma \) (which depends on amino acid composition only) fixed. The elementary MC move in sequence space that preserves aminoacid composition is swap of two randomly chosen residues. The acceptance of a "move" is controlled by applying a standard Metropolis criterion with "selective temperature" \( T \) that serves as a measure of a degree of evolutionary pressure toward sequences having larger energy gaps. The contact energies \( B_{\sigma_i\sigma_j} \) were taken from Table 5 of Myazawa and Jernigan. Folding of a a typical sequence generated at \( T = 0.01 \) (denoted as \( S_{48} \)) was studied by using standard lattice Monte-Carlo folding simulations and the thermodynamic and kinetic parameters for that sequence were determined. In the units we are considering \( (RT_{room} = 0.6 \text{ kcal/mol}) \), the energy of \( S_{48} \) in its native conformation is \( E_{nat} = -26.97 \), whereas the value \( E_c \) of the lowest dissimilar conformation is \( E_c = -21.50 \), leading to a gap \( \delta = 5.47 \) and to an "order
parameter $\xi = 18.2$. We are thus guaranteed that $S_{48}^{*}$ folds fast. Indeed, its mean first passage time to the native conformation is $3.3 \times 10^6$ MC steps at the optimal folding temperature. Also, any sequence $S'_{48}$ obtained from mutations that do not alter significantly the chain composition, in the native conformation, has an energy smaller than $E_n$, (i.e., an excitation energy lying within the gap $\delta$) will fold.\textsuperscript{16}

The sequence design procedure was performed at different selective temperatures $T$, recording each time 1,000 sequences with energy lower than a given threshold $E_{th} = E_n$ and separated by 1,000 MC steps. All these sequences fold to the native state within $10^8$ MC steps. From the collected sequences, it was possible to calculate the distribution $P(q_s)$ at different temperatures. The results are shown in Figure 1 for several simulation temperatures ranging from $T = 0.01$ to $T = 0.08$. The $P(q_s)$ plot recorded at $T = 0.01$ features a number of peaks most pronounced centered at $q_s = 0.1$, $q_s = 0.55$, and $q_s = 0.95$.\textsuperscript{5} The pronounced peak at $q = 0.55$ disappears in the $P(q_s)$ plots taken at higher temperatures. For example, $T = 0.08$ only two peaks are apparent: one at $q = 0.1$ and a set of smaller peaks in the high-$q$ region.

Because our MC design sampling probes different regions of sequence space at low-temperature simulations and at high-temperature ones, comparison of their results gives a deep insight into the nature of energy landscape in sequence space in a range of energies. Apparently, the low-temperature simulation at $T = 0.01$ samples the bottom part of the energy spectrum in sequence space. The peaks of $P(q_s)$ at $q_s \approx 1$ and at $q_s = 0.55$ indicate that the low-energy part of space of designed sequences can be divided into clusters, each cluster containing sequences that differ only by a few mutations. Furthermore, within each cluster, there are six to eight sites in which the associated residues are 100% conserved. These sites are the ones denoted as “hot” in Tiana et al.\textsuperscript{16} The size of these clusters varies considerably, but no sequence is isolated. This means that, for each sequence, it is possible to introduce a swapping of amino acids between some sites and still obtain a good folding sequence belonging to the same cluster. The maximum number of sequence changes that preserve a sequence within a cluster is around 15% of the length of the chain. If this limit is exceeded, a wider cooperative rearrangement of residues (between 40% and 60%, corresponding to the second peak in Fig. 1) is needed to produce a sequence that again displays an energy within the gap $\delta$ in the native conformation. These greater sequence rearrangements correspond to moves in the space of sequences, from one cluster to the next one.

Because all sequences belonging to a cluster are equivalent, i.e., they display >90% overlap among them, it is possible to compare different clusters by comparing sequences chosen at random from them. In other words, each cluster behaves like a class of equivalence. From a detailed analysis of the sequences belonging to different clusters, it turns out that clusters group into “superclusters.” Clusters within each supercluster are associated with a similarity parameter $q_s \sim 0.55$, whereas clusters belonging to different superclusters have $q_s \sim 0.1$.

Within each supercluster, “hot” sites are in the same position, and the type of residues occupying each of these sites is conserved in 100% of the sequences. This fact emphasizes once more the importance of “hot” sites to govern the low-energy properties of protein sequences. Furthermore, half of the “hot” sites of sequences belonging to the first supercluster are in the same positions as the “hot” sites of sequences of the second supercluster and the other half move to neighboring sites. The common “hot” sites are different for residues that are very similar (meaning that the columns of MJ matrix indicating their interaction with all the other monomers, are very similar, i.e., D with E and Q, K with R). This finding suggests that the presence of the two superclusters may be due to the quasi-degeneracy of these matrix elements. The substitution of a monomer in a “hot” site with a very similar one causes to a small extent a rearrangement of the other “hot” sites, and to a much larger degree the rearrangement of the other (“cold”) residues. To substantiate this scenario, we have repeated the calculation with a randomly generated matrix having the same distribution of elements of the MJ matrix. Because in this case all the columns of the matrix are different, we expect no superclusters. In fact, the distribution $P(q_s)$ for this case is very similar to the distribution shown in Figure 1, except for the absence of the peak at $q_s = 0.1$ (data not shown).

The schematic illustration of the low-energy part of the sequence space is sketched in Figure 2A. For each folder sequence, it is possible to mutate a few “cold” sites, up to 15% of them and still obtain a good folding sequence. This produces a cluster. If more “cold” sites are changed,
HIKING IN THE ENERGY LANDSCAPE IN SEQUENCE SPACE

Figure 2.
there is a partial rearrangement of many other “cold” sites to keep the energy low, leading to different clusters. When a “hot” site is mutated with a similarly strongly interacting residue (if the interaction matrix allows for such a possibility), the other “hot” sites become slightly rearranged, whereas the “cold” sites are heavily rearranged, the net result being a new supercluster.

An important aspect of this analysis concerns the height of the “barriers” in the sequence space. This question is closely related to the issue of connectivity of clusters via neutral network, i.e., set of mutations that preserve the ability of all intermediate sequences to fold into the same structure. Although the question of what constitutes a viable folding sequence is a complicated one having a number of aspects (thermodynamic, kinetic, and functional) a necessary condition for a sequence to fold into a given structure is that its energy in this structure is lower than $E_c$. From this “minimalistic” point of view, the clusters are connected by neutral networks, whereas superclusters are not—the top of the barrier in sequence space between them lies above the level of $E_c$, i.e., the mutation “pathway” between superclusters passes through the “sea” of non-folding sequences. This is schematically illustrated in Figure 2A.

The schematic representation of the structure and connectivity of the part of the sequence space sampled at higher temperatures (i.e., that fits the native structure with higher energy but still lower than $E_c$) is shown in Figure 2B. The main difference between $T = 0.01$ and higher temperature simulations is that the peak of $P(q_s)$ at $q_s = 0.55$ does not exist at higher temperature sampling. This indicates that the structure of the sequence space sampled at $T = 0.08$ does not include apparent clusters. However, the superclusters remain the same with the same “hot” sites conserved in each of them. Figure 2B(c) schematically represents the energy landscape in the sequence space obtained both from low-temperature sampling and high-temperature sampling. It clearly shows that there are two characteristic barriers: between clusters and between superclusters. When the sampling temperature exceeds the former, the apparent picture of clusters disappears and the simulations sample broad basins each corresponding to a supercluster, in our terminology.

**Stability of the Results**

To proceed further, the universality of these findings has to be tested against the conformation chosen as native and to the interaction matrix. We have analyzed the distribution $P(q_s)$ for 10 different compact conformations of a 48-monomer chain. In all cases the distribution $P(q_s)$ taken at low temperature $T = 0.01$ exhibits three peaks and the last of them is centered at $q_s \approx 0.9$. The two peaks observed at lower values of $q_s$ are always well separated from the other peaks by a minimum at $q_s \approx 0.7$. The position and width of the other peaks seem to depend on the specific conformation. This is in agreement with the interpretation we gave of the existence of these peaks. If they are caused by rearrangements of a substantial part of the sequence, the number of amino acids involved in these rearrangements can depend on the details of the geometry of the system and thus the degree of similarity that the mutated sequences will display with each other.

To study the dependence of $P(q_s)$ on the interaction matrix, we repeated the same calculations with a randomly generated matrix (see above) and with the matrix published by Kolinski et al. Unlike the MJ matrix, these do not display any similar columns and, consequently, the distribution $P(q_s)$ shows only two peaks. The peaks are located at $q_s = 0.55$ and $q_s = 0.95$ when simulations are run at $T = 0.01$. At higher temperature the “left” peak moves to lower values of $q_s \approx 0.2$. Note, in this case for each structure, the type of residue occupying “hot” sites is unique because with that set of parameters only one type of interactions (hydrophobic) can provide substantial stabilization to the hot sites. Hence, in this case the superclusters probably do not exist in the energy landscape in sequence space.

In the case of models containing only two kinds of residues (hydrophobic and polar), the assertion that $E_c$, the lowest boundary energy of states that are structurally different from the native one, depends only on the composition of the chain, may not hold. Therefore, in this case it may not be possible to describe in a simple fashion the space of large gap sequences in terms of ground-state properties. Nonetheless, it is still necessary (but now not sufficient) for designed sequences to have, in the native conformation, low energy compared, for example, to the average energy of compact configurations. Calculations performed with the matrix elements used in Ref. show that $P(q_s)$ has only one peak close to $q_s = 1$.

We also analyzed sequences composed of three kinds of residues. The behavior in this case is more similar to that found for 20 kinds of residues. Three seems then to be the minimum quantity of different residues necessary to have a non-trivial shape of $P(q_s)$.

Furthermore, we repeated the calculations by optimizing the sequences with an algorithm that does not require the constraint of constant amino acid composition of the chain. This algorithm maximizes the stability of the sequences, approximating the free energy of non-native states with a high-temperature expansion. The resulting shape of $P(q)$ is very similar to the one shown in Figure 1, except for the fact that the low-$q$ peak is very small. This last feature seems to be a non-physical consequence of the high-temperature approximation.

**Comparison With Real Sequences**

An analysis of the distribution of the parameter $q_s$ associated with real proteins was conducted by Rost. The values of $q_s$ are obtained by the comparison of pairs of sequences that fold to similar three-dimensional structures. Sequences and structures were selected from the PDB database in a way to minimize bias due to the finiteness of the set. The function $P(q_s)$ thus obtained displays two well-defined peaks centered at $q_s = 0.08$ and $q_s \approx 1$ and a continuum of smaller peaks in between. The overline on the distribution of $q_s$ indicates that the results
of Rost having the meaning of an average of $P(q_s)$, as defined in Equation (2), over many different target structures.

The results of Rost for real sequences seem to match better the pattern of $P(q_s)$ obtained in our simulations at $T = 0.08$ mutations. As was already stated, the energy landscape in sequence space is complex so that how it is sampled may affect the apparent observed correlations. Specifically, it is important to make proper weighting of close homologues. Clearly, exclusion or underweighting of sampled may affect the apparent observed correlations.

Within each family ("supercluster" ("hot" sites, see above). At high temperature (total 51 families listed in Fig. 2 of Mirny and Shakhnovich). Proteins within each family are related by sequence homology not less than 25%. However, sequence homology between families is low, less than 25% despite the fact that structurally proteins belonging to those 51 families are similar. To cope with this fact, we align families against each other using structural alignment. One should be careful to weigh large and small families equally. First, we compute $p_i^m(l)$, the frequency of residue type $i$ at position $l$ within each family $m$. Next, we compute the across-family frequency

$$P_i(l) = \frac{1}{51} \sum_{m=1}^{51} p_i^m(l),$$

and the across-family entropy

$$S_{across}(l) = -\sum_{i=1}^{20} P_i(l)\log P_i(l).$$

$$S_{across}(l) = -\sum_{i=1}^{20} p_i^m(l)\log p_i^m(l).$$

$$S_{across}(l) = -\sum_{i=1}^{20} p_i(l)\log p_i(l).$$

Fig. 3b shows the resulting distribution of sequence entropies $S_{across}$ in structurally aligned families of proteins having Immunoglobulin (Ig) fold. We see a remarkable qualitative agreement between low-temperature lattice model results for designed sequences (presented in Fig. 3a) and that for natural sequences of real proteins having Ig-fold. Both show a pronounced bimodal distribu-
tion. As explained above, for the case of lattice sequences, the bimodal distribution of individual entropies is related to the complicated structure of the energy landscape in sequence space. The same feature observed in one of the most populated protein families is suggestive of complex energy landscape in sequence space of real proteins.

**DISCUSSION**

Our analysis suggests that the space of sequences that fold into the same native conformations is quite complicated: it is partitioned into clusters and superclusters. Sequences that belong to different clusters have little similarity despite the fact that they are able to fold into the same native conformation. Nevertheless, sequences belonging to the same supercluster (but different clusters) have same amino acids in a few strategic “hot” core positions. Finally, amino acids in core positions vary between superclusters.

These results bear striking similarity to recent findings for real proteins. The analysis presented in Mirny et al. and Mirny and Shakhnovich revealed universally conserved core positions in several protein folds. In some cases the analogues had the same types of amino acids in these positions despite the lack of sequence similarity anywhere else. However, in several cases correlated mutations in the core that preserved strong stabilization of the core but with different sets of amino acids were observed. For example, in some cases strong aliphatic hydrophobic contact was replaced by disulfide bond. Another example is aromatic rescue of glycine in β-sheets when two aliphatic hydrophobic groups in cores of β-sheets can be replaced by glycine and an aromatic group without loss of stability. These situations correspond to transitions from one supercluster to another, using the terminology of the present article.

A non-trivial distribution of barriers translates into complex dynamics with multiple discrete timescales. In our case, the results relate to evolutionary dynamics. In particular, we found that evolutionary dynamics may proceed with distinctly different rates for different parts of a protein.

Our study may provide a meaningful “toy” model of convergent and divergent evolution. The analysis shows that transitions between clusters occur over the barriers that do not reach $E_c$. In other words, clusters may be connected by neutral networks of sequences that all fold into the same native conformations, albeit with different stabilities. However, superclusters are separated by barriers that are high enough to exceed $E_c$ at their tops (see Fig. 2). This means that it is not possible to pass from supercluster to another supercluster via a neutral network. Correspondingly, sequences that belong to the same cluster could have appeared as a result of divergent evolution, whereas sequences that belong to different superclusters are likely to appear as a result of convergent evolution. Translating this to the language of evolutionary physical analysis of real proteins, we assert that in cases when protein analogues contain different types of amino acids in universally conserved cores that may have been the product of convergent evolution, whereas similar amino acids in the core may be suggestive of divergent evolution origin of analogs.

In this article mutations were generated by local “moves” in sequence space: amino acid swaps. Using point mutations as elementary moves allows the search of the sequence space in a very detailed way because many mutations are accepted. The Boltzmann-based criterion of acceptance/rejection of mutations is a most computationally efficient “canonical” way to enforce energetic stability condition for accepted sequences (more detailed discussion of this issue is in Shakhnovich and Gutin and Shakhnovich). It provides a satisfactory description of the energy landscape in sequence space. Because no constraints are imposed in sequence space (like polymer bond constraints in configurational space) the ergodicity of the simulation in sequence space is not an issue. Apparently, any two sequences can be connected by a set of point mutations. This provides further corroboration to our approach to scan sequence space of model proteins. However, point mutations may not be the only way how actual biological evolution of protein sequences occurred, especially in higher organisms. More complex “moves” in sequence space such as gene recombinations and duplication may be needed to provide more realistic models of biological evolution that resulted in modern proteins in modern organisms. This may be the subject of future studies.

Another possible physical evolutionary pressure that was not taken explicitly into account in our present calculations is the one toward fast-folding kinetics. It was argued in Mirny and Shakhnovich that such factors as the need to prevent aggregation or proteolysis may lead to effective pressure toward fast-folding sequences. However, it was shown in simulations and for some proteins (J. Clarke, personal communication) that the core residues conserved within clusters are also the ones that participate in nucleation and thus determine the folding rate. Therefore, for such proteins stability will be highly correlated with folding rate, and evolutionary regulation of both factors may result in stabilization of the same set of contacts. However, it was noted in Mirny and Shakhnovich that this is not always the case: a counterexample is cold-shock protein family where stability and fast folding are likely to be determined by different sets of contacts. In these cases, the evolutionary optimization of folding kinetics may result in additional conserved cluster in the nucleation region. Obviously, future experimental and theoretical studies will help to identify the physical and biological reason for sequence conservation in structures of most protein families.

**ACKNOWLEDGMENTS**

This work has been partially supported by the Graduate School of Biophysics, University of Copenhagen (to G.T.) and National Institutes of Health (grant GM52126 to E.S.).
REFERENCES

1. Branden C, Tooze J. Introduction to protein structure. New York: Garland Publishing, Inc.; 1998.
2. Brenner S, Chothia C, Hubbard T. Population statistics of protein structures: lessons from structural classifications. Curr Opin Struct Biol 1997;7:369–376.
3. Rost B. Protein structures sustain evolutionary drift. Folding Design 1997;2:S19–S24.
4. Govindarajan S, Goldstein R. Evolution of model proteins on a foldability landscape. Proteins 1997;29:461–466.
5. Shakhnovich E, Gutin A. Exhaustive enumeration of all conformations of compact heteropolymers with quenched disordered sequence of links. J Chem Phys 1990;93:5967–5971.
6. Sali A, Shakhnovich E, Karplus M. Kinetics of protein folding: a lattice model study for the requirements for folding to the native state. J Mol Biol 1994;235:1614–1636.
7. Myazawa S, Jernigan R. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1988;18:534–552.
8. Goldstein R, Luthey-Schulten Z, Wolynes P. Optimal protein-folding codes from spin-glass theory. Proc Natl Acad Sci USA 1992;89:4918–4922.
9. Bowie J, Luthey R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–169.
10. Shakhnovich E. Proteins with selected sequences fold to their unique native conformation. Phys Rev Lett 1994;72:3907–3910.
11. Shakhnovich E, Gutin A. Engineering of stable and fast-folding sequences of model proteins. Proc Natl Acad Sci USA 1993;90:7195–7199.
12. Shakhnovich E, Gutin A. A novel approach to design of stable proteins. Protein Eng 1993;6:793–800.
13. Creighton T. Proteins: structure and molecular properties. W.H. Freeman; New York, 1992.
14. Hilhorst H, Deutch J. Analysis of monte-carlo results on the kinetics of lattice polymer chains with excluded volume. J Chem Phys 1975;63:5153–5161.
15. Broglia R, Tian G, Roman H, Vigezzi E, Shakhnovich E. Stability of designed proteins against mutations. Phys Rev Lett 1999;82:4727–4730.
16. Tian G, Broglia R, Roman H, Vigezzi E, Shakhnovich E. Folding and misfolding of designed proteinlike chains with mutations. J Chem Phys 1998;108:757–761.
17. Mirny L, Abkevich V, Shakhnovich E. How evolution makes proteins fold quickly. Proc Natl Acad Sci USA 1998;95:4976–4981.
18. Govindarajan S, Goldstein RA. On the thermodynamic hypothesis of protein folding. Proc Natl Acad Sci USA 1998;95:5545–5549.
19. Kolinski A, Godzik A, Skolnick J. The general method for the prediction of the three-dimensional structure and folding pathway of globular proteins: application to designed helical proteins. J Chem Phys 1993;98:7420–7433.
20. Sfatos C, Gutin A, Shakhnovich E. Phase diagram of random copolymers. Phys Rev E 1993;48:465.
21. Deutsch J, Kurosky T. New algorithm for protein design. Phys Rev Lett 1996;76:323–326.
22. Li H, Wingreen N, Tang C. Emergence of preferred structures in a simple model of protein folding. Science 1996;273:666–669.
23. Morrissey M, Shakhnovich E. Design of proteins with selected thermal properties. Folding Design 1996;1:391–406.
24. Shakhnovich E, Abkevich V, Pitsyn O. Conserved residues and the mechanism of protein folding. Nature 1996;379:98–98.
25. Mirny L, Shakhnovich E. Universally conserved residues in protein folds: reading evolutionary signals about protein function, stability and folding kinetics. J Mol Biol 1999;291:177–196.
26. Merkel J, Reagan L. Aromatic rescue of glycine in β-sheets. Folding Design 1998;3:449–455.
27. Perl D, Wolker C, Schindler T, Shroder K, Marahiel M, Jaenicke R, Schmid F. Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. Nat Struct Biol 1998;5:229–235.