1-1-2003

Long- and short-range interactions in native protein structures are consistent/minimally frustrated in sequence space

Sanzo Miyazawa
National Institutes of Health

Robert L. Jernigan
Iowa State University, jernigan@iastate.edu

Long- and short-range interactions in native protein structures are consistent/minimally frustrated in sequence space

Abstract
We show that long- and short-range interactions in almost all protein native structures are actually consistent with each other for coarse-grained energy scales; specifically we mean the long-range inter-residue contact energies and the short-range secondary structure energies based on peptide dihedral angles, which are potentials of mean force evaluated from residue distributions observed in protein native structures. This consistency is observed at equilibrium in sequence space rather than in conformational space. Statistical ensembles of sequences are generated by exchanging residues for each of 797 protein native structures with the Metropolis method. It is shown that adding the other category of interaction to either the short- or long-range interactions decreases the means and variances of those energies for essentially all protein native structures, indicating that both interactions consistently work by more-or-less restricting sequence spaces available to one of the interactions. In addition to this consistency, independence by these interaction classes is also indicated by the fact that there are almost no correlations between them when equilibrated using both interactions and significant but small, positive correlations at equilibrium using only one of the interactions. Evidence is provided that protein native sequences can be regarded approximately as samples from the statistical ensembles of sequences with these energy scales and that all proteins have the same effective conformational temperature. Designing protein structures and sequences to be consistent and minimally frustrated among the various interactions is a most effective way to increase protein stability and foldability.

Keywords
empirical potentials, inverse protein folding, protein folding, protein sequence design, protein sequence-structure compatibility

Comments
This article is published as Miyazawa, S. and Jernigan, R. L. (2003), Long- and short-range interactions in native protein structures are consistent/minimally frustrated in sequence space. Proteins, 50: 35–43. doi: 10.1002/prot.10242.

PROTEINS: Structure, Function, and Genetics 50:35–43 (2003)

Long- and Short-range Interactions in Native Protein Structures Are Consistent/Minimally Frustrated in Sequence Space

Sanzo Miyazawa[1,2,3] **and Robert L. Jernigan**[2,3]
[1]*Faculty of Technology, Gunma University, Kiryu, Gunma, Japan*
[2]*Laboratory of Experimental and Computational Biology, CCR, National Cancer Institute, National Institutes of Health, Bethesda, Maryland*
[3]*Laurence H. Baker Center for Bioinformatics and Biological Statistics, Plant Science Institute and Biochemistry, Biophysics, and Molecular Biology, Iowa State University, Ames, Iowa*

ABSTRACT **We show that long- and short-range interactions in almost all protein native structures are actually consistent with each other for coarse-grained energy scales; specifically we mean the long-range inter-residue contact energies and the short-range secondary structure energies based on peptide dihedral angles, which are potentials of mean force evaluated from residue distributions observed in protein native structures. This consistency is observed at equilibrium in sequence space rather than in conformational space. Statistical ensembles of sequences are generated by exchanging residues for each of 797 protein native structures with the Metropolis method. It is shown that adding the other category of interaction to either the short- or long-range interactions decreases the means and variances of those energies for essentially all protein native structures, indicating that both interactions consistently work by more-or-less restricting sequence spaces available to one of the interactions. In addition to this consistency, independence by these interaction classes is also indicated by the fact that there are almost no correlations between them when equilibrated using both interactions and significant but small, positive correlations at equilibrium using only one of the interactions. Evidence is provided that protein native sequences can be regarded approximately as samples from the statistical ensembles of sequences with these energy scales and that all proteins have the same effective conformational temperature. Designing protein structures and sequences to be consistent and minimally frustrated among the various interactions is a most effective way to increase protein stability and foldability. Proteins 2003;50:35–43.** © 2002 Wiley-Liss, Inc.

Key words: empirical potentials; inverse protein folding; protein folding; protein sequence design; protein sequence-structure compatibility

INTRODUCTION

The feature of consistency among interactions in protein native structures was first noticed by Ptitsyn and Finkelstein[1] and by Go[2] as an effective way for proteins to increase their structural stabilities. There are several pieces of evidence that explicitly show such a consistency between interactions. It was reported that native side-chain conformations could be well predicted by taking account of only side chain-backbone interactions if all backbone atoms are fixed in their native conformation,[3] although the side chain-side chain interactions contribute to the stabilization of the native conformations of side chains.[4] This fact supports an overall consistency between side chain-side chain and side chain-backbone interactions.

On the other hand, Bryngelson and Wolynes[5] observed the complexity of the energy surface in protein conformational space and pointed out that the energy landscape for natural proteins must be minimally frustrated between smooth and rough energy landscapes and must resemble funnels for proteins to fold within reasonable times. A rough energy landscape, a frustrated situation that is caused by many competing interactions, is a characteristic of random heteropolymers, which often exhibit glass transitions where the system can be trapped in one of the lower energy states at a transition temperature. To the contrary, natural proteins have a unique feature of smoothing the energy landscape on a coarse-grained conformational scale and hence are capable of folding into single stable structures within a limited time.[6] Thus, minimal frustration and consistency between interactions are essential for the stability and foldability of protein structures. This concept was also used to define independently folding units.[7]

Here, we examine whether or not short- and long-range interactions between residues, i.e., locally or distantly located along amino acid chains, in protein native struc-

Correspondence to: Sanzo Miyazawa, Faculty of Technology, Gunma University, Kiryu, Gunma 376-8515, Japan.
E-mail: miyazawa@smlab.sci.gunma-u.ac.jp

Received 1 April 2002; Accepted 16 July 2002

tures are consistent with each other for sequence selection of the more stable sequences for each protein. Sequence space is searched instead of conformational space by exchanging amino acids within each protein. This maintains the composition. Among many kinds of interactions operative in protein structures, the two classes of interactions playing different roles in protein folding are chosen to be examined here. Then, consistencies between short- and long-range interactions are examined for their effects on the mean and the variance of interaction energies at statistical equilibrium in sequence space. Random sequence samples are generated with the Metropolis Monte Carlo method[8] by exchanging amino acids in the native structure of each protein. The long- and short-range interaction potentials used here are the contact potentials[9–11] between the 20 kinds of amino acids and the secondary structure potentials,[12] based on peptide dihedral angles evaluated according to the methods previously reported by us.

It is shown that adding the other class of interaction to either contact energy or secondary structure energy decreases the mean and also the variance of those energies at equilibrium in sequence space for the almost all protein native structures. There is virtually no correlation between these types of energies at statistical equilibrium when the equilibration uses both classes of interactions, but almost all proteins show positive covariances between the two classes for the statistical equilibrium using only one class. These facts indicate that both classes of interactions work consistently by restricting the sequence space available to one class of interactions but that they act almost independently near the minimum in the total energy surface.

It is also shown here that contact energies and secondary structure energies of almost all native proteins fall within ± 2 SDs of the equilibrium mean for each protein and that there is no correlation between them. These features indicate that protein native sequences can be regarded approximately as samples from the statistical ensembles of sequences with these energy scales and that in addition all proteins have the same conformational temperature.[13]

MATERIALS AND METHODS

Stability of Protein Sequence and Structure

The stability of a specific conformation for a protein sequence is determined relative to the whole ensemble of protein conformations. Let us define an effective free energy, $\mathcal{F}$, to represent the stability of a sequence-structure pair, i.e., probability $P(s, i)$ of a specific conformation s for sequence i, as

$$\beta\mathcal{F}(s|i) \equiv -\log(P(s|i)) \tag{1}$$

$$= \beta E^{conf}(s, i) + \log\left(\sum_s \exp(-\beta E^{conf}(s, i))\right) \tag{2}$$

where β is equal to $1/kT$, k is the Boltzmann constant, T is temperature, $E^{conf}(s, i)$ is the conformational energy of the conformational state s of sequence i, and the sum is taken over all possible conformations.

How can we estimate the second term of Equation 2, which serves as a reference energy for $\mathcal{F}$ to measure relative protein stability? Analyses using the Random Energy Model approximation suggest that the contribution to the partition function from non-native conformations depends primarily on amino acid composition rather than on sequence at sufficiently high temperature $T > T_c$, where T_c is the temperature of the "freezing" transition in a random heteropolymer having the same amino acid composition.[18,19] For sequence space optimization of simple lattice proteins, estimating Equation 2 has been attempted directly. The partition function was estimated by dual Monte-Carlo simulations,[15] by the first cumulant in a high-temperature expansion,[16] and by a cumulant expansion approximation.[17] Z scores have been successfully used instead of energy for sequence space optimization of simple lattice proteins under the unrestrictive condition of amino acid composition.[14]

Here, the second term in Equation 2 is approximated by including only dominant terms in the summation of Boltzmann factors over all conformations, i.e., only native-like compact conformations. Native-like conformations mean compact conformations in which nonpolar residues tend to be located inside the globules. In other words, we consider only conditions under which such native-like conformations are dominant in the conformational ensemble. Then the log function is evaluated in a high-temperature expansion approximation; this is similar in concept to the work of Deutsch and Kurosky.[16]

$$\begin{aligned}
&\log\left(\sum_s \exp(-\beta E^{conf}(s, i))\right) \\
&\quad \simeq \log\left(\sum_{s \in \{\text{native-like}\}} \exp(-\beta E^{conf}(s, i))\right) \\
&\quad \simeq \log\left(\sum_{s \in \{\text{native-like}\}} 1\right) \\
&\qquad - \beta \sum_{s \in \{\text{native-like}\}} E^{conf}(s, i) \Big/ \left(\sum_{s \in \{\text{native-like}\}} 1\right) \\
&\quad \simeq n_r\sigma - \beta\langle E^{conf}(s, i)\rangle_{\beta = 0,\ \text{native-like conf.}}
\end{aligned} \tag{3}$$

where n_r is the sequence length of a protein and σ is a constant to represent the conformational entropy per residue in k units for native-like structures. Thus, the effective free energy $\mathcal{F}(s, i)$ to represent the stability of conformation s and sequence i may be approximated as

$$\beta\mathcal{F}(s|i) \simeq \beta\mathcal{E}(s, i) + n_r\sigma \tag{4}$$

where $\mathcal{E}(s, i)$ is the conformational energy relative to the average over native-like conformations;

$$\beta\mathcal{E}(s, i) \equiv \beta E^{conf}(s, i) - \beta\,\langle E^{conf}(s, i)\rangle_{\beta = 0,\ \text{native-like conf.}} \tag{5}$$

Coarse-grained Conformational Energy

Short- and long-range interaction energies considered here are specifically those for secondary structure interactions and pairwise interactions at the residue level in protein structures. Both interactions are estimated by using simple empirical potentials, a secondary structure potential[12] and a contact potential,[9–11] which are potentials of mean force evaluated from the observed distributions of peptide dihedral angles (ϕ, ψ) and contacting residue pairs. For the secondary structure potential, the interaction potential between a tripeptide in a conformational state (s_{-1}, s_0, s_1) and a side chain of type i_0 located at its center, $e^s \equiv \delta e^s(s_{-1}, i_0, s_0, s_1)$ defined in Equation 6 of Miyazawa and Jernigan,[12] is employed here; peptide conformations were classified here into five states, α, β, proline β, left-handed α, and left-handed β, based on peptide dihedral angles. Intrinsic energies of backbone-backbone interactions are not included in the secondary structure energies, because they are invariant for a given structure. The total secondary structure energy $E^s(s, i)$ of conformation s for sequence i is represented as

$$E^s(s, i) = \sum_p \delta e^s(s_{p-1}, i_p, s_p, s_{p+1}) \tag{6}$$

$$\langle E^s(s, i)\rangle_{\beta=0,\ \text{native-like conf.}} \simeq \sum_p \langle \delta e^s(s_{p-1}, i_p, s_p, s_{p+1})\rangle_{\text{all natives}} \tag{7}$$

where i_p and s_p mean the type of amino acid and the conformational state of residue at position p in sequence. The unweighted averages of $\delta e^s(s_{p-1}, i_p, s_p, s_{p+1})$ over native-like conformations are approximated as the averages over all positions of residue type i_p in all protein structures.

Likewise contact energies having as a reference the collapse energy e_{rr}, that is, $e_{ij} - e_{rr}$ are used here; the total collapse energy depends on structures but not on sequences.[10,11,20,21] r is an average residue and i and j are specific residue types in contact. The total contact energy $E^c(s, i)$ is represented as

$$E^c(s, i) = \frac{1}{2}\sum_p \sum_j n^c_{i_p j}(e_{i_p j} - e_{rr}) \tag{8}$$

$$\langle E^c(s, i)\rangle_{\beta=0,\ \text{native-like conf.}} \simeq \frac{1}{2}\sum_p \langle n^c_{i_p j}\rangle_{\text{all natives}}(e_{i_p j} - e_{rr}) \tag{9}$$

where $n^c_{i_p j}$ is the number of residues of type j in contact with residue of type i at position p. The unweighted averages of $n^c_{i_p j}$ over native-like conformations are approximated as the averages over all positions of residue type i_p in the native structures of all proteins.

It must be noted here that the approximations of Equation 7 and Equation 9 can be applied to sequences that can fold into a stable structure but that they cannot be used for a function, Equation 5, for sequence space optimization of proteins under the unrestricted condition of amino acid composition, because the unweighted averages of $\delta e^s(s_{p-1}, i_p, s_p, s_{p+1})$ and $n^c_{i_p j}$ over native-like compact structures for most sequences in sequence space may depart significantly from those for known protein structures. It was shown for lattice proteins that optimizing the Z score instead of energy can yield more stable and more foldable sequences.[14] Here ensembles of sequences whose amino acid compositions are taken to be those of native proteins are used.

These secondary structure potentials for the 20 types of amino acids and contact energies between them are re-estimated here from the new set of 2129 protein species representatives with the sampling method[10] and with the parameters evaluated in Miyazawa and Jernigan[11] to statistically estimate contact energies; refer to the section "Datasets of protein structures used" for the protein selection. Sampling weights for proteins are calculated according to the method reported in Miyazawa and Jernigan,[10] but identities <0.2 are regarded as 0 for alignments between proteins; as a result, the effective number of proteins used is reduced to 1658.

Statistical Ensemble of Sequences

Let us consider a statistical ensemble of sequences having probabilities $P(i|s)$, which are the conditional probabilities of sequence i for a given structure s, represented according to Bayes' rule.

$$P(i|s) = P(s|i)P(i) \Big/ \sum_i P(s|i)P(i) \tag{10}$$

$$P(i) = \text{constant} \tag{11}$$

where the sum over i means the sum over all sequences with fixed length for a given structure; here, we consider only sequences having the same amino acid composition as the native sequence. The conditional probability $P(s|i)$ of a specific conformation s for a given sequence i is defined by Equation 2 and is calculated from Equation 4 and Equation 5. $P(i)$ is the *a priori* probability of a sequence i and is taken to be the same for all sequences. Then, the probability of sequence i for a given structure s can be calculated as

$$P(i|s) = \frac{1}{\mathcal{Z}} \exp(-\beta \mathcal{E}(s, i)) \tag{12}$$

$$\mathcal{Z} \equiv \sum_i \exp(-\beta \mathcal{E}(s, i)) \tag{13}$$

The second term in Equation 4 is ignored here because it depends only on sequence length and does not depend on amino acid composition and is therefore constant for the same structure. Although the second term, the average conformational energy of native-like structures, in Equation 5 is also constant for a given structure, it is included to permit comparisons of these effective energies between different proteins; here it should be noted that terms are included if they depend on amino acid composition, even though they are constant for a given structure; thus, backbone-backbone interactions and the collapse energy e_{rr} are removed in the estimation of secondary structure energies of Equation 6 and for contact energies of Equation 8, respectively. $\mathcal{Z}$ is the partition function for the present statistical ensemble of sequences.

Here we examine whether or not short- and long-range interactions are frustrated in sequence space rather than in conformational space by calculating various types of statistical averages of short- and long-range interaction energies, over sequences using probability $P(i|s)$ of Equation 12, at equilibrium in sequence space. In the following, we use the notation of the statistical averages $\langle X\rangle_Y$ below:

$$\langle X\rangle_Y \equiv \frac{1}{\mathcal{Z}(Y)} \sum_i X(s, i)\exp(-\beta Y(s, i)) \qquad (14)$$

$$\mathcal{Z}(Y) \equiv \sum_i \exp(-\beta Y(s, i))$$

X and Y can be short-range, long-range interaction energy, or the sum of both energies. Here it should be noted that in Equation 14 energies are averaged in sequence space with a statistical weight (Boltzmann factor) and so only sequences which are similar to the native sequences and are compatible with the native structures will contribute significantly to the statistical averages of the energies.

For example, one of the statistical averages calculated will be

$$\langle \mathcal{E}^c\rangle_{\mathcal{E}^s+\mathcal{E}^c} \equiv \frac{1}{\mathcal{Z}} \sum_i \mathcal{E}^c(s, i)\exp(-\beta(\mathcal{E}^s(s, i) + \mathcal{E}^c(s, i))) \qquad (15)$$

where the superscripts, s and c, refer to secondary structure energy and contact energy, respectively. Also, the variance of energies such as $\langle(\Delta\mathcal{E}^c)^2\rangle_{\mathcal{E}^s+\mathcal{E}^c}$ is calculated, where

$$\Delta\mathcal{E} \equiv \mathcal{E} - \langle\mathcal{E}\rangle \qquad (16)$$

These statistical averages in sequence space are based on sets of equilibrium ensembles of sequences generated with Monte Carlo simulations for 797 representative proteins taken from the list of protein families in the SCOP database.[22]

Monte Carlo Simulations to Generate the Statistical Ensemble of Sequences

The statistical ensemble of sequences for each protein is generated with the Metropolis method[8] by exchanging pairs of amino acids in each protein, with 100,000 residue exchange trials per residue. Energies are evaluated in the multimeric state of a whole protein structure for each protein domain. The temperature $1/\beta$ is always taken to be one for the present energy scale, so that the sum of the equilibrium distributions over all proteins for contacting residues and tripeptide conformations are close to those observed in their native structures; this temperature was called a conformational temperature by Finkelstein et al.[13] Similar methods were previously used for optimizing protein sequences for a given structure.[23,24]

Datasets of Protein Structures Used

Proteins each of which represent a different protein fold were collected. Release 1.53 of the SCOP database[22] was used for the classification of protein folds. Representatives of families or species are the first entries in the protein lists for each family or each species in the SCOP; if these first proteins in the lists are not appropriate (see below) to use for the present purpose, then the second ones are chosen. These families and species are all those belonging to the protein classes 1 to 5; that is, classes of all α, all β, α/β, α + β, and multidomain proteins. Classes of membrane and cell surface proteins, small proteins, peptides, and designed proteins are not used. Proteins whose structures[25] were determined by NMR or with resolution worse than 2.5 Å are removed to assure that the quality of proteins used is high. Also, proteins whose coordinate sets either consist of only C^α atoms or include many unknown residues or lack many atoms or residues are removed. Proteins shorter than 50 residues are also removed. As a result, the sets of family representatives and species representatives include 797 proteins and 2129 proteins, respectively. The set of family representatives is used for analyses, where each protein is to have a completely different protein fold.

RESULTS AND DISCUSSION

First, the statistical averages of contact energies of proteins are calculated for an equilibrium ensemble of sequences with two different systems: (1) interactions consist of contact interactions only and (2) interactions include both contact and secondary structure potentials. These values are compared in Figure 1(A). If there were no correlation between those two types of interactions, the addition of secondary structure interactions would cause no systematic deviations in the statistical averages of contact energies of proteins. If secondary structure interactions were to work against pairwise contact interactions, the statistical averages of contact energies would increase due to the addition of secondary structure interactions. However, the observation is that 769 of the 797 proteins fall below the line having unit slope, showing that mean contact energies almost always decrease when secondary structure interactions are added. Since those proteins are the representatives from each protein family in the protein fold database of SCOP-1.53,[22] the decrease in the mean contact energies by addition of secondary structure potentials is a general characteristic of protein native structures. This indicates that in protein native structures the contact interactions are consistent with the secondary structure interactions, at least at the level of mean energies.

The effects of adding secondary structure interactions on the variance of contact energies are presented in Figure 1(B). For 791 of the 797 proteins, the variances of contact energies also decrease when secondary structure interactions are added. The inclusion of the next neighbor interactions between a tripeptide and a side chain in the secondary structure potential decreases even further those means and variances of contact energies (results not shown). Thus, in almost all protein structures, the secondary structure interactions are consistent with the contact

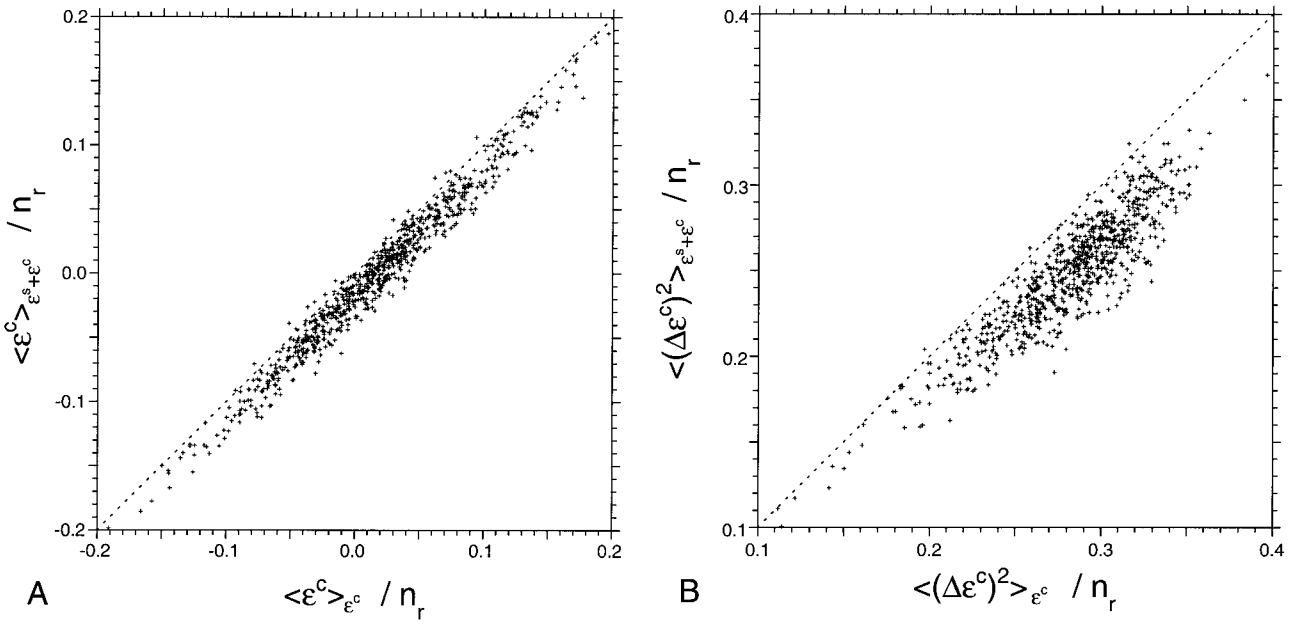

Fig. 1. Comparisons of mean contact energies per residue (A) and variances per residue (B), averaged in two ways, at statistical equilibrium for sequences generated with contact interactions only and at equilibrium with both the secondary structure energies and contact energies included. Shown are 797 proteins of family representatives in the SCOP-1.53 database of which 769 averages and 791 variances of contact energies are reduced when averaged with both classes of energies. The dotted line shows a line with equal values for both axes. See Equation 14 for the definition of the notation of the statistical averages indicated on each axis.

interactions, decreasing not only the means but the variances of contact energies at equilibrium.

The converse relationship, that is, the effects of contact interactions on secondary structure energies are shown in Figure 2; the means and variances of secondary structure energies at statistical equilibrium over sequences generated with secondary structure interactions only are plotted against those with both secondary structure and contact

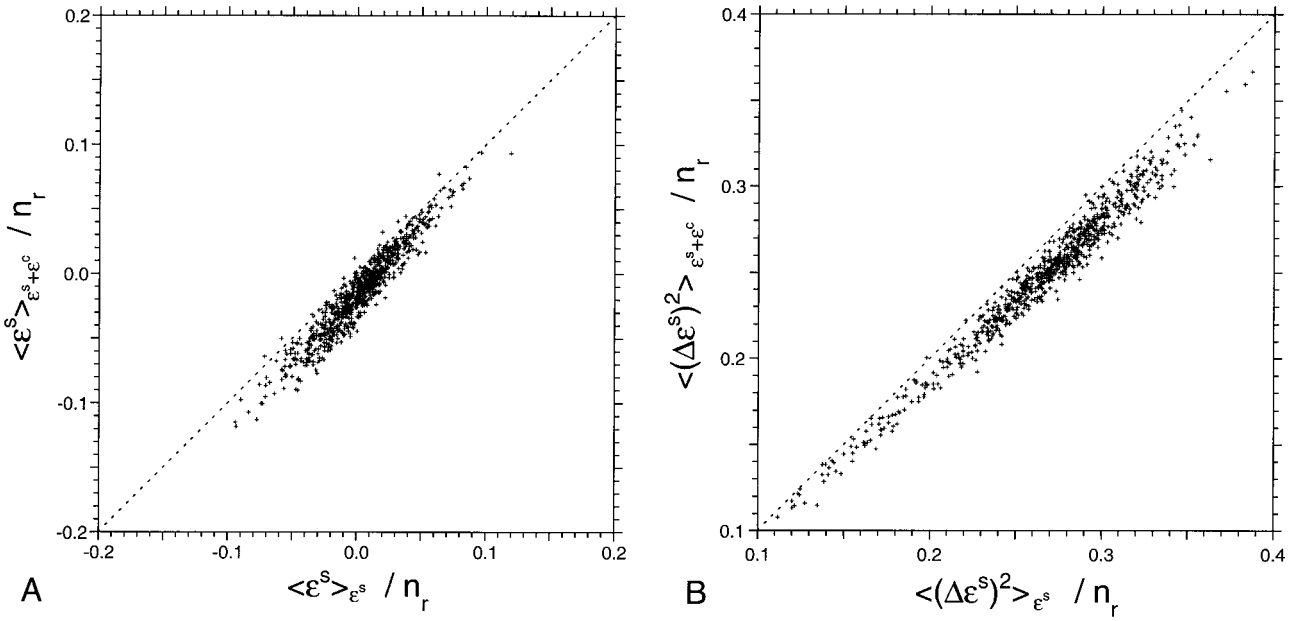

Fig. 2. Comparisons of mean secondary structure energies per residue (A) and variances per residue (B), averaged in two ways, at statistical equilibrium of sequences with secondary structure interactions only and at equilibrium with both secondary structure energies and contact energies included. Shown are 797 proteins of family representatives in the SCOP-1.53 database. Similarly to the results in Figure 1, 771 averages and 792 variances of secondary structure energies are reduced when both categories of interactions are considered. The dotted line shows a line with equal values for both axes. See Equation 14 for the definition of the notation of the statistical averages indicated on each axis.

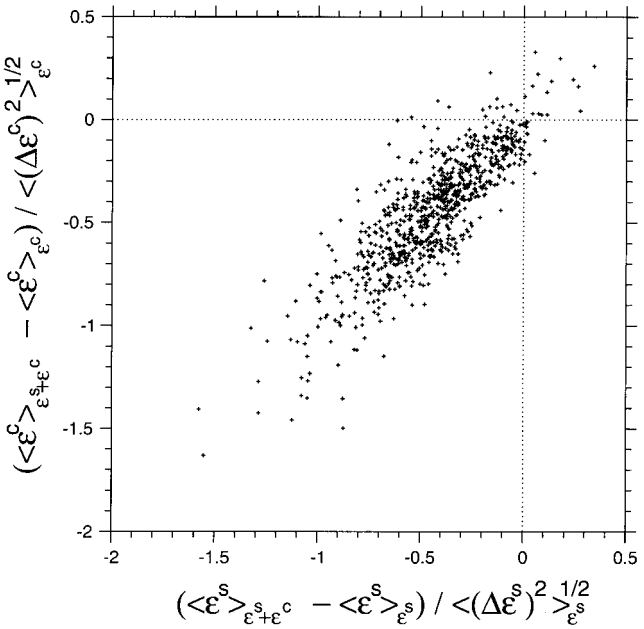

Fig. 3. The changes of mean interaction energies caused by averaging using the total energy are shown in standard deviation units of the interaction energies for each of the short-range secondary structure and the long-range contact interactions. Shown are 797 proteins of family representatives in the SCOP-1.53 database. See Equation 14 for the definition of the notation of the statistical averages indicated on each axis. The dotted lines show lines with zero values for each axis.

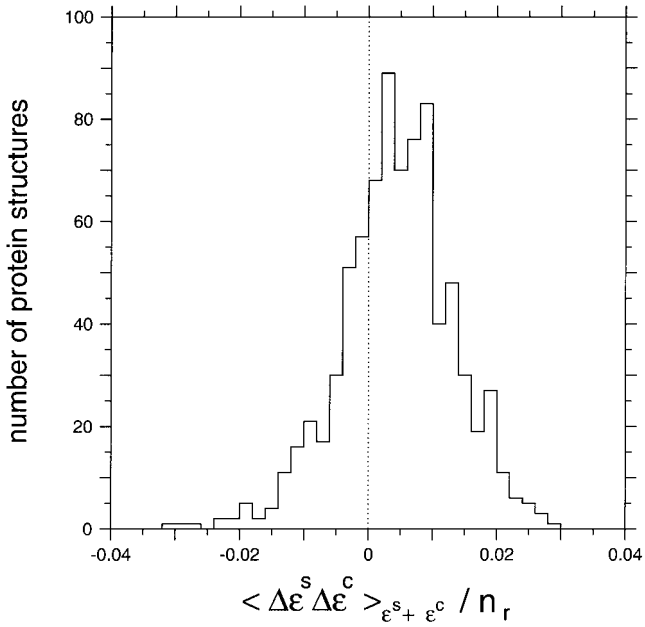

Fig. 4. Frequency of proteins, for intervals of 0.002 in the covariances per residue between secondary structure energies and contact energies averaged with statistical ensembles of sequences generated with both secondary structure energies and contact energies. Used here are 797 proteins of family representatives in the SCOP-1.53 database. See Equation 14 for the notation of the statistical average indicated on the abscissa. A dotted line marks zero covariance.

potentials. The decreases to the means and variances of secondary structure energies by the addition of contact interactions are likewise clearly observed; the means decrease for 771 of the 797 proteins and the variances decrease for 792 of the 797 proteins.

The decreases in the means of one class of interaction energies upon adding the other class of interactions are not small in comparison with the variances of the interaction energies. As shown in Figure 3, these changes caused by averaging using the total energy fall mostly in the range of 0 to -1, with a weak dependence ($\propto -n_r^{0.5}$) on protein length, of the standard deviations of the interaction energies for both classes of the energies. As a result, these observations indicate that in protein native structures the long-range contact interactions and the short-range secondary structure interactions are consistent with each other at the level of mean energies and variances.

To further understand common features between the contact and secondary structure interactions in protein structures, covariances between contact energies and secondary structure energies have been calculated for the 797 proteins. The dependence of the statistical average of energy on one class of interaction is represented as

$$\int_0^1 \frac{\partial \langle \mathscr{E}^c \rangle_{x\mathscr{E}^s + \mathscr{E}^c}}{\partial x}\, dx = -\beta \int_0^1 \langle \Delta\mathscr{E}^s \Delta\mathscr{E}^c \rangle_{x\mathscr{E}^s + \mathscr{E}^c}\, dx \qquad (17)$$

$$\int_0^1 \frac{\partial \langle \mathscr{E}^s \rangle_{\mathscr{E}^s + y\mathscr{E}^c}}{\partial y}\, dy = -\beta \int_0^1 \langle \Delta\mathscr{E}^s \Delta\mathscr{E}^c \rangle_{\mathscr{E}^s + y\mathscr{E}^c}\, dy \qquad (18)$$

with parameters x and y; see Equation 14 for the definition of these statistical averages. These covariances, the integrand with $x = 1$ or $y = 1$ on the right hand side, are calculated first with the statistical equilibrium generated with both interactions. Figure 4 shows the frequencies of proteins whose covariances per residue fall into the various intervals of size 0.002. The covariances per residue scatter about zero with a small shift toward positive values, indicating that there is virtually no correlation between the two types of energies in the statistical ensemble of sequences selected using both interactions; the correlation coefficients between these interactions are mostly between ± 0.1. Then, these covariances are calculated for systems with secondary structure interactions only ($y = 0$) and with contact interactions only ($x = 0$), shown in Figure 5. Almost all proteins show positive covariances between these two interactions, consistent with the results in Figures 1 and 2. The correlations between both interactions are significant; however, it should be noted here that the values of these covariances are less than one fifth of the values of the variances, $\langle (\Delta\mathscr{E}^c)^2 \rangle_{\mathscr{E}^c}$ and $\langle (\Delta\mathscr{E}^s)^2 \rangle_{\mathscr{E}^s}$ shown in Figures 1(B) and 2(B); the correlation coefficients between these interactions are almost all positive but below 0.2.

What landscape of an energy surface in sequence space would have such features as seen in Figures 3 and 4? One

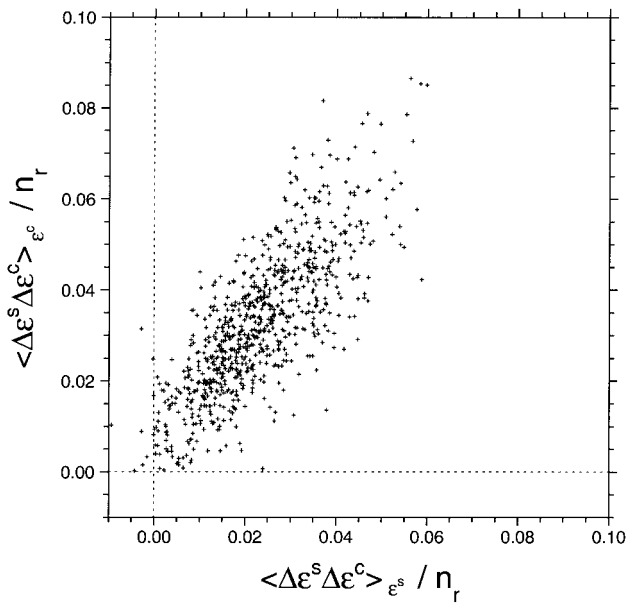

Fig. 5. Two kinds of statistical averages for covariances per residue between secondary structure energies and contact energies are compared, in one of which statistical ensembles of sequences are generated with contact energies only and in the other, secondary structure energies alone are used. The dotted lines show lines with zero values for each axis. See Equation 14 for the meaning of the notation of the statistical averages indicated on each axis.

of the characteristics of an energy surface that could lead to these features would be for each interaction to tend to reduce the sequence space available to the other. Sequence spaces that are inaccessible for one of the interactions would also exist in relatively high energy regions for the other class of interaction, and as a result, the statistical averages of interaction energies would decrease when the former class of interaction was added. This is reasonable, because the contact interactions, for example, will tend to restrict polar residues to the surface and nonpolar residues to the interior of protein structures and thus substantially reduce the sequence space accessible to polar and nonpolar residues. If backbone conformations of interior residues are suitable for these nonpolar residues and those for exterior residues are also favorable for polar residues, then the addition of contact interactions will be favorable for secondary structure interactions and will reduce the mean energies of secondary structure interactions as shown in Figure 2. Conversely, the addition of secondary structure interactions will tend to yield favorable contact interactions and will decrease the mean energies of contact interactions as shown in Figure 1. However, both interaction classes act almost independently near the minimum in the total energy surface, as can be seen in Figure 4. Even far from the minimum, the correlation coefficients between both interactions are below 0.2; see Figures 1(B), 2(B), and 4. These results are consistent with an approximation of mean force for the secondary structure potentials and contact potentials, in which all other classes of interactions are included only as a mean field[9–12]; in Miyazawa and Jernigan,[11] the effects of secondary structure interactions were taken into account to estimate contact energies from those values predicted by the Bethe approximation.

These calculations have also been repeated using a secondary structure intra-residue potential based on data collected for 10° intervals of (ϕ, ψ) angles, and nearly identical results were observed.

Can Native Protein Sequences Be Regarded as Samples at Equilibrium in Sequence Space?

We have examined the consistency between the contact and secondary structure energies of sequences at statistical equilibrium in sequence space. These analyses would be most meaningful if the native sequences of the proteins can be regarded as samples at equilibrium in sequence space. We reported previously[11] that the total contact frequencies between the 20 kinds of amino acids observed in many protein native structures can be regarded with small relative errors ($<10\%$) as contact frequencies at statistical equilibrium in sequence space. Here, it is shown that contact energies and secondary structure energies of most native proteins lie mostly within the statistical fluctuations around equilibrium in sequence space.

In Figure 6, the frequency distributions of the contact energies, secondary structure energies and their sums of native protein sequences are shown. Their energies are given in standard deviation (SD) units away from the mean of each interaction energy in the statistical ensemble for each protein. The statistical ensemble and the statistical averages of the interaction energies are calculated in each case by using both classes of interaction energies. Although the frequency distributions for the contact energies and for the secondary structure energies are slightly shifted from the origin in opposite directions, the frequency distribution for the total energies is similar to a Gaussian distribution; there is no clear reason that they must obey a Gaussian distribution. For reference, a Gaussian distribution is shown as a dotted line in Figure 6(B). The native sequences and structures of proteins are both close to the statistical equilibrium in sequence space, at least to the extent that the energies for most native proteins fall mostly within ± 2 SD of the equilibrium mean for each protein.

In Figure 7, the contact energy and secondary structure energy, measured in SD units from the means of each native protein sequence are compared. There is clearly no correlation between the deviations of secondary structure and contact energies for each native protein from their statistical averages; the correlation coefficient is 0.11. This is expected from the results described in the previous section that show both interactions to act nearly independently near the minimum in the total energy surface, insofar as native proteins can be regarded as samples at equilibrium in sequence space.

In addition, these two facts, shown in Figures 6 and 7, indicate that the conformational temperature $1/\beta$ is the same for all proteins; otherwise, contact energies and

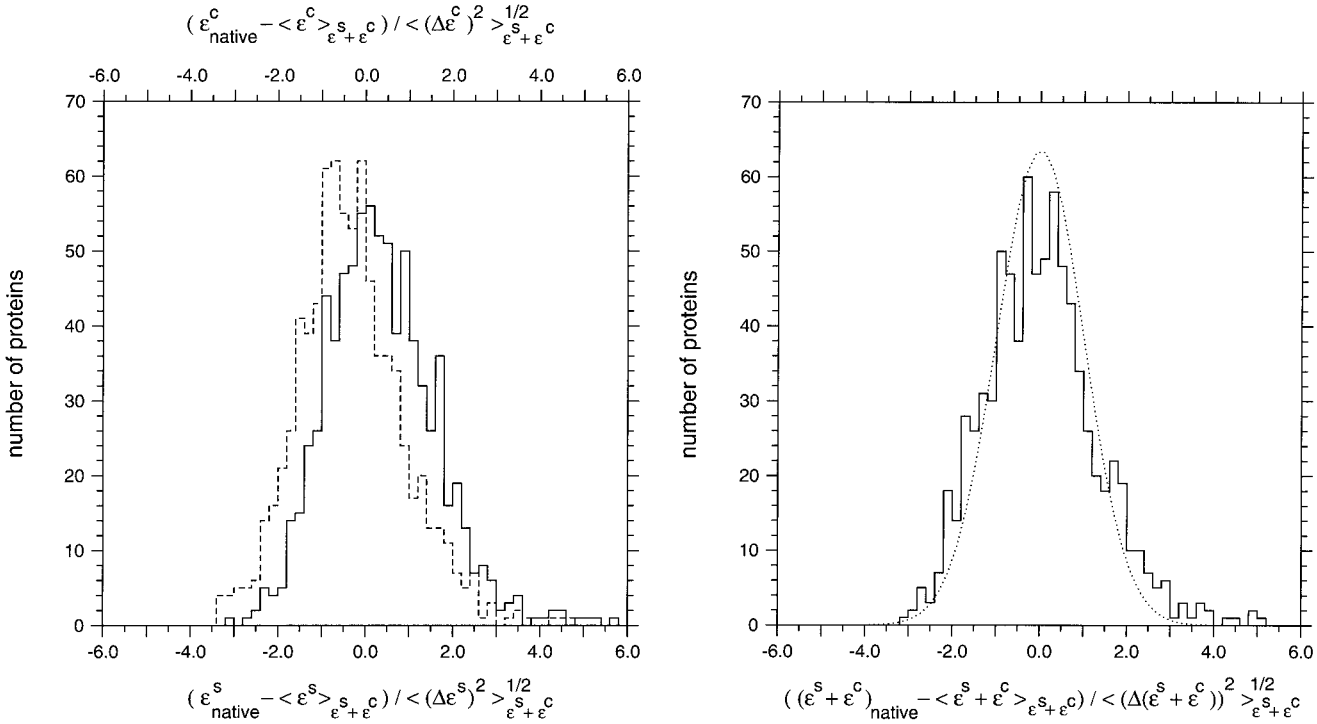

Fig. 6. Frequency of proteins is shown for each interval, 0.2, for native secondary structure energies as a solid line in (A) and for native contact energies as a broken line in (A), and with their sums indicated by a solid line in (B), which are measured in standard deviation units away from their means calculated using a statistical equilibrium of sequences generated with both interaction energies for each protein. Used here are 797 proteins of family representatives in the SCOP-1.53 database. For reference, a Gaussian distribution is shown as a dotted curve in (B). See Equation 14 for the definition of the notations of the statistical averages indicated on each axis.

secondary structure energies of the native proteins would exhibit greater scatter and show a positive correlation between them. This result is consistent with the study of Finkelstein et al.,[13] who showed that the occurrence of various structural elements in stable folds of random copolymers is exponentially dependent on the intrinsic energy of the element and that the same "conformational temperature," equal to the freezing temperature, holds for any structural element, independently of whether it is a small detail or a large-scale motif of overall chain folding.

CONCLUSIONS

For coarse-grained energy potentials, i.e., statistical potentials at the residue level, it has been shown, for most protein structures, that short-range secondary structure interactions and long-range contact interactions are consistent with each other for a statistical equilibrium with residue exchanges in protein sequences. The statistical averages of one class of interaction energies almost always decrease when the other class of interactions are included. Also, the decrease in the mean of one class of interaction energies by including the other class of interactions is about 0.4, on average, of the standard deviation of the energies. The decreases in the variance of secondary structure energies by the addition of inter-residue contact interactions indicate that these long-range interactions tend to reduce the available range of peptide dihedral

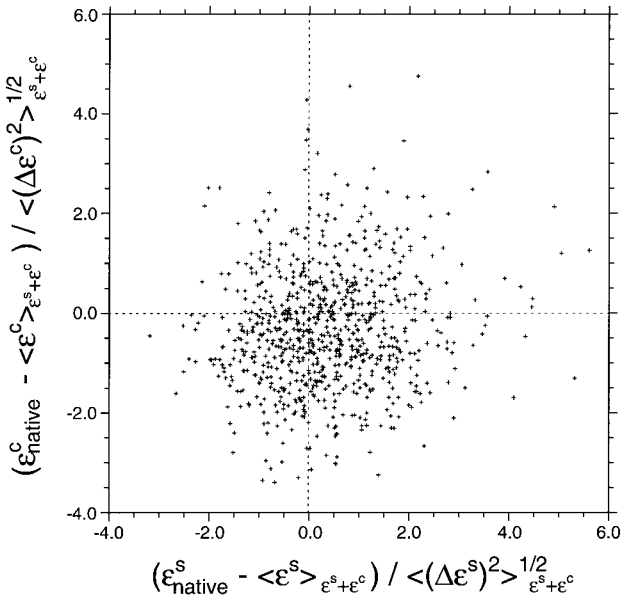

Fig. 7. Contact energies of native protein sequences are plotted against their secondary structure energies. Both classes of energies are measured in standard deviation units away from their means in the equilibrium ensemble of sequences generated using both interaction energies for each protein. Plotted here are 797 protein family representatives from the SCOP-1.53 database. See Equation 14 for the definition of the notations of the statistical averages indicated on each axis.

angle space. Also decreases in variance of contact energies by the addition of secondary structure interactions indicate that the short-range interactions can likewise restrict protein native structures. However, it should be noted here that both interactions act almost independently near the minimum in the total energy surface, and even far from the minimum the correlation coefficients between them are below 0.2, supporting the mean force approximation for the secondary structure potentials and contact potentials, at least as a first approximation; although the decrease mean energy can be as large as the standard deviation of the energies. This fact indicates that peptide conformations for interior residues are more or less adjusted to be suitable for nonpolar residues, and inversely those for residues on protein surfaces are adjusted for polar residues. Thus, this consistency between short- and long-range interactions has been shown in sequence space and also implies that the energy landscape of short- and long-range interactions in protein native sequences is minimally frustrated near protein native structures in conformational space. Such consistency and minimal frustration in interactions is not found for random heteropolymers, and this is an important distinctive characteristic of proteins, which causes them to fold into single stable structures within reasonable times. Proteins must have achieved these unique characteristics of smoothing the energy landscape on a coarse-grained conformational scale over the course of molecular evolution.

REFERENCES

1. Ptitsyn OB, Finkelstein AV. Relation of the secondary structure of globular proteins with their primary structure. Biofizika (Moscow) 1970;15:757–767.
2. Go N. Theoretical studies of protein folding. Annu Rev Biophys Bioeng 1983;12:183–210.
3. Eisenmenger F, Argos P, Abagyan R. A method to configure protein side-chains from the main-chain trace in homology modeling. J Mol Biol 1993;231:849–860.
4. Tanimura R, Kidera A, Nakamura H. Determinants of protein side-chain packing. Prot Sci 1994;3:2358–2365.
5. Bryngelson JD, Wolynes PG. Spin glasses and the statistical mechanics of protein folding. Proc Natl Acad Sci USA 1987;84: 7524–7528.
6. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, pathways and the energy landscape of protein folding: a synthesis. Proteins 1995;21:167–195.
7. Panchenko AR, Luthey-Schulten Z, Cole R, Wolynes PG. The foldon universe: a survey of structural similarity and self-recognition of independently folding units. J Mol Biol 1997;272:95–105.
8. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. J Chem Phys 1953;21:1087–1092.
9. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
10. Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term for simulation and threading. J Mol Biol 1996;256: 623–644.
11. Miyazawa S, Jernigan RL. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. Proteins 1999;34:49–68.
12. Miyazawa S, Jernigan RL. Evaluation of short-range interactions as secondary structure energies for protein fold and sequence recognition. Proteins 1999;36:347–356.
13. Finkelstein AV, Badretdinov AY, Gutin AM. Why do protein architectures have Boltzmann-like statistics? Proteins 1995;23: 142–150.
14. Mirny L, Abkevich V, Shakhnovich EI. Universality and diversity of the protein folding scenarios: a comprehensive analysis with the aid of lattice model. Folding Design 1996;1:103–116.
15. Seno F, Vendruscolo M, Maritan A, Banavar JR. Optimal protein design procedure. Phys Rev Lett 1996;77:1901–1904.
16. Deutsch JM, Kurosky T. New algorithm for protein design. Phys Rev Lett 1996;76:323–326.
17. Morrissey M, Shakhnovich EI. Design of proteins with selected thermal properties. Folding Design 1996;1:391–406.
18. Pande VS, Grosberg AY, Tanaka T. Statistical mechanics of simple models of protein folding and design. Biophys J 1997;73: 3192–3210.
19. Shakhnovich EI. Protein design: a perspective from simple tractable models. Folding Design 1998;3:R45–R58.
20. Miyazawa S, Jernigan RL. An empirical energy potential with a reference state for protein fold and sequence recognition. Proteins 1999;36:357–369.
21. Miyazawa S, Jernigan RL. Identifying sequence-structure pairs undetected by sequence alignments. Prot Eng 2000;13:459–475.
22. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
23. Shakhnovich EI, Gutin AM. A new approach to the design of stable proteins. Prot Eng 1993;6:793–800.
24. Shakhnovich EI, Gutin AM. Engineering of stable and fast-folding sequences of model proteins. Proc Natl Acad Sci USA 1993;90: 7195–7199.
25. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.