PROTEINS: Structure, Function, and Genetics 36:357–369 (1999)

An Empirical Energy Potential With a Reference State for Protein Fold and Sequence Recognition

Sanzo Miyazawa[1,2]* **and Robert L. Jernigan**[2]
[1]*Faculty of Technology, Gunma University, Kiryu, Gunma, Japan*
[2]*Laboratory of Experimental and Computational Biology, DBS, National Cancer Institute, National Institutes of Health, Bethesda, Maryland*

ABSTRACT **We consider modifications of an empirical energy potential for fold and sequence recognition to represent approximately the stabilities of proteins in various environments. A potential used here includes a secondary structure potential representing short-range interactions for secondary structures of proteins, and a tertiary structure potential consisting of a long-range, pairwise contact potential and a repulsive packing potential. This potential is devised to evaluate together the total conformational energy of a protein at the coarse grained residue level. It was previously estimated from the observed frequencies of secondary structures, from contact frequencies between residues, and from the distributions of the number of residues in contact in known protein structures by regarding those distributions as the equilibrium distributions with the Boltzmann factor of these interaction energies. The stability of native structures is assumed as a primary requirement for proteins to fold into their native structures. A collapse energy is subtracted from the contact energies to remove the protein size dependence and to represent protein stabilities for monomeric and multimeric states. The free energy of the whole ensemble of protein conformations that is subtracted from the conformational energy to represent protein stability is approximated as the average energy expected for a typical native structure with the same amino acid composition. This term may be constant in fold recognition but essentially varies in sequence recognition. A simple test of threading sequences into structures without gaps is employed to demonstrate the importance of the present modifications that permit the same potential to be utilized for both fold and sequence recognition. Proteins 1999;36:357–369.**

Key words: optimal protein sequence; protein inverse folding; protein sequence threading; scoring function; protein sequence design; sequence-structure compatibility; statistical potentials; 3D-1D

INTRODUCTION

It was reported that conventional intra-protein energy potentials at the atomic level do not necessarily give lower energies for the native structures than for non-native folds,[1] unless proper account is taken of solvent effects.[2] This result provided a landmark demonstration of the important role of hydrophobic interactions. However, properly accounting for all electrostatic interactions, hydrogen bonds and non-bonded interactions between a protein and water, and estimating precisely the entropic and energetic effects of solvent are extremely complex, making impossible a rigorous evaluation of the free energies of folds. Consequently simplified, less detailed models are needed to study protein folding, with simplifications needed in both the geometry and the potential functions. Fortunately, a simple model of solvation energy was shown to be able to distinguish misfolded proteins from native structures.[3] Since then, a number of other simple methods were developed to distinguish native protein structures from non-native folds and also to recognize folds compatible with sequences.[4–14] These subjects have been discussed extensively in the CASP meetings[15,16] that are held regularly on the critical assessment of methods of protein structure prediction.

Even if the positions of all backbone atoms in a protein are fixed, the energy surface for conventional potentials at the atomic level still has many local minima corresponding to different positions of side chain atoms. These local minima are separated by relatively high energy barriers, arising from van der Waals repulsions between atoms. In order to understand the dependency of the potentials on backbone geometry alone, the potential of mean force is calculated by effectively integrating a Boltzmann factor over all possible side chain atom positions and all solvent positions for fixed backbone atoms. This coarse graining aids in evaluating protein folds, but it is important that the process be done in a careful, self-consistent way.

Any energy potential designed to describe protein folds completely on a physical basis needs to approximate the potential of mean force for backbone conformations using a simple function of backbone geometry. One simple way is to represent the free energies of backbone folds as a sum of short-range interactions, long-range pairwise residue-residue interactions, and higher order contributions. In general, such a mean force potential would include interactions higher in order than two-body.[17]

*Correspondence to: Sanzo Miyazawa, Faculty of Technology, Gunma University, Kiryu, Gunma 376-8515, Japan. E-Mail: miyazawa@smlab.sci.gunma-u.ac.jp
Received 24 September 1998; Accepted 8 April 1999

Previously, we empirically evaluated a set of effective inter-residue contact energies for all twenty types of amino acids using the Bethe approximation based upon the numbers of residue-residue close contacts observed in protein crystal structures.[18] Recently, the validity and stability of these contact energies were demonstrated by using many more protein structures and demonstrating that they were virtually unchanged.[19] These empirically derived energy functions include solvent effects and provide an estimate of the long-range component of conformational energies without atomic details.[20] Atomic contact energies were recently estimated by a similar method.[21] Thomas and Dill[22] showed, however, that input contact potentials can be approximately extracted from an ensemble of lattice conformers with this procedure but that the extracted contact potentials depend on the proteins used. On the other hand, Mirny and Shakhnovich[23] showed that contact energies do reflect well the input contact potentials in 3-dimensional lattice conformers. Also, Miyazawa and Jernigan[24] showed that the correlation between input and predicted contact potentials is highly significant and larger than 0.9.

Pairwise potentials were also estimated by Sippl[25] from radial distributions of all pairs of 20 types of amino acids as potentials of mean force; they were estimated as a function of distance between residues to be the logarithm of the ratio of the radial distribution function of a given type of residue pair to the overall radial distribution function. Pairwise potentials were also estimated as potentials of mean force in similar ways by others.[10,11] Sippl's original expression of the pairwise potentials does not include hydrophobic interaction energies[24,26]; refer to Miyazawa and Jernigan[24] for a discussion of the differences between Sippl's potentials and our contact energies. Potentials that represent the relative hydrophobicities of residues have been devised and added to Sippl's type of pairwise potentials especially for fold recognition as: a function of residue accessibility,[7] a function of accessible surface area of a residue,[10] or a function of the number of residues within a given shell surrounding a given residue.[9]

Detecting compatibilities between sequences and structures is a highly practical and important problem, especially since recent evidence from accumulated data on protein structures and sequences strongly indicates that the number of families of protein folds appears to be limited, with suggested values ranging from[27] 1000 to about[28] 7900. This implies that prediction of a protein fold from its sequence could be based on searches over this limited set of folds rather than by thoroughly searching over the entire vast conformational space of a protein. Therefore, scoring functions that can discriminate the native structures among non-native folds and detect compatible protein folds ought to be extremely useful. One of our purposes here is to develop a unified set of empirical potentials to approximate actual conformational energies without all of the atomic details. Finally, we assess their effectiveness for fold recognition.

Long-range interactions among residues are certainly principal forces for protein sequences to recognize their native folds because they are responsible for proteins cooperatively folding into their unique native structures. However, short-range interactions ought not to be neglected even in fold-sequence recognition because they contribute significantly to the formation of secondary structures in proteins, which are also essential parts of protein structures. Secondary structure potentials have been used in some works[9,10,29] on fold recognition. The effects of short-range interactions on secondary structures have been evaluated[30] from the observed frequencies of secondary structures in known protein structures which are assumed to have an equilibrium distribution, following the Boltzmann factor of secondary structure energies. In this work, unlike previous works, interactions are decoupled into intrinsic potentials of residues, potentials of backbone-backbone interactions, and of side chain-backbone interactions. Interactions are also decoupled into one-body, two-body, and higher-order interactions between backbone and side chain and between peptide backbones. These decouplings are essential to correctly evaluate the total secondary structure energy of a protein structure without multiple counts of interactions. Each interaction potential was evaluated separately by taking account of the correlation in the amino acid order of protein sequences. Interactions among side chains were neglected because of the relatively limited number of protein structures. These short-range potentials[30] for secondary structure are devised to be used additively with the long-range contact energies and repulsive packing energies[19] for evaluating the total conformational energies of proteins at the residue level; they are devised in order to avoid multiple counts of any interaction. Here the enhancement of secondary structure potentials for fold recognition will be examined.

Different folds of the same sequence can be compared directly with an energy scale. However, if the compatibilities of different sequences for a given fold are to be discussed, or if deletions and insertions are to be allowed in sequence-structure alignments, then usually an energy potential cannot properly reflect the stabilities of proteins and so cannot be used as a scoring function for such comparisons, unless the zero energy is properly chosen by means of an appropriate reference state. Even when different folds of the same sequence are compared, there can be a problem if one fold in a multimeric state is to be compared with the energy of another fold in a monomeric state. For multimeric proteins, the total stability, rather than that of the isolated monomer, must be considered.

Below, we discuss how to modify these energy potentials to represent approximately the stabilities of proteins in various environments, for multimers, as well as for monomers. Importantly, we also describe how to define a single reference state for these potentials that is appropriate for sequence recognition; protein native sequences are assumed to be well designed to fold into their native structures, and then the stability of native structures is assumed as a primary requirement for proteins to fold. The stability of a protein conformation can be measured by the free energy of the whole ensemble of protein conformations

subtracted from its conformational energy. Thus, the free energy of all protein conformations serves as a reference energy for an energy potential to measure protein stability. It is approximated here as the average energy expected for a typical native structure with the same amino acid composition. This term is constant in fold recognition unless deletions and additions are allowed, but essentially varies in sequence recognition.

A simple test of threading sequences into structures without gaps is used to demonstrate the usefulness of these potentials for both discriminating a native fold from non-native folds and also a native sequence from non-native sequences. Tests for both types of recognition are needed in order to justify the modifications of these potentials.

MATERIALS AND METHODS

Conformational Energy

The total conformational energy of a protein is represented here as a sum over contributions from residues along the sequence as

$$E^{conf} \equiv \sum_p E_p^{conf} \quad (1)$$

Each residue's contribution is further divided into two terms, for secondary structure and for tertiary structure.

$$E_p^{conf} \equiv E_p^{sec} + E_p^{tert} \quad (2)$$

where p indexes residue position.

The secondary structure energies used here have been estimated[30] on the basis of short-range interactions, ignoring the effects of long-range interactions. However, this does not mean that either structure is arrived at independently of the other. The classification into short-range and long-range terms here is based on the distance of separation between residues along a protein sequence and not on the physical range of interactions; short-range interactions are those between residues close along the protein sequence, and long-range interactions are those between sequentially distant residues.

The tertiary structure energies have previously been estimated as a sum of pairwise residue-residue contact energies and repulsive residue packing energies for volume exclusion, together termed long-range interaction energies.[19]

$$E_p^{tert} = E_p^c + E_p^r \quad (3)$$

The contact energy E_p^c and the repulsive packing energy E_p^r of a residue at position p are defined by Eqs. 18–19 and Eq. 40 in one of our previous papers.[19]

Secondary Structure Potential

The contribution of the pth residue to secondary structures is approximated to originate only in the short-range interactions.

$$E_p^{sec} \simeq e^s(\ldots; i_{p-1}, s_{p-1}; i_p, s_p; i_{p+1}, s_{p+1}; \ldots) \quad (4)$$

$e^s(\ldots; i_{p-1}, s_{p-1}; i_p, s_p; i_{p+1}, s_{p+1}; \ldots)$ is the short-range interaction energy within the secondary structure, $(\ldots; i_{p-1}, s_{p-1}; i_p, s_p; i_{p+1}, s_{p+1}; \ldots)$, where i_p is the residue type at p, and s_p means the secondary structure of that residue. Thus, s_p designates a backbone conformation and i_p the residue type at position p. The ellipses indicate those yet unspecified, but nonetheless of limited range.

The effects of short-range interactions on secondary structures have been estimated[30] by a potential of mean force from the observed frequencies of secondary structures in known protein structures, which is assumed to be an equilibrium distribution with the Boltzmann factor of their secondary structure energies. The correlations between long- and short-range interactions are neglected, and the effects of long-range interactions are taken into account only as a mean field. Because of the limited number of available protein structures, the secondary structure potential, e^s, is approximated as a sum of additive contributions from neighboring residues along a sequence, with neglect of side chain-side chain interactions. Non-additive contributions are simply neglected. In addition, the effects here from neighboring residues are limited to a dependence on their amino acid type but not on their secondary structures. The conformational specification is limited to a tripeptide.

This previous estimate[30] of secondary structure potentials is used here. Thus, the secondary structure potential, e^s, is approximated as a sum of the following contributions

$$e^s(\ldots; i_{-1}, s_{-1}; i_0, s_0; i_1, s_1; \ldots) \simeq e^s(s_{-1}, s_0, s_1) + \sum_{-3\le p\le 3} \delta e^s(s_{p-1}, s_p, s_{p+1}, i_0) \quad (5)$$

or

$$\simeq e^s(s_{-1}, s_0, s_1) + \sum_{-3\le p\le 3} \delta e^s(s_{-1}, s_0, s_1, i_p) \quad (6)$$

The residue under consideration is indexed as zero, and negative and positive numbers represent relative residue positions towards the N-terminal and the C-terminal sides. The first terms in Eqs. 5 and 6 represent the backbone-backbone interactions and the second terms correspond to side chain-backbone interactions either within a residue or among residues. Altogether side chain-backbone interactions within five consecutive backbone units on each side of a side chain are included in the short-range interactions. Here it should be noted that two-body and higher order interactions between side chains and backbones of triplets are counted only once in the estimation of each term in Eqs. 5–6 to add to the total short-range interaction. The first term $e^s(s_{-1}, s_0, s_1)$ is also defined[30] to include only half of the two-body interactions between nearest neighbors in order to avoid multiple

counts of nearest neighbor interactions in the estimation of the total secondary structure energy of Eq. 1.

Any zero energy state for each energy term in Eqs. 5–6 could be taken for convenience as energy functions, but it is useful to set the statistical averages of these energies to zero for use as scoring functions for compatibilities between sequences and structures as will be described later.

Contact Energies and Repulsive Packing Energies

The contact energies (e_{ij}) for all pairs of the twenty types of residues which are applied to residue-residue close contacts, and the repulsive packing energies for the twenty types of residues which are a function of the number of residues in contact, previously estimated by us,[19] are employed here. The contact energy E_p^c and repulsive packing energy E_p^r of residues at each position in structures as required in Eq. 3 are calculated according to their Eqs. 18–19 and Eqs. 40–43. However, the hard core repulsion term is not included here; i.e., e^{hc} is set to zero in their Eq. 41, since there should not be overly dense regions in properly refined structures.

Alignment Energy for Scoring of Sequence-Structure Compatibility

In the following, a reference state for zero energy is defined and each energy potential is modified to represent the stabilities of protein structures for measuring sequence-structure compatibilities. The stability of native structure is assumed as a primary requirement for proteins to fold into their native structures.

Reference State

The stability of a specific conformation for a protein sequence is determined relative to the whole ensemble of protein conformations, i.e., the partition function

$$-\log(\text{probability of a specific conformation s in a sequence i}) = \beta E^{conf}(s, i) + \log\left(\sum_s \exp(-\beta E^{conf}(s, i))\right) \quad (7)$$

where β is equal to 1/kT, the variable i means a specific sequence, s means a conformational state, $E^{conf}(s, i)$ is the conformational energy of state s of sequence i, and the sum is taken over all possible conformations. Therefore, the free energy of the whole ensemble can be regarded as a zero energy state, i.e., a reference state for an energy potential to represent protein stability. The free energy of the protein ensemble varies unless the protein sequence is the same. Thus, in order to discuss the compatibilities of different protein sequences with a given fold, it must be taken into account, as well as the conformational energy. Even in the case of searching for folds compatible with a given sequence, if deletions in the sequence are allowed in sequence-structure alignments, then the change of the whole ensemble of protein conformations must be taken into account.

How can we estimate the second term of Eq. 7 which serves as a reference energy for an energy potential to measure protein stability? Analyses using the Random Energy Model (REM) approximation suggest that the contribution to the partition function from non-native-like conformations depends primarily on amino acid composition rather than on sequence at high enough temperature $T > T_c$, where T_c is the temperature of the "freezing" transition in a random heteropolymer having the same amino acid composition.[31,32] Although this result from the mean-field heteropolymer theory must be examined, it indicates that the change in the conformational partition function may be neglected unless the amino acid composition changes; otherwise, it must be taken into account. In sequence space optimization for simple lattice proteins, estimating Eq. 7 has been attempted. The Z score was used[33] instead of native energy. The partition function was estimated by dual Monte-Carlo simulations,[34] by taking account of the first cumulant in a high-temperature approximation,[35] and by using a cumulant expansion approximation.[36]

Here, the second term in Eq. 7 is approximated as follows; in the summation of Boltzmann factors over all conformations only dominant terms, i.e., native-like compact conformations are taken into account, and then the log function is evaluated in a high temperature approximation.

$$\log\left(\sum_s \exp(-\beta E^{conf}(s, i))\right) \simeq \log\left(\sum_{s\in[\text{native-like conformations}]} \exp(-\beta E^{conf}(s, i))\right) \quad (8)$$

$$\simeq \log\left(\sum_{s\in[\text{native-like}]} 1\right) - \beta < E^{conf}(s, i) >_{\beta=0,\ \text{native-lie}} \quad (9)$$

$$\simeq \log n_r \sigma - \beta(E^{conf} \text{ of a typical native} \cdot \text{structure with the same amino acid composition}) \quad (10)$$

where n_r is the sequence length of a protein and σ is a constant to represent the conformational entropy per residue for native-like structures. The unweighted average of $E^{conf}(s, i)$ over native-like conformations is approximated as the conformational energy expected for a typical native structure with the given amino acid composition, which depends only on amino acid composition. Thus, a scoring function to evaluate compatibilities between sequence and structure is represented as follows

$$-\log(\text{probability of a specific conformation s in a} \cdot \text{sequence i}) \approx \beta E^{conf}(s, i) - \beta(E^{conf} \text{ of a typical native} \cdot \text{structure with the same amino acid composition}) + n\sigma \quad (11)$$

Because judgements on insertions and deletions in sequence-structure alignments are made for every residue, these reference energies must be taken into account for every residue. The energy potentials are modified so that the reference state now corresponds to the zero energy of the potentials. σ in the last term of Eq. 11 is taken to be

independent of residue type and is related to the deletion penalty parameter in sequence-structure alignments.

In the case of short-range energies, the reference energy for each of the short-range potentials defined in Eq. 5 to Eq. 6 would be its average energy over all proteins; see Eqs. 19–24 in Miyazawa and Jernigan.[30] For example, the energy difference with the following reference state should be used for the interaction energy $\delta e^s(s_{q-1}, s_q, s_{q+1}, i_p)$ between backbone and side chain.

$$\delta e^s(i_p, s_{q-1}, s_q, s_{q+1}) - \sum_{s_{q-1}} \sum_{s_q} \sum_{s_{q+1}} \frac{N(i_p, s_{q-1}, s_q, s_q, s_{q+1})}{N(i_p)} \delta e^s(i_p, s_{q-1}, s_q, s_{q+1}) \quad (12)$$

It is achieved here by setting the constant terms in the potentials to force the average energies for all proteins to be zero, that is, to satisfy the following equation for the term above.

$$\sum_{s_{q-1}} \sum_{s_q} \sum_{s_{q+1}} \frac{N(i_p, s_{q-1}, s_q, s_{q+1})}{N(i_p)} \delta e^s(i_p, s_{q-1}, s_q, s_{q+1}) = 0 \quad (13)$$

$N(i_p)$ is the number of residues of type i_p in all protein structures. $N(i_p, s_{q-1}, s_q, s_{q+1})$ is the number of occurrences of a residue of type i_p at residue position p, with the segment at position $(q - 1)$ to $(q + 1)$ in the conformational state (s_{q-1}, s_q, s_{q+1}). Indexes, p and q, are taken to be relative to the 0th residue, the residue under consideration.

For the tertiary structure energies, the reference energy is taken as the average tertiary structure energy per residue for each type of residue in the native protein structures. That is, the following difference in the tertiary structure energy is considered.

$$\Delta E_p^{tert} \equiv \Delta E_p^c + \Delta E_p^r \quad (14)$$

$$\equiv (E_p^c - \langle E_{i_p}^c \rangle) + (E_p^r - \langle E_{i_p}^r \rangle) \quad (15)$$

The second term and the fourth term in Eq. 15 are the average contact energy per residue of type i_p and the average repulsive energy per residue of type i_p in native structures.

If environments surrounding proteins are the same, the stabilities of those proteins can be compared by energy pontentials with their properly evaluated zero energy states. However, in fold and sequence recognition the environments surrounding protein structures are not always the same. Thus, energy potentials need to be modified to approximately measure protein stabilities even for proteins in different environments.

Excluding Intrinsic and Backbone-Backbone Secondary Structure Energies

As already noted, the intrinsic potential and backbone-backbone interaction potentials for secondary structures estimated here depend strongly on the types of protein structures used. If more α proteins were used than β proteins, then the intrinsic potential would be estimated to be lower for α conformations than for β ones. Any set of known protein structures may have such biases. In addition, the secondary potentials are estimated here on the basis of only short-range interactions. There are other types of interactions that affect the stability of secondary structures such as hydrogen bond interactions between β strands. Thus, including only energy terms dependent on residue type may be better for fold recognition; that is, for fold recognition it may be inappropriate to include the backbone-backbone interaction energies, $e^s(s_{-1}, s_0, s_1)$. These will be removed from consideration here.

Subtracting a Collapse Energy From Contact Energies

Many proteins exist in multimeric states. The binding surfaces of such protein monomers are usually just as hydrophobic as the average protein interior.[18] In order to discuss the stabilities of such proteins, the environments surrounding them must be taken into account. For example, the native fold for hemoglobin subunit α can be more properly assessed in the tetrameric state consisting of two α and two β subunits than in the monomeric state. On the other hand, the native structure of myoglobin is appropriately assessed in the monomeric state. For fold recognition, both the native structures, hemoglobin α in the tetrameric state and myoglobin in its monomeric state, are compared with a given sequence. Even if the energy of the former is lower than that of the latter, we cannot say that the former is more stable than the latter, unless the negative binding energies between hemoglobin α and β molecules can overcome the translational and rotational entropy loss due to the formation of a tetramer. Thus, a rigorous assessment of protein stability for fold recognition is not so simple.

To avoid some of these difficulties for fold recognition, only the part, $e_{ij} - e_{rr}$ of the contact energy that depends on the specific side chains, is to be included to assess the compatibilities between sequences and structures. e_{ij} is the contact energy for a pair of residues of type i and j, and e_{rr} reflects the overall compactness of proteins and is defined[18] as an average interaction between residues

$$\exp(-e_{rr}) \equiv \frac{\overline{n_{rr}}\ \overline{n_{00}}}{\overline{n_{r0}}\ \overline{n_{r0}}} \quad (16)$$

$$= \left[\frac{\sum_{i=1} \sum_{j=1} \overline{n_{ij}} \exp(e_{ij})}{n_{rr}} \right]^{-1} \quad (17)$$

$$= \frac{\sum_{i=1} \sum_{j=1} \overline{n_{i0} n_{0j}} \exp(-e_{ij})}{\overline{n_{r0}}\ \overline{n_{0r}}} \quad (18)$$

in RT units, where $\overline{n_{rr}}$, $\overline{n_{r0}}$, and $\overline{n_{00}}$ are the statistical averages of the total number of contacts between residues n_{rr}, the total number of contacts between residues and

effective solvents n_{r0}, and the total number of contacts between effective solvents n_{00}, respectively; refer to Miyazawa and Jernigan[18,19] for details. $e_{ij} - e_{rr}$ has removed the homogeneous energy for protein collapse and consists only of the remaining energy depending on the specific types of side chains. This quantity takes positive values for contacts between polar residues and negative values for hydrophobic pairs. If e_{rr} were not removed, then the total contact energy would give the lower values to conformations with larger numbers of contacts, e.g., the energies of folds in multimers would usually be lower than those in monomers.

After all of the considerations above are included, the following quantity is considered to be appropriate for assessing compatibilities between protein sequences and structures.

$$\Delta E_p^{conf}(e_{i_p j} - e_{rr}) \equiv \Delta E_p^{sec} + \Delta E_p^{tert}(e_{i_p j} - e_{rr}) \qquad (19)$$

where $e_{ij} - e_{rr}$ within parenthesis means it is the argument of the function. The intrinsic and backbone-backbone interaction energies are excluded in the first term.

$$\Delta E_p^{sec} \simeq \Delta e^s(\ldots, s_{p-1}, i_p, s_p, s_{p+1}, \ldots) \equiv \sum_{p-3 \le q \le p+3} \delta e^s(i_p, s_{q-1}, s_q, s_{q+1}) \qquad (20)$$

Parenthetically, it should be noted that the reference energies of Eq. 13 and of the second and fourth terms in Eq. 15 do not depend on conformation but only on the amino acid composition of sequence. Therefore, if deletions and insertions in sequence-structure alignments are not allowed and energy values are compared among different conformations for an identical sequence, subtracting those reference energies does not actually matter, but they are necessary when sequences are varied.

RESULTS

Total Alignment Energies and a Reference State

In Figure 1, estimates of total alignment energies per residue, $\Delta E^{conf}(e_{ij} - e_{rr})/ n_r$ which include contributions of secondary structure energy, contact energy, and repulsive packing energy and are defined by Eqs. 19 for 189 protein representatives are plotted against $n_r^{-1/3}$; where n_r is the sequence length of each protein. These representatives of protein structures differ from each other substantially and have less than 35% sequence identity as selected by Orengo et al.[38]; see their Table I. Coordinate files with too many unknown atomic coordinates are excluded from these datasets. In the estimates of alignment energies, contact energies e_{ij} for a pair of residues of type i and j are modified by subtracting a collapse energy per contact, e_{rr}. The long-range tertiary structure energies are calculated for multimeric states only if the coordinates of the other bound molecules are given in the PDB[37] file. See Figure 6B of Miyazawa and Jernigan[19] for the chain length dependence of the tertiary structure energies, that is, for plots of $\Delta E^{tert}(e_{ij} - e_{rr})/n_r$ alone versus $n_r^{-1/3}$. Including the short-

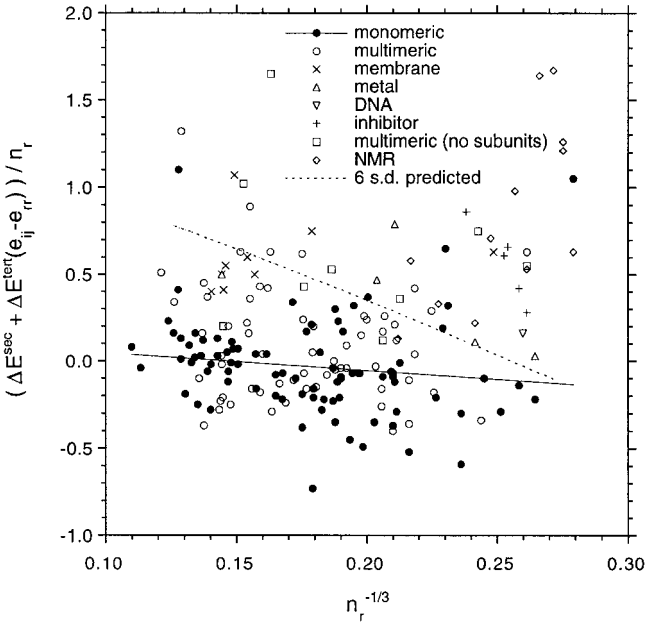

Fig. 1. The total alignment energy per residue with a collapse energy subtracted to remove the protein size dependence; see Eq. 19 for $\Delta E^{conf}(e_{ij} - e_{rr})$. The tertiary structure energies include the contact energies and the repulsive packing energies but not the hard core repulsion energies, i.e., $e^{hc} = 0$ in Eq. 41 of Miyazawa and Jernigan.[19] The long-range tertiary structure energies are calculated in a multimeric state, only if the coordinates of other bound molecules are given in the PDB file. All energies are given here in RT units. The representative protein structures used here are 189 protein structures that differ from each other by having no more than 35% sequence identity and are those selected by Orengo, et al.[38]; see their Table I. Proteins with many unknown atomic coordinates are not included. The solid circles show the values for monomeric proteins determined by X-ray, not including membrane proteins, metal binding proteins, DNA binding proteins, and inhibitors and multimeric proteins not given in their complete assembly in the coordinate files. The open circles are proteins whose structures are given in at least partial, if not full assembly of subunits. A solid line shows the regression line for the monomeric proteins. A collapse energy has been removed, so that the regression line is almost flat, $\Delta E^{conf}(e_{ij} - e_{rr})/n_r = 0.15 - 1.0 n_r^{-1/3}$, and the correlation coefficient is 0.13. The dotted line shows energy values that correspond to -6 in standard deviation units from the mean in the distributions of threadings predicted by Eq. 21; the equation of the dotted line is $(2.89 \cdot n_r^{0.932} - 6 \cdot 1.34 \cdot n_r^{0.682})/ n_r$. The entry names and sequence identifiers of the PDB files used in this figure are:

Membrane proteins:
1PRC-L 1PRC-M 1PRC-C 2POR 1SN3 1VSG-A 1HGE-A 1HGE-B 1PRC-H

Metal binding proteins:
1CY3 1PRC-C 5RXN 2HIP-A 2CDV

DNA binding proteins:
1HDD-C

Inhibitors without an enzyme:
1HOE 1PI2 3EBX 2OVO 5PTI

Multimeric proteins without subunit interactions:
2WRP-R 1UTG 1ROP-A 2TMV-P 2RHE 2STV 3PGM 6LDH 1PYP

Structures determined by NMR:
1C5A 1HCC 1ATX 1SH1 2SH1 1EPG 4TGF 3TRX 1EGO
1APS 1IL8-A 2GB1

Other monomeric proteins:

1MBC	1MBA	1ECD	2LH3	2LHB	1R69	4ICB	4CPV	1LE2	1YCC
1CC5	451C	1IFC	1RBP	1SGT	4PTP	2SGA	2ALP	2SNV	1CD8
1CD4	1ACX	1PAZ	1PCY	1GCR	2CNA	3PSG	1F3G	8I1B	1ALD
1PII	6XIA	2TAA-A	4ENL	5P21	4FXN	2FCR	2FX2	3CHY	5CPA
8DFR	3DFR	3ADK	1GKY	1RHD	4PFK	3PGK	2GBP	8ABP	2LIV
1TRB	1IPD	4ICD	1PGD	8ADH	2TS1	1PHH	3LZM	1LZ1	1RNH
7RSA	1CRN	1CTF	1FXD	2FXB	4FD1	1FDX	4CLA	9RNT	1RNB-A
1FKF	1SNC	1UBQ	3B5C	9PAP	3BLM	2CPP	1CSC	1ACE	1COX
1GLY	1LAP	2CYP	8ACN	2CA2					

Other multimeric proteins:

1HBB-A	2SDH-A	1ITH-A	1COL-A	1LMB-A	3SDP-A	2SCP-A	2HMZ-A	256B-A	2CCY-A
1GMF-A	1BBP-A	2FB4-H	3HLA-B	1COB-A	2AZA-A	2PAB-A	1BMV-1	1BMV-2	2PLV-1
1TNF-A	2MEV-1	2MEV-2	2MEV-3	2PLV-2	2PLV-3	2LTN-A	2RSP-A	2ER7-E	5HVP-A
1NSB-A	5TIM-A	2TRX-A	1CSE-E	1GP1-A	4DFR-A	8CAT-A	4MDH-A	1GD1-0	7AAT-A
1HRH-A	1RVE-A	2SIC-1	8ATC-B	2TSC-A	2SAR-A	1MSB-A	1BOV-A	1FXI-A	1TGS-I
1TPK-A	9WGA-A	3HLA-A	8ATC-A	2CPK-E	1GST-A	10VA-A	7API-A	1WSY-B	2GLS-A
2PMG-A	6TMN-E	3GAP-A							

range energies of secondary structures does not change these characteristics. As expected, overall, there is no correlation between the two quantities for monomeric proteins, and the mean energy for monomeric proteins is centered about zero, an average protein being taken as a reference state for the alignment energies. Therefore, subtracting a collapse energy, e_{rr}, from the contact energies has removed most of the dependence of the total contact energies on the surface area of the proteins. Membrane proteins, which are shown as crosses, tend to have much higher values of $\Delta E^{conf}(e_{ij} - e_{rr})/n_r$ than for average monomeric proteins. This is expected because the hydrophobic environment surrounding membrane proteins is not taken into account and in the present calculation they are incorrectly presumed to be in water. Similarly, an exception is seen for metal binding proteins and DNA binding proteins in which the metals and DNA have been treated here only as holes filled with water. Also, the multimeric cases given as open circles tend to be located above the solid line, probably because the coordinates of intermolecular neighbors in some PDB files are given incompletely as partially assembled structures. On the other hand, the high values of energies for proteins whose structures were determined by NMR may indicate the relatively poorer resolution of these structures.[39]

Predicted energy values that correspond to −6 standard deviation units from the mean in the distributions of threadings are shown by the dotted line in Figure 1. Proteins above the dotted line have significantly high conformational energies, probably because some interactions are not properly taken into account in the present estimations, such as intermolecular interactions and disulphide bonds, especially for small proteins.

Simple Threading Without Gaps

Eighty-eight proteins determined to a resolution better than 2.5 Å by X-ray analyses that are structurally dissimilar to each other (with values smaller than 80 on the scale of Orengo, et al.[38] for structure similarity) are threaded into each of the 189 representatives of protein structures. The 88 proteins are a subset of the 189 protein representatives whose names are listed in the caption of Figure 1. Proteins classified within the multidomain group by those authors[38] are excluded from the set of sequences to be threaded.

The total alignment energy $\Delta E^{conf}(e_{ij} - e_{rr})$ is calculated for protein sequences threaded at all possible positions in all other protein structures, and their means and standard deviations are also calculated; no gaps in either the sequences or the structures are allowed. The long-range tertiary structure energies are calculated for multimeric states only if the coordinates of the other bound subunits are given in the PDB file. Then, the positions of the native energies in the distributions of all threadings are measured in units of standard deviations (s.d.) where negative values indicate that the native energies are below the mean. Table I lists energies per residue, $\Delta E^{conf}(e_{ij} - e_{rr})/n_r$, for the set of protein sequences threaded into their own set of native structures, as well as the ranks and z-scores, i.e., positions of the native energies from the mean in units of s.d., in the distributions of all threadings; proteins are listed in increasing order of the values of $\Delta E^{conf}(e_{ij} - e_{rr})/n_r$ in units of standard deviation.

For most proteins, the native structures have significantly low values in s.d. units of $\Delta E^{conf}(e_{ij} - e_{rr})/n_r$ and rank lowest in energy. These good results are obtained by subtracting the collapse energies from the contact energies and excluding the side-chain independent terms in the secondary structure potentials; otherwise, structures with more contacts such as multimeric proteins and with more secondary structures of specific types would tend to be more highly ranked. However, there are some proteins for which the native structures are not the best or not even significantly better than most others. As pointed out previously,[19] proteins with values worse (higher) than −6.0 s.d., are always membrane proteins, proteins such as small inhibitors, single subunits whose coordinates are given in isolated forms without their bound counterparts, or proteins bound to metal ions or other molecules. Those proteins are marked by daggers in Table I. For these small proteins, the relative proportions of binding regions on their surfaces may be unusually large.

Table I also shows the results where only secondary structure energies or only tertiary structure energies are used for ranking. The proteins, beef liver catalase (8CAT-A), cytochrome C_3 (1CY3), and wheat germ agglutinin (9WGA-A), whose total secondary structure energies in s.d. units are significantly low, are proteins that contain unusually large numbers of Pro's in 8CAT-A and 1CY3, or Gly's in 9WGA-A. Proteins containing many Pro's and/or Gly's tend to have low values of the secondary structure energies in s.d. units, because of the distinct conformational characteristics of Pro's and Gly's. In comparison to the tertiary structure energies, the secondary structure energies do not contribute so much to selecting the native folds over other folds. The secondary structure energies of most native folds are within −6 s.d. units of the mean energies. The correlation of the energy values of native folds in s.d. units between the short-range secondary structure energies and the long-range tertiary structure energies is weak, with a correlation coefficient of 0.32 for all proteins and 0.54 even if the Pro/Gly rich proteins of 8CAT-A, 1CY3, and 9WGA-A are excluded. It is reasonable in view of the expectation that overall folds are recognized by long-range interaction energies, and local conformations are recognized by short-range interaction energies.

As shown in Figure 2A, the addition of the secondary structure energies to the tertiary structure energies almost always improves the discrimination of the native structures from other folds; however, the improvement in the case of threading is not large, because in threading the average of secondary structure energies of native structures in s.d. units is much less negative than that of tertiary structure energies; see Figures 3A and 3B. This does not mean that secondary structure energies need not be taken into account. As noted in the next section, the

TABLE I. Positions of Native Sequence-Structure Pairs in the Energy Distributions of Threadings and Inverse Threadings

PDB name	n_r	#threadings	$\Delta E^{sec}/n_r$ (RT)[a]	Threading Rank[b]	Threading (s.d.)[c]	Inverse threading Rank[d]	Inverse threading (s.d.)[e]	$\Delta E^{tert}(e_{i_p j} - e_{rr})/n_r$ (RT)[a]	Threading Rank[b]	Threading (s.d.)[c]	Inverse threading Rank[d]	Inverse threading (s.d.)[e]	$\Delta E^{conf}(e_{i_p j} - e_{rr})/n_r$ (RT)[a]	Threading Rank[b]	Threading (s.d.)[c]	Inverse threading Rank[d]	Inverse threading (s.d.)[e]
1PGD	469	765	0.11	1	−7.3	1	−10.7	−0.10	1	−10.9	1	−10.4	0.01	1	−12.5	1	−13.3
8CAT-A	498	548	0.04	1	−11.7	1	−11.6	0.30	1	−8.6	1	−7.5	0.34	1	−12.3	1	−12.4
1PII	452	953	−0.00	1	−5.8	1	−12.4	−0.19	1	−10.1	1	−9.8	−0.19	1	−11.0	1	−13.4
3PGK	415	1426	−0.08	1	−5.8	1	−13.9	0.10	1	−9.1	1	−9.1	0.02	1	−10.5	1	−15.3
9WGA-A	170	13451	−0.52	1	−8.8	1	−12.0	0.37	1	−6.1	14	−3.8	−0.15	1	−10.5	1	−10.7
2ER7-E	330	3548	−0.08	1	−5.8	1	−12.1	−0.12	1	−9.8	1	−8.2	−0.21	1	−10.4	1	−12.3
4ENL	436	1146	0.02	1	−5.3	1	−11.6	0.07	1	−9.7	1	−7.8	0.09	1	−10.2	1	−11.6
4PTP	223	9338	−0.27	1	−6.8	1	−11.9	0.07	1	−8.1	1	−5.9	−0.20	1	−10.1	1	−10.9
2LIV	344	3084	−0.01	1	−5.0	1	−11.0	0.04	1	−10.0	1	−7.8	0.03	1	−9.9	1	−11.4
1ALD	363	2538	−0.08	1	−5.3	1	−11.7	0.06	1	−9.5	1	−7.3	−0.03	1	−9.9	1	−11.7
1GCR	174	13171	−0.53	1	−6.8	1	−12.2	−0.20	1	−7.8	1	−7.9	−0.73	1	−9.8	1	−13.2
8ADH	374	2254	0.01	1	−5.0	1	−11.1	−0.07	1	−9.9	1	−8.7	−0.06	1	−9.6	1	−12.3
2GBP	309	4402	−0.02	1	−4.6	1	−9.8	0.01	1	−9.9	1	−7.3	−0.01	1	−9.6	1	−10.5
7AAT-A	401	1681	0.07	1	−4.2	1	−10.3	−0.17	1	−11.1	1	−9.5	−0.10	1	−9.5	1	−12.4
2FCR	173	13264	−0.04	1	−5.2	1	−7.9	−0.17	1	−8.3	1	−6.2	−0.21	1	−9.2	1	−9.0
1GKY	186	12058	−0.06	1	−4.8	1	−8.2	−0.14	1	−8.8	1	−7.2	−0.19	1	−9.2	1	−9.9
1GD1-O	334	3406	−0.03	1	−4.7	1	−10.8	0.06	1	−9.5	1	−7.2	0.03	1	−9.1	1	−10.7
4PFK	319	3972	−0.02	1	−5.1	1	−10.5	0.07	1	−8.7	1	−7.4	0.05	1	−8.9	1	−11.2
1F3G	151	15407	−0.20	1	−5.7	1	−9.5	−0.14	1	−7.6	1	−6.0	−0.35	1	−8.9	1	−9.9
5TIM-A	249	7560	−0.02	1	−4.6	1	−9.9	−0.16	1	−8.8	1	−7.3	−0.18	1	−8.9	1	−10.4
1COB-A	151	15407	−0.07	1	−5.8	1	−8.1	0.02	1	−7.4	1	−5.4	−0.05	1	−8.9	1	−8.8
1PHH	394	1808	0.08	1	−4.1	1	−10.5	−0.05	1	−9.9	1	−9.3	0.03	1	−8.8	1	−12.5
1NSB-A	390	1889	−0.08	1	−5.2	1	−12.9	0.24	1	−8.3	1	−6.1	0.16	1	−8.8	1	−11.4
6XIA	387	1953	−0.01	1	−4.8	1	−11.7	0.12	1	−8.5	1	−8.2	0.11	1	−8.8	1	−12.1
4FXN	138	16873	0.01	1	−4.1	1	−7.0	−0.46	1	−9.3	1	−7.2	−0.45	1	−8.8	1	−9.4
1IPD	345	3052	0.07	1	−4.9	1	−9.9	0.06	1	−8.6	1	−7.9	0.13	1	−8.8	1	−11.2
3ADK	194	11413	0.11	1	−4.0	1	−6.7	−0.21	1	−9.1	1	−7.7	−0.10	1	−8.8	1	−9.8
2TS1	317	3972	0.03	1	−4.0	1	−9.6	−0.14	1	−9.6	1	−9.1	−0.12	1	−8.7	1	−12.0
1CSE-E	274	6151	−0.05	1	−5.8	1	−10.6	0.27	1	−7.1	1	−5.6	0.22	1	−8.6	1	−9.6
1RHD	293	5173	0.01	1	−4.6	1	−10.3	−0.03	1	−8.7	1	−8.0	−0.02	1	−8.5	1	−11.8
1PAZ	120	19074	−0.11	1	−4.7	1	−7.6	−0.24	1	−7.9	1	−6.2	−0.35	1	−8.3	1	−9.0
2CNA	237	8389	0.05	1	−4.6	1	−8.7	−0.00	1	−7.7	1	−6.8	0.04	1	−8.3	1	−10.0
1ACX	108	20669	−0.14	1	−5.7	1	−8.3	0.07	1	−6.7	1	−5.0	−0.07	1	−8.3	1	−8.7
3LZM	164	14113	−0.02	1	−4.0	1	−7.7	−0.26	1	−8.4	1	−7.7	−0.28	1	−8.2	1	−10.5
2LTN-A	181	12563	−0.08	1	−4.9	1	−8.8	−0.09	1	−7.4	1	−6.8	−0.16	1	−8.2	1	−10.2
3CHY	128	18067	−0.04	1	−4.0	1	−7.0	−0.45	1	−8.2	1	−7.4	−0.49	1	−8.2	1	−9.3
1COL-A	197	10951	0.13	1	−3.6	1	−6.0	−0.25	1	−8.2	1	−7.0	−0.11	1	−8.2	1	−8.7
1RVE-A	244	7876	0.13	1	−3.8	1	−8.3	−0.09	1	−8.7	1	−8.0	0.04	1	−8.1	1	−10.9
4CPV	108	20530	−0.17	1	−4.4	1	−7.5	−0.20	1	−7.6	1	−6.0	−0.37	1	−8.1	1	−8.7
2TSC-A	264	6713	0.05	1	−4.2	1	−8.8	−0.21	1	−8.3	1	−8.4	−0.16	1	−8.1	1	−11.3
2TRX-A	108	20669	0.04	1	−3.8	1	−5.5	−0.44	1	−7.8	1	−6.8	−0.40	1	−7.9	1	−8.1
1MBC	153	15196	0.13	1	−3.2	1	−5.3	−0.36	1	−8.5	1	−7.8	−0.23	1	−7.8	1	−9.3
4DFR-A	159	14598	0.07	1	−4.0	1	−6.9	−0.15	1	−7.7	1	−6.6	−0.08	1	−7.7	1	−8.7
1MSB-A	115	19724	−0.08	1	−4.3	1	−7.3	−0.08	1	−7.4	1	−5.0	−0.16	1	−7.7	1	−7.8
4ICB	76	25579	−0.10	1	−3.9	1	−5.9	−0.49	1	−7.3	1	−6.8	−0.59	1	−7.7	1	−8.5
8ATC-B	146	15196	0.16	1	−3.4	1	−5.6	−0.20	1	−8.2	1	−7.0	−0.04	1	−7.7	1	−8.6
1YCC	108	20669	−0.11	1	−4.6	1	−7.2	0.03	1	−6.8	1	−4.8	−0.09	1	−7.7	1	−7.6
1FKF	107	20812	−0.05	1	−4.5	1	−6.4	−0.07	1	−6.9	1	−5.4	−0.12	1	−7.6	1	−7.9
9RNT	104	21250	−0.13	1	−5.0	1	−7.3	0.11	1	−6.4	1	−4.7	−0.01	1	−7.6	1	−7.7
4CLA	213	10054	0.05	1	−3.8	1	−7.8	−0.27	1	−7.4	1	−8.8	−0.22	1	−7.5	1	−10.9
2RSP-A	115	18565	−0.14	1	−4.7	1	−7.8	−0.11	1	−6.4	1	−5.7	−0.26	1	−7.4	1	−8.1
1LZ1	130	17821	−0.05	1	−4.0	1	−7.2	−0.03	1	−7.4	1	−5.5	−0.07	1	−7.3	1	−8.5
5CPA	307	4493	0.06	1	−4.0	1	−9.7	0.05	1	−7.3	1	−6.8	0.11	1	−7.2	1	−10.3
2RHE[f]	114	19857	−0.10	1	−5.3	1	−6.9	0.23	1	−5.4	1	−4.6	0.12	1	−7.2	1	−7.4
5P21	166	13922	0.09	1	−3.5	1	−6.3	−0.04	1	−7.2	2	−5.8	0.05	1	−7.1	1	−8.0
6LDH[f]	329	3548	0.08	1	−3.9	1	−9.7	0.12	1	−7.2	1	−7.8	0.20	1	−7.1	1	−11.2
1UBQ	76	25579	−0.09	1	−3.7	1	−5.6	−0.21	1	−6.8	1	−5.9	−0.30	1	−7.1	1	−7.3
1BOV-A	69	26735	−0.05	1	−3.4	1	−5.3	−0.29	1	−6.9	2	−4.8	−0.34	1	−7.0	1	−6.4
1PRC-C[f]	333	3441	0.02	1	−4.6	1	−10.6	0.48	1	−5.7	1	−5.3	0.50	1	−6.9	1	−10.0
2AZA-A	129	17943	0.03	1	−3.7	1	−6.5	0.12	1	−6.8	4	−4.5	0.15	1	−6.9	1	−6.9
1RBP	175	13081	0.08	1	−3.6	1	−7.2	0.12	1	−6.8	1	−5.6	0.21	1	−6.8	1	−8.1
1RNH	148	15300	0.18	1	−3.2	1	−4.8	0.05	1	−7.2	1	−5.4	0.23	1	−6.8	1	−6.8
2SAR-A	96	22443	−0.07	1	−4.4	1	−6.5	0.10	1	−5.8	1	−4.5	0.04	1	−6.8	1	−7.2

TABLE I. (Continued)

PDB name	n_r	#threadings	$\Delta E^{sec}/n_r$ (RT)[a]	Threading Rank[b]	Threading (s.d.)[c]	Inverse threading Rank[d]	Inverse threading (s.d.)[e]	$\Delta E^{tert}(e_{i_p j} - r_{rr})/n_r$ (RT)[a]	Threading Rank[b]	Threading (s.d.)[c]	Inverse threading Rank[d]	Inverse threading (s.d.)[e]	$\Delta E^{conf}(e_{i_p j} - r_{rr})/n_r$ (RT)[a]	Threading Rank[b]	Threading (s.d.)[c]	Inverse threading Rank[d]	Inverse threading (s.d.)[e]
1LMB-A	87	23056	−0.03	1	−3.3	1	−5.4	−0.16	1	−6.5	1	−5.3	−0.18	1	−6.7	1	−7.0
2SIC-I	107	20812	0.12	1	−3.8	1	−5.4	0.08	1	−6.1	2	−5.0	0.20	1	−6.7	1	−7.0
3B5C	86	23674	−0.07	1	−3.6	1	−6.0	−0.14	1	−6.5	1	−6.2	−0.21	1	−6.6	1	−8.0
1FXD	58	28625	−0.14	1	−3.6	1	−5.4	0.00	1	−6.6	16	−3.6	−0.14	1	−6.6	1	−5.6
1GMF-A	119	19203	0.09	1	−3.4	1	−5.6	−0.12	1	−6.4	1	−6.1	−0.03	1	−6.6	1	−7.7
256B-A	106	20957	0.08	1	−3.1	1	−4.7	0.04	1	−6.7	1	−5.2	0.12	1	−6.6	1	−6.6
2PAB-A	114	18692	0.05	1	−3.7	1	−5.9	0.12	1	−5.7	1	−5.2	0.17	1	−6.3	1	−7.2
1CY3[f]	118	19333	−0.58	1	−7.2	1	−10.6	1.04	>100	−1.4	>100	−2.0	0.47	1	−6.3	1	−9.3
5RXN[f]	54	29347	−0.14	1	−4.3	1	−5.7	0.17	1	−5.0	2	−4.0	0.03	1	−6.2	1	−6.2
7RSA	124	18565	0.08	1	−3.2	1	−5.8	0.29	1	−6.1	3	−4.0	0.37	1	−6.0	1	−5.9
2HIP-A[f]	71	26400	−0.02	1	−3.6	1	−5.5	0.13	1	−5.5	4	−3.9	0.11	1	−5.9	1	−6.2
1TPK-A[f]	88	23674	−0.12	1	−4.4	1	−6.5	0.41	11	−4.1	63	−3.0	0.29	1	−5.7	1	−5.8
2STV[f]	184	11413	0.07	1	−3.9	1	−7.2	0.36	1	−4.5	3	−4.5	0.43	1	−5.6	1	−7.3
2OVO[f]	56	28983	0.02	1	−3.2	1	−4.3	0.26	1	−4.9	24	−3.2	0.28	1	−5.5	1	−4.8
2WRP-R[f]	104	20812	0.14	1	−2.7	1	−4.2	0.22	1	−4.5	1	−5.7	0.36	1	−4.8	1	−6.9
5PTI[f]	58	28625	−0.04	1	−3.5	1	−4.8	0.46	44	−3.6	>100	−2.6	0.42	1	−4.8	1	−4.9
1SN3[f]	65	27410	0.22	1	−3.0	1	−3.8	0.41	5	−4.1	29	−3.0	0.63	1	−4.7	1	−4.5
2CDV[f]	107	20812	−0.01	1	−4.0	1	−6.5	0.80	49	−2.8	17	−3.0	0.79	1	−4.7	1	−5.9
3EBX[f]	62	27925	0.07	1	−3.1	1	−4.5	0.53	14	−3.8	>100	−2.3	0.61	1	−4.6	1	−4.2
1PI2[f]	61	27752	0.13	1	−2.8	1	−3.9	0.54	26	−3.7	35	−3.1	0.67	1	−4.2	1	−4.6
1UTG[f]	70	26567	0.21	1	−2.5	23	−2.9	0.54	15	−3.7	7	−3.5	0.75	1	−4.2	1	−4.3
2POR[f]	301	4778	0.35	1	−2.8	1	−6.5	0.72	3	−3.3	12	−3.3	1.07	1	−4.1	1	−6.4
1PRC-L[f]	273	6205	0.18	1	−4.1	1	−7.5	0.42	>100	−2.0	37	−3.2	0.60	1	−4.0	1	−6.7
1HOE[f]	74	25905	0.08	1	−3.3	1	−4.6	0.78	>100	−2.2	>100	−1.7	0.86	1	−3.7	9	−3.6
1CRN[f]	46	30832	0.35	7	−2.3	>100	−1.7	0.70	>100	−2.6	>100	−1.7	1.05	9	−3.3	>100	−2.2

[a]Energy of the native structure in *RT* units.
[b]Rank of the native structure in the energy distribution of all threaded structures.
[c]Energy of the native structure in standard deviation units from the mean in the energy distribution of all threaded structures.
[d]Rank of the native sequence in the energy distribution of all inverse threadings.
[e]Energy of the native sequence in standard deviation units from the mean in the energy distribution of all inverse threadings.
[f]Proteins such as membrane proteins, metal binding proteins, inhibitors without an enzyme, and multimeric proteins without subunit interactions; see Table 5B of Miyazawa and Jernigan.[19]

average of secondary structure energies of native sequences in s.d. units is significantly large and negative in the case of inverse threading. Also, it will be shown in another paper that the secondary structure energies, which recognize local conformations, are useful to achieve a correct positioning of residues in sequence-structure alignments.

The native energies, in standard deviation units, can be predicted without carrying out sequence threadings. The mean and variation of energies of random threadings can be approximated by simple functions of sequence length. If surface effects on the number of contacts are ignored, their means and variations will be roughly proportional to sequence length. Thus, the means and variations of energies of threadings for those proteins listed in Table I can be fit with a function of the form $a \cdot n_r^b$ with a log-log plot, where a and b are two parameters. These evaluated parameters are listed in Table II. The correlation coefficients are all larger than 0.96, indicating this fitting function to be appropriate. Estimated values of the exponents for standard deviations are much larger than 0.5 and those for the means are slightly less than 1.0, probably due to variations in the number of contacts per residue or surface effects on the number of contacts. As a result, the native conformational energies in standard deviation units can be predicted by using these estimated relationships as follows.

$$\text{(Native energy in standard deviation units)} \sim ((\text{native energy}) - 2.89 n_r^{0.932})/(1.34 n_r^{0.682}) \quad (21)$$

In the equation above, the native energies are total energies including both secondary and tertiary structure energies. The correlation between predicted and observed values is highly significant with a correlation coefficient of 0.90 and a slope of 0.98. Two exceptions for which the observed values do differ significantly from the predicted values are 8CAT-A and 1PGD which are the two longest proteins in this protein set. Eq. 21 can be generally useful to provide estimates without detailed calculations.

Inverse Threading Without Gaps

In order to examine how the present energy function recognizes native sequences for a given protein structure, a given protein structure is threaded with all possible partial sequences of other proteins, and their alignment energies $\Delta E^{conf}(e_{ij} - e_{rr})$ are compared. It should be noted that the total average of each type of interaction energy over all native structures is set to zero by subtracting the

average energy of the native structures as a reference state. This inverse threading is carried out for the structures of the same set of proteins whose sequences have been used in threading in the previous section.

Table I also shows the ranks of the native sequences and their energies in standard deviation units from the mean in the energy distributions of all inverse threadings. Unlike the usual threading, the native alignment energies in s.d. units for inverse threading depend on the choice of the reference state for the secondary structure and tertiary structure potential energies, which is defined as the zero energy state; that is, they depend on the second and fourth terms of Eq. 15 for the tertiary structure potential and on Eq. 12 for the secondary structure potentials. On the other hand, subtracting the collapse energies from the contact energies and excluding the side-chain independent terms in the secondary structure potentials do not change the native alignment energies in s.d. units. Therefore, inverse threading provides a test for examining whether the choice of the reference state, i.e., the approximation of Eq. 11, is appropriate for native sequence recognition with a structure. On the other hand, threading is a test for the adequacy of the conformation dependent terms, i.e., the subtraction of the collapse energies from the contact energies and of the exclusion of the side-chain independent terms in the secondary structure potentials.

In both the secondary and the tertiary structure potentials, most native sequences rank at the lowest energy. The correlation of the energy values of native structure-sequence pairs in s.d. units between the short-range secondary structure energies and the long-range tertiary structure energies is higher than for the case of normal threading, with a correlation coefficient of 0.62 for all proteins and 0.70 for the set of proteins excluding the Pro/Gly rich proteins 8CAT-A, 1CY3, and 9WGA-A. As for threading, membrane proteins and proteins, given in isolated forms without their binding counterparts, such as small inhibitors and other subunits, and proteins binding metallic ions or other molecules, are also exceptions in inverse threading. These results indicate that the choice of the reference state at zero energy is quite good for both secondary structure potentials and the contact energy potentials in order to recognize native sequences from other non-native sequences for given protein structures.

Figure 3 shows comparisons of native energies in s.d. units for inverse threading and conventional threading. One of interesting features is that the secondary structure energies of native structure-sequence pairs in s.d. units

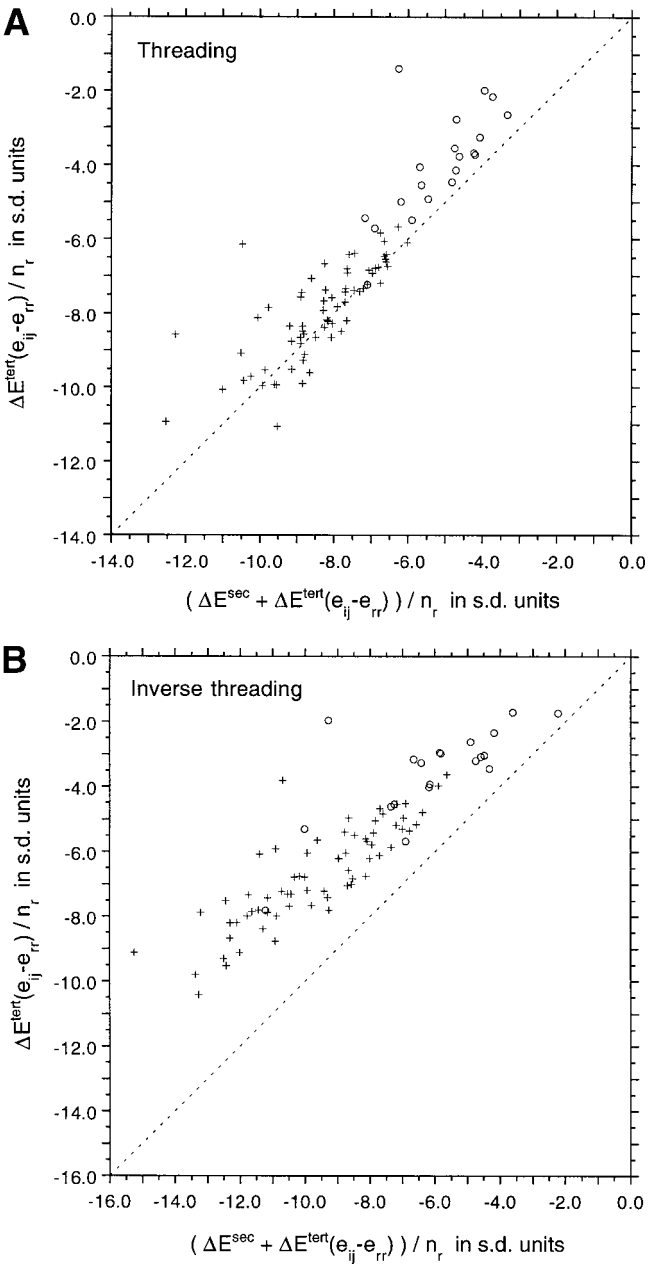

Fig. 2. **A:** The effects of the secondary structure energies on the discrimination of native structures among other folds with a given sequence and (**B**) inversely on the recognition of native sequences among other non-native sequences with a given structure. The total alignment energies per residue $\Delta E^{conf}(e_{ij} - e_{rr})/n_r$ including the secondary structure energies, and the tertiary structure energies per residue $\Delta E^{tert}(e_{ij} - e_{rr})/ n_r$ alone of the native structures or sequences of 88 proteins, in standard deviation units from the mean in the energy distribution of random threadings, are plotted against each other; see Eq. 19. Sequences of 88 proteins whose structures were determined to a resolution better than 2.5 Å by X-ray analyses and are structurally dissimilar to each other are passed through each of the 189 representative proteins, that differ from each other by having no more than 35% sequence identity, as selected by Orengo, et al.[38] Inversely the 88 protein structures are threaded by the 189 protein sequences in the case of inverse threading. Then, the total conformational energies as well as the tertiary structure energies according to Eq. 19 for all threadings, with no gaps allowed, and its mean and standard deviation are calculated. The long-range tertiary structure energies are calculated in a multimeric state, only if the coordinates of other bound molecules are given in a PDB file. The 189 protein representatives are the same ones used for Figure 1. The 88 proteins are a subset of these 189 protein representatives and are listed in Table I; proteins[38] classified within the multidomain group are excluded from this set of 88 proteins. Coordinate files with too many unknown atomic coordinates are also excluded from these data sets. Proteins, which were classified in Figure 1 and marked by footnote f in Table I as membrane proteins, metal or DNA binding proteins, inhibitors without an enzyme, and multimeric proteins without subunit interactions, are plotted here as open circles.

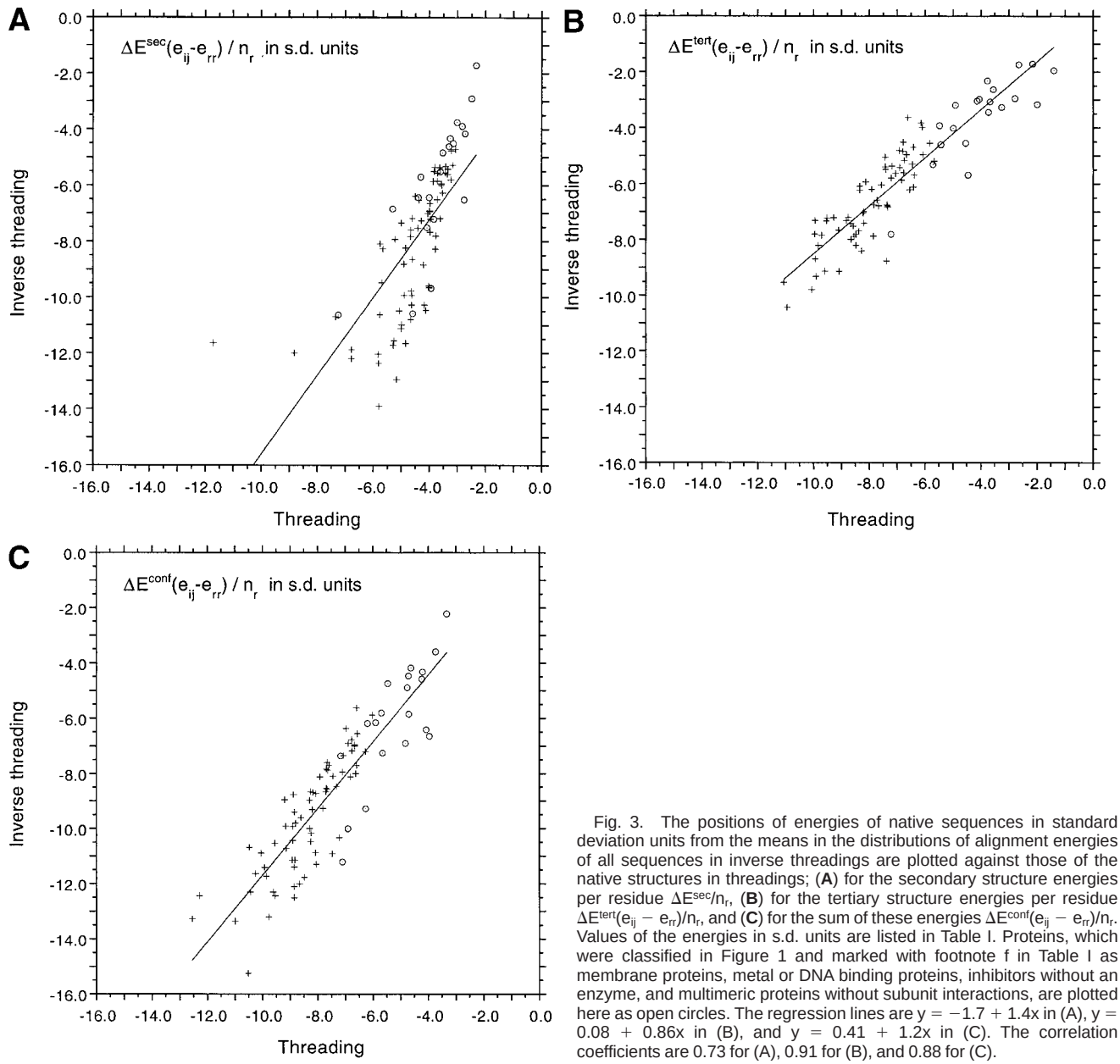

Fig. 3. The positions of energies of native sequences in standard deviation units from the means in the distributions of alignment energies of all sequences in inverse threadings are plotted against those of the native structures in threadings; (**A**) for the secondary structure energies per residue $\Delta E^{sec}/n_r$, (**B**) for the tertiary structure energies per residue $\Delta E^{tert}(e_{ij} - e_{rr})/n_r$, and (**C**) for the sum of these energies $\Delta E^{conf}(e_{ij} - e_{rr})/n_r$. Values of the energies in s.d. units are listed in Table I. Proteins, which were classified in Figure 1 and marked with footnote f in Table I as membrane proteins, metal or DNA binding proteins, inhibitors without an enzyme, and multimeric proteins without subunit interactions, are plotted here as open circles. The regression lines are $y = -1.7 + 1.4x$ in (A), $y = 0.08 + 0.86x$ in (B), and $y = 0.41 + 1.2x$ in (C). The correlation coefficients are 0.73 for (A), 0.91 for (B), and 0.88 for (C).

tend to be significantly more negative in inverse threading than in normal threading; see Figure 3A. Similar behavior was also reported by DeWitte & Shakhnovich[29]; in their analyses, the average z-score where only local interactions were taken into account is about -2 for the discrimination of native sequences from a random collection of shuffled sequences, a larger value than about -0.5 for discriminating native structures from a group of "random structures." The mean of the energies of random threadings ought to be the same for both inverse threading and normal threading, and it is actually almost the same for the present protein set. The origin of this difference is the fact that the standard deviations of energies of random threadings in inverse threading is about half of that for normal threading; see Table II. This indicates that the differences in secondary structures among proteins yield larger variations in energy than do the differences of amino acid sequences. As a result, the addition of secondary structure energies to tertiary structure energies significantly improves the positions of native energies in standard deviation units; see Figure 2B. On the other hand, the tertiary structure energies of the native sequences in standard deviation units in inverse threading are slightly higher than in normal threading; see Figure 3B and 3C.

The means and standard deviations of energies of inverse threadings for those proteins listed in Table I are also fitted to the function $a \cdot n_r^b$ with a log-log plot, where a and b are two parameters. Estimated values of these parameters are given in Table II. Except for the standard deviation of the secondary structure energy, the mean and

TABLE II. Observed Equations for the Means and Standard Deviations in the Distributions of Energies of Random Threadings, Which are Estimated from Linear Fitting of the Log-log Plots of Energy Versus Sequence Length

	Threading				Inverse threading			
	Mean		Standard deviation		Mean		Standard deviation	
	a	b	a	b	a	b	a	b
Total secondary structure energy	$1.01 \cdot n_r^{0.963}$	0.99	$0.808 \cdot n_r^{0.720}$	0.96	$0.661 \cdot n_r^{1.04}$	0.96	$0.992 \cdot n_r^{0.561}$	0.91
Total tertiary structure energy	$1.89 \cdot n_r^{0.912}$	0.99	$0.940 \cdot n_r^{0.657}$	0.97	$1.35 \cdot n_r^{0.966}$	0.99	$1.32 \cdot n_r^{0.614}$	0.98
Total conformational energy	$2.89 \cdot n_r^{0.932}$	1.00	$1.34 \cdot n_r^{0.682}$	0.98	$2.02 \cdot n_r^{0.993}$	0.99	$1.68 \cdot n_r^{0.601}$	0.98

[a]Equations estimated from linear fitting in log-log plots of energy versus sequence length; n_r is sequence length. These average energies for random threadings are positive, because the total average of interaction energies of each type over all native structures is calibrated at zero.
[b]Correlation coefficients in log-log plots of energy versus sequence length.

even the standard deviation of all energies for inverse threading are similar to those for normal threading. The largest difference is that the power dependence of the standard deviation of the secondary structure energy on sequence length is 0.56 closer to ½, as expected, for inverse threading than to the values of 0.72 found for normal threading. Then, the native energies in standard deviation units for inverse threading can be predicted by using those estimated relationships in a way similar to Eq. 21 for normal threading. The correlation between the predicted and observed values is highly significant with an excellent correlation coefficient of 0.96 and a slope of 0.94.

DISCUSSION

Contact energies were demonstrated to discriminate between native-like conformations and incorrectly folded conformations in five small lattice-proteins,[40] and also to be useful in ranking potential binding peptides to MHC molecules.[41] It was also demonstrated[19] that the tertiary potential function consisting of the contact energies and the repulsive packing energies could select native structures from non-native folds, which were generated by threading sequences into other structures in all possible ways without gaps. In the present paper, the same type of test has been carried out to see the effects of the secondary structure potentials on the discrimination of the native folds among non-native folds, and it has been shown that their inclusion can substantially improve the capability for discrimination for almost all of the proteins tested here. DeWitte and Shakhnovich[29] tried to optimize the relative weight for the contribution of local interactions to enhance fold or sequence recognition. According to their analyses, the optimum ratio of local interactions to contact interactions was one for inverse threading and 0.3 to 0.7 for threading. Here instead, we did not optimize the relative weight of secondary structure energies to tertiary structure energies for a scoring function, but rather they are summed up with an equal weight, assuming that both types of empirical energies properly reflect the actual strength of those interactions.

Because non-native folds have been generated here by simply threading a sequence into other non-native structures without gaps, subtracting the reference energy, which does not depend on protein conformation, from each potential does not change the values of the energies in s.d. units from their means in normal threading, although it does affect the absolute values of the energies; see Eqs. 11–15. Therefore, this threading check of the potentials can only test other corrections to the potentials such as the subtraction of collapse energies from contact energies and the removal of intrinsic and backbone-backbone interactions from the secondary structure potentials. In the present test, the energies of intrinsic residue and backbone-backbone interactions for secondary structures have been excluded. The exclusion of these energy terms is not required but is better at the present level in which only the effects of short-range interactions on secondary structures are taken into account. When other long-range effects such as the more complex hydrogen bond interactions between β strands and other interactions are properly included, these additional terms should be included in the estimation of secondary structure energies, even for fold recognition. On the other hand, subtracting the collapse energy from the estimation of the total contact energy is required to make it possible to compare energies among monomeric and multimeric proteins. However, all these energies must be taken into account to estimate conformational energies for simulations of protein conformations.

On the other hand, the adequacy of an average native structure as the reference state for the present energy potentials has been successfully tested by inverse threading, which clearly shows that native sequences can easily be identified for a given protein structure to have the lowest alignment energies among non-native sequences. But, most remarkably it has been shown that through the proper choice of a reference state, both threading and inverse threading can succeed with the same set of potentials.

An implicit assumption in the present scoring function for sequence recognition with a given structure is that native sequences are well designed to fold into specific native structures. With this basic assumption, a scoring function for sequence recognition becomes just the same as that for sequence design in the inverse folding problem in which a sequence is designed to better fold to a specific structure. Here, by assuming the stability of native structures as a primary requirement for a sequence to fold into a structure, the present scoring function in Eq. 11 has been devised. This scoring function that approximately measures the stability of a structure could also be used for sequence design. However, it should be noted that in principle native structures must be the lowest energy folds for their sequences but that native sequences need not be best designed for their native structures, even though it is highly probable; proteins might evolve toward more com-

patible sequences, if their sequences are not quite compatible with their structures.

CONCLUSION

Here, a simple test by threading sequences into structures without gaps has been used to show the suitability of the modifications to the original energy potentials for scoring in both fold and sequence recognition. The present case of forbidding gaps is essential in order to demonstrate separately the suitability of the reference state for inverse threading and other corrections to the energy potentials for normal threading. The result that native folds are discriminated from non-native folds for given sequences in normal threading justifies the subtraction of a collapse energy from contact energies. The adequacy of an average native structure as the reference state for both secondary structure potentials and tertiary structure potentials has been shown by the fact that native sequences can be identified by given structures in inverse threading. Thus, both threading and inverse threading can succeed with the same set of potentials. Also, it has been shown that the inclusion of the secondary structure potentials can substantially improve the quality of discrimination for almost all of the proteins tested here, expecially in inverse threading.

REFERENCES

1. Novotny J, Bruccoleri RE, Karplus M. An analysis of incorrectly folded protein models; implications for structure predictions. J Mol Biol 1984;177:787–818.
2. Novotny J, Rashin AA, Bruccoleri RE. Criteria that discriminate between native proteins and incorrectly folded models. Proteins 1988;4:19–30.
3. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. Nature 1986;319:199–203.
4. Hendlich M, Lackner P, Weitckus S, Floechner H, Froschauer R, Gottsbachner K, Casari G, Sippl MJ. Identification of native protein folds amongst a large number of incorrect models; the calculation of low energy conformations from potentials of mean force. J Mol Biol 1990;216:167–180.
5. Bowie JU, Lüthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
6. Sippl MJ, Weitckus S. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. Proteins 1992;13:258–271.
7. Jones DT, Taylor WR, Thornton JM. A New Approach to Protein Fold Recognition. Nature 1992;358:86–89.
8. Godzik A, Skolnick J. Sequence-structure matching in globular proteins: application to super-secondary and tertiary structure determination. Proc Natl Acad Sci USA 1992;89:12098–12102.
9. Nishikawa K, Matsuo Y. Development of pseudoenergy potentials for assessing protein 3D-1D compatibility and detecting weak homologies. Prot Eng 1993;6:811–820.
10. Kocher J-PA, Rooman MJ, Wodak SJ. Factors Influencing the Ability of Knowledge-based Potentials to Identify Native Sequence-Structure Matches. J Mol Biol 1994;235:1598–1613.
11. Matsuo Y, Nakamura H, Nishikawa K. Detection of protein 3D-1D compatibility characterized by the evaluation of side-chain packing and electrostatic interactions. J Biochem 1995;118:137–148.
12. Park B, Levitt M. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. J Mol Biol 1996; 258:367–392.
13. Park BH, Huang ES, Levitt M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. J Mol Biol 1997;266:831–846.
14. Miyazawa S, Jernigan RL. A Scoring function with structure-dependent gap penalties for identifying protein sequence-structure compatibilities. Submitted, 1999.
15. Moult J, Hubbard T, Bryant SH, Fidelis K, Pedersen T. A large-scale experiment to assess protein structure prediction methods. Proteins 1995;23:ii.
16. Moult J, Hubbard T, Bryant SH, Fidelis K, Pedersen T. Critical Assessment of Methods of Protein Structure Prediction (CASP): Round II. Proteins 1997;supplement 1:2-6.
17. Munson PJ, Singh RK. Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment. Prot Sci 1997;6:1467–1481.
18. Miyazawa S, Jernigan RL. Interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
19. 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.
20. Miyazawa S, Jernigan RL. Substitution mutants and the extent of local compactness in the denatured state. Prot Eng 1994;7:1209–1220.
21. Zhang C, Vasmatzis G, Cornette JL, DeLisi C. Determination of atomic desolvation energies from the structures of crystallized proteins. J Mol Biol 1997;267:707–726.
22. Thomas PD, Dill K. Statistical potentials extracted from protein structures: how accurate are they? J Mol Biol 1996;257:457–469.
23. Mirny LA, Shakhnovich EI. How to derive a protein folding potential? A new approach to an old problem. J Mol Biol 1996;264: 1164–1179.
24. 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.
25. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. J Mol Biol 1990;213:859–883.
26. Sippl MJ. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structure. J Comp-Aided Mol Design 1993;7:473–501.
27. Chothia C. One thousand families for the molecular biologist. Nature 1992;357:543–544.
28. Orengo CA, Jones DT, Thornton JM. Protein superfamilies and domain superfolds. Nature 1994;372:631–634.
29. DeWitte RS, Shakhnovich EI. Pseudodihedrals: simplified protein backbone representation with knowledge-based energy. Protein Sci 1994;3:1570–1581.
30. Miyazawa S, Jernigan RL. Evaluation of short-range interactions as secondary structure energies for protein fold and sequence recognition. Proteins 1999;36:347–356.
31. Pande VS, Grosberg AY, Tanaka T. Statistical mechanics of simple models of protein folding and design. Biophys J 1997;73:3192–3210.
32. Shakhnovich EI. Protein design: a perspective from simple tractable models. Folding & Design 1998;3:R45–R58.
33. 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.
34. Seno F, Vendruscolo M, Maritan A, Banavar J. Optimal protein design procedure. Phys Rev Lett 1996;77:1901–1904.
35. Deutsch JM, Kurosky T. New algorithm for protein design. Phys Rev Lett 1996;76:323–326.
36. Morrissey M, Shakhnovich EI. Design of proteins with selected thermal properties. Folding & Design 1996;1:391–406.
37. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The protein data bank: a computer-based archival file for macromolecular structures. J Mol Biol 1997;112:535–542.
38. Orengo CA, Flores TP, Taylor WR, Thornton JM. Identification and classification of protein fold families. Prot Eng 1993;6:485–500.
39. Godzik A, Koliński A, Skolnick J. Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. Protein Sci 1995;4:2107–2117.
40. Covell DG, Jernigan RL. Conformations of folded proteins in restricted spaces. Biochemistry 1990;29:3287–3294.
41. Altuvia Y, Schueler O, Margalit H. Ranking potential binding peptides to MHC molecules by a computational threading approach. J Mol Biol 1995;249:244–250.