Self-Consistent Estimation of Inter-Residue Protein Contact Energies Based on an Equilibrium Mixture Approximation of Residues

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 **Pairwise contact energies for 20 types of residues are estimated self-consistently from the actual observed frequencies of contacts with regression coefficients that are obtained by comparing "input" and predicted values with the Bethe approximation for the equilibrium mixtures of residues interacting. This is premised on the fact that correlations between the "input" and the predicted values are sufficiently high although the regression coefficients themselves can depend to some extent on protein structures as well as interaction strengths. Residue coordination numbers are optimized to obtain the best correlation between "input" and predicted values for the partition energies. The contact energies self-consistently estimated this way indicate that the partition energies predicted with the Bethe approximation should be reduced by a factor of about 0.3 and the intrinsic pairwise energies by a factor of about 0.6. The observed distribution of contacts can be approximated with a small relative error of only about 0.08 as an equilibrium mixture of residues, if many proteins were employed to collect more than 20,000 contacts. Including repulsive packing interactions and secondary structure interactions further reduces the relative errors. These new contact energies are demonstrated by threading to have improved their ability to discriminate native structures from other non-native folds. Proteins 1999;34:49–68.**
© 1999 Wiley-Liss, Inc.

Key words: empirical potential; quasi-chemical approximation; statistical potential; knowledge-based potential; protein fold recognition

INTRODUCTION

Simulating complete protein folding processes, occurring on a time scale ranging from milliseconds to seconds, would require enormous computational power because of the high dimensionality of the space of protein conformations and the complexity of their energy surfaces. If potentials with full atomic representations of proteins are used, present-day computer capabilities usually limit the time scale of molecular dynamics simulations to nanoseconds. Consequently simplified models are required at an appropriate level of coarse graining for complete descriptions of the energy landscape, from the denatured to the native state. Simplifications must be made consistently to both geometry and potential functions. Pairwise contact energies,[1–4] and pairwise potentials of mean force[5,6] are types of such simplified potential functions that have been used to provide a crude estimate of conformational energy at the residue level[7] or to distinguish between native and non-native folds.[8–13] Similarities and differences among pairwise interaction energies have been analyzed.[14] Pseudopotentials were devised to find out which amino acid sequences fold into a known three-dimensional structure.[15–17] An empirical method to evaluate the correctness of protein models was developed.[18] Statistical potentials extracted from known protein structures are also used to predict the docking of protein structures,[19] to predict protein binding,[20,21] and to simulate protein folding.[22–25]

To extract these statistical potentials[1–6] from protein structures, the pairwise density or higher order cluster of residues observed in protein native structures was approximated or assumed to obey the Boltzmann distribution with the corresponding interaction energy for the specific pairs or clusters. Boltzmann statistics was also assumed for the distribution of residues between the interior and exterior of protein molecules.[26] Extracted partition energies between interior and exterior of protein molecules correlate well with transfer experiments. For other properties, Boltzmann-like distributions are observed for the distributions of backbone and side-chain dihedral angles,[27–29] ion-pair substructures in proteins,[30] cis and trans conformations of proline residues,[31] and the sizes of empty cavities.[32] Thomas and Dill[33] pointed out that for many different protein structures the apparent temperature is of the same order of magnitude, lying between about 150 K and 600 K. They estimated that the temperatures of Boltzmann-like distributions for the interior-exterior partitioning of residues in protein structures range from 640 K to 1800 K, depending on the length, amino acid composition and compactness of the protein.

Thomas and Dill[33] also analyzed whether or not pairwise distributions of residues are Boltzmann-like in the unique,

*Correspondence to: Robert L. Jernigan, Room B-116, Building 12B, MSC 5677, Laboratory of Experimental and Computational Biology, DBS, National Cancer Institute, Bethesda, MD 28092-5677. E-mail: jernigan@structure.nci.nih.gov

Received 7 April 1998; Accepted 17 August 1998

lowest energy conformations of lattice proteins of the AB model, which consist of two monomer types A and B for short chains on a two-dimensional square lattice. 1) First, they assumed a set of true contact free energies. 2) For each chain length, they performed an exhaustive search of conformational space, and found the lowest energy states, the native states, for all sequences. 3) A "database" of the unique native structures was constructed. 4) From this database, contact potentials were extracted by two representative methods, the contact energy approach of Miyazawa and Jernigan[1] and the distance-dependent approach for pairwise potentials of mean force by Sippl.[5] 5) Then, these extracted potentials were compared with the original "input" contact energies, in order to assess how the derived statistical potentials reflect the true contact energies in real proteins. In the two residue, hydrophobic and polar, HP model, a simpler version of the AB model, the extracted contact energies are non-zero whereas "input" ones were zero, and these also showed unrealistic dependences on chain length or on the surface-to-volume ratio of proteins. The extracted energies also depended on the average partition propensity of the structures in the database of lattice proteins; the average partition propensity being defined as the total number of contacts divided by the total number of coordination sites of hydrophobic residues. Based on these results, they suggested that the Boltzmann distribution law may be inappropriate especially for converting pair frequencies of residues in protein structures into energies.

On the other hand, Mirny and Shakhnovich[34] pointed out that there were good correlations between "input" pairwise energies and contact energies extracted by the methods of Miyazawa and Jernigan[1] and Hinds and Levitt,[4] although the method of Miyazawa and Jernigan[1] yields extremely strong non-specific attractions between residues. Unlike Thomas and Dill,[33] they used $3 \times 3 \times 3$ cubic lattice proteins consisting of 20 types of residues, whose structures are the most compact forms, and whose sequences are designed to not only be the lowest energy form for thermodynamic stability, but also to have maximal $|Z|$ scores, that is, large energy gaps to ensure kinetic accessibility, for each structure.

Here, we analyze how well our method[1] for extracting contact energies from contact frequencies of residue pairs in protein structures reproduces "input" contact energies in protein structures. There are two questions: the first question is whether assumptions in our method are actually satisfied in real proteins, and the second question is how well can the Bethe approximation (quasi-chemical approximation) reproduce "input" contact energies when those assumptions are satisfied. One basic assumption in our method is that inter-residue contacts in a large enough sample can be regarded as an equilibrium ensemble of unconnected residues which interact with one another through pairwise potentials specific to their residue types.[1] This assumption will be examined directly by comparing the observed numbers of inter-residue contacts with those predicted with this assumption. The predicted numbers of contacts for each protein are calculated from the equilibrium mixtures of residues interacting with the estimated contact energies in a protein. The reproducibility of "input" contact energies by the Bethe approximation is analyzed by comparing "input" contact energies with those extracted from the equilibrium mixtures of residues. This way of testing the self-consistency of extracted potentials was proposed by both Thomas and Dill[33] and Mirny and Shakhnovich.[34]

The equilibrium mixtures of residues are generated in a Monte Carlo simulation by shuffling residues in the protein structures. Residues are shuffled in each protein structure, so that the total numbers of residue-to-residue contacts and residue-to-solvent contacts are fixed for each protein. Therefore, the effects of chain connectivity[1] imposing a limit to the size of the system, i.e., the total number of lattice sites or the number of effective solvent molecules, are not considered at all. In this case, the absolute values of contact energies cannot be estimated but only the alignment energies, i.e., the relative contact energy defined as contact energies less a collapse energy, $e_{ij} - e_{rr}$; see Materials and Methods section for details. To examine how the reproducibilities of "input" energies by the Bethe approximation depend on the model systems, amino acid mixtures on simple cubic lattices and on face-centered cubic lattices are also examined. Lattice calculations indicate that the reproducibility depends strongly on the lattice structures and interaction strengths. Simulations on protein structures reveal that the Bethe approximation overestimates the strengths of contact energies, especially of partition energies, $e_{ir} - e_{rr}$. However, simulations show that the correlation between "input" energies and predicted values is greater than 0.9 for most proteins. Because of these high correlations, the present simulations permit us to evaluate quantitatively how much the Bethe approximation overestimates contact energies and to correct our original estimates. When these corrected values of contact energies are taken iteratively as "input" energies, contact energies can be self-consistently estimated by the present method from the observed distribution of contacts.

The equilibrium distributions of contacts generated with the corrected contact energies are compared with what is actually observed in proteins. The relative errors between the equilibrium distributions and the observed ones decrease with about the -0.4 power of the total number of contacts, which is close to the power dependence, -0.5, to be expected a for random sample and converges to a value of about 0.08 for more than 20,000 contacts, indicating that the actual observed distributions of contacts may be approximated as equilibrium distributions of residue mixtures interacting with contact energies.

In addition, the effects of other interactions operative in protein structures, such as repulsive packing interactions,[2] secondary structure interactions,[13] and random noise to simulate all other interactions are also examined for the estimation of contact energies. The corrections for contact energies are evaluated in the presence of these interactions to yield a self-consistent set of values.

MATERIALS AND METHODS

What Interaction Energies Can Be Estimated for This System?

Let us consider the system of a protein surrounded by effective solvent molecules and assume that the neighboring pairs of residues of the i and j types interact with energies E_{ij}. The total energy of this system is

$$\text{Total energy} = \sum_{i=0}\sum_{j=0} E_{ij} n_{ij} \tag{1}$$

$$= \sum_{i=0} (2E_{i0} - E_{00}) q_i n_i / 2 + \sum_{i=1}\sum_{j=1} e_{ij} n_{ij} \tag{2}$$

where n_{ii} and $n_{ij} + n_{ji}$ are the numbers of contacts between i type residues and between i and j types of residues. Subscript 0 is used to represent effective solvent, and subscripts 1 through 20 are for residue types. Energies E_{ij} and the numbers of contacts n_{ij} are both symmetrical;

$$E_{ij} = E_{ji} \tag{3}$$

$$n_{ij} = n_{ji} \tag{4}$$

q_i is coordination number for an i type residue, and satisfies

$$\sum_{j=0} n_{ij} = \frac{q_i n_i}{2} \tag{5}$$

where n_i is the number of residues of type i. The contact energy e_{ij} is defined as

$$e_{ij} \equiv E_{ij} + E_{00} - E_{i0} - E_{0j} \tag{6}$$

$$= e_{ji} \tag{7}$$

It is clear that only the second term in Eq. 2 depends on the protein conformation, i.e. n_{ij}. Therefore, in the following parts of this paper, the phrase of “the total contact energy” is used to mean only the second term of this equation.

Then, let us assume that residues can be exchanged within a native protein structure. In the Bethe approximation, the partition function of this system is estimated as

$$Z = \text{const} \sum_{\{n_{ij}\}} \frac{n_{r0}!\, n_{0r}!\, n_{rr}!}{\prod_{i=1} n_{i0}! \prod_{j=1} n_{0j}! \prod_{i=1}\prod_{j=1} n_{ij}!} \exp\left(- \sum_{i=1}\sum_{j=1} e_{ij} n_{ij} \right) \tag{8}$$

where n_{r0} is the total number of residue-to-solvent contacts, and n_{rr} is the total number of contacts. Note that all energies here are taken to be dimensionless, i.e. in units of RT. Because only shuffling of residues within native protein structures is considered, the total number of contacts n_{rr} and then the total number of residue-to-solvent contacts n_{r0} are fixed for this system. Thus, constraints for this system are

$$\sum_{i=1}\sum_{j=1} n_{ij} = n_{rr} \tag{9}$$

$$\sum_{i=1} n_{i0} = n_{r0}. \tag{10}$$

$$\sum_{j=1} n_{0j} = n_{0r}. \tag{11}$$

By maximizing the partition function with respect to n_{ij} with these constraints and Eqs. 4 and 5, the statistical average, $\overline{n_{ij}}$, of n_{ij} is derived:

$$\frac{\overline{n_{ij}} n_{r0} n_{0r}}{n_{rr} \overline{n_{i0} n_{0j}}} = \exp(-\Delta e_{ij}) \tag{12}$$

$$\Delta e_{ij} \equiv e_{ij} - e_{rr}. \tag{13}$$

A constant e_{rr} that is called the collapse energy is defined as

$$\exp(-e_{rr}) \equiv \left[\frac{\sum_{i=1}\sum_{j=1} \overline{n_{ij}} \exp(e_{ij})}{n_{rr}} \right]^{-1} \tag{14}$$

$$= \frac{\sum_{i=1}\sum_{j=1} \overline{n_{i0} n_{0j}} \exp(-e_{ij})}{n_{r0} n_{0r}}. \tag{15}$$

Eq. 12 indicates that applying the Bethe approximation to this system cannot provide estimates of contact energies e_{ij} but only of relative contact energies Δe_{ij} ($\equiv e_{ij} - e_{rr}$). This is physically reasonable, since structures are fixed in the native state.

The partition energy or hydrophobic energy, e_{ir}, which is defined as

$$\exp(-e_{ir}) \equiv \left[\frac{\sum_{j=1} \overline{n_{ij}} \exp(e_{ij})}{\overline{n_{ir}}} \right]^{-1} \tag{16}$$

$$= \frac{\sum_{j=1} \overline{n_{0j}} \exp(-e_{ij})}{n_{0r}} \tag{17}$$

is related to statistical averages of the numbers of contacts as follows:

$$\overline{n_{i0}} \Big/ \left[\frac{\overline{n_{ir} n_{r0}}}{n_{rr}} \right] = \exp(\Delta e_{ir}) \tag{18}$$

$$\Delta e_{ir} \equiv e_{ir} - e_{rr} \tag{19}$$

where n_{ir} is the sum of n_{ij} over all types of amino acids

$$n_{ir} \equiv \sum_{j=1} n_{ij} \tag{20}$$

Δe_{ir} may be called the relative partition energy or relative hydrophobic energy.

Also, intrinsic inter-residue interaction energies, δe_{ij}, between 20 types of residues, which do not include hydrophobic energies, are defined as

$$\delta e_{ij} \equiv \Delta e_{ij} - \Delta e_{ir} - \Delta e_{rj} = e_{ij} + e_{rr} - e_{ir} - e_{rj} \qquad (21)$$

and satisfy the following equation

$$\overline{n_{ij}} \Big/ \left[\frac{\overline{n_{ir} n_{rj}}}{n_{rr}} \right] = \exp(-\delta e_{ij}). \qquad (22)$$

The above equation means that the ratio of the observed number of i–j contacts in protein structures to their expected number for random mixing with fixed n_{ir} and n_{rj} is equal to the Boltzmann factor of their intrinsic inter-residue energy. These energies are similar to contact energies used by Park and Levitt; their way to evaluate the expected number is different from Eq. 22.[35] Similarities and differences between Sippl's potentials of mean force[5] and contact energies are described in detail in the Appendix.

Here, intrinsic pairwise energies and relative partition energies predicted by the Bethe approximation are calculated as follows from the observed frequencies of contacts in protein structures or from the equilibrium frequencies of residue pairs in contact for the same set of proteins. First, residues in protein structures are represented by single points at the centers of their side chain atom positions; the positions of C^{α} atoms are used for glycines. Residues whose centers are closer than R_c are defined to be in contact. The limiting value $R_c = 6.5$ Å for contacts was chosen on the basis of the occurrence of the first peak in the radial distribution of residues in the interior of proteins.[1] The coordination number for each type of residue was estimated from residue volumes and the volume of the first shell.[1,2] The numbers of residue-to-solvent contacts are evaluated from Eq. 5 with these estimated values of coordination numbers. Then, those energies are calculated from

$$\exp(-\delta e_{ij}) = \frac{N_{ij}}{C_{ij}} \qquad (23)$$

$$\exp(\Delta e_{ir}) = \frac{N_{i0}}{C_{i0}} \qquad (24)$$

where N_{ij} is defined as the total number of i–j contacts in all protein structures used, and C_{ij} is its expected value for random mixing.

$$N_{ij} \equiv \sum_{p} \overline{n_{ij;p}} \qquad (25)$$

$$C_{ij} \equiv \sum_{p} \frac{1}{2} \left[\frac{\overline{n_{ir;p} n_{rj;p}}}{n_{rr;p}} \cdot \frac{1 - \delta_{ij}/n_{i;p}}{1 - \overline{n_{ri;p}}/n_{rr;p}/n_{i;p}} + \frac{\overline{n_{jr;p} n_{ri;p}}}{n_{rr;p}} \cdot \frac{1 - \delta_{ji}/n_{j;p}}{1 - \overline{n_{rj;p}}/n_{rr;p}/n_{j;p}} \right] \qquad (26)$$

where p indicates the pth protein, and δ_{ij} is the Kronecker δ. The definition of C_{ij}, Eq. 26, includes a correction for infrequent residues, which was missing in the original definition.[1] Eq. 24 to estimate the partition energy is different from the one in Miyazawa and Jernigan,[1,2] although their estimated values are only slightly different.

Then, relative contact energies Δe_{ij} are calculated from intrinsic pairwise energies and relative partition energies as follows:

$$\Delta e_{ij} = \Delta e_{ir} + \Delta e_{rj} + \delta e_{ij}. \qquad (27)$$

The relative contact energies Δe_{ij} calculated from the observed frequencies of contacts in proteins with Eqs. 23, 24, and 27 are the initial estimates of contact energies in the present iterative procedure.

Iterative Procedure to Self-Consistently Estimate Contact Energies

The number of relative contact energies for all amino acid pairs that must be determined is 210. However, as presented in the Results section later, the results for lattice monomers indicate that the correlations between real energies and predicted ones with the Bethe approximation described above are high for both the partition energies and intrinsic pairwise energies in the actual range of interaction strength. If these correlation coefficients are high enough, one may reduce the number of parameters to obtain, from 210 to two regression coefficients for the partition energies and intrinsic pairwise energies by utilizing values predicted with the Bethe approximation. In other words, contact energies are estimated from values calculated from the actual distribution of contacts in proteins with the Bethe approximation and the regression coefficients between "input" and predicted energies. Because these regression coefficients depend significantly on interaction strength and on structure, they are self-consistently calculated in an iterative procedure in which the equilibrium distribution of contacts in each protein is generated with the previous estimates of contact energies by a Monte Carlo simulation, and better estimates of the regression coefficients are calculated.

The iterative procedure is:

1. Assume the relative contact energies Δe_{ij} calculated from the observed frequencies of contacts in protein structures with Eqs. 23, 24, and 27 as "input" energies.
2. Perform a Monte Carlo simulation with the Metropolis method[36] for each protein in which residues in a protein are assumed to interact with each other with the pairwise contact energies and are shuffled to obtain an equilibrium distribution of contacts. It should be noted here that the equilibrium distribution of contacts in a protein does not depend on the collapse energy, because protein structures (n_{r0} and n_{rr}) are fixed and residues are shuffled only within a protein; see Eq. 12. Relative temperature T_{rel} is taken to be one. In some cases, repulsive packing energies and secondary structure energies[2,13] are also taken into account to generate an

equilibrium mixture of residues; see the sub-section "Conformational Energy."

3. Calculate predicted contact energies by using the Bethe approximation from the total equilibrium distribution of contacts; see Eqs. 23, 24, and 27.
4. Calculate regression coefficients of the "input" versus "predicted" values for both types of energies, $\Delta e_{ir}(\equiv e_{ir} - e_{rr})$ and $\delta e_{ij}(\equiv e_{ij} + e_{rr} - e_{ir} - e_{rj})$ by comparing the predicted and the "input" energies.
5. Calculate better estimates of real contact energies by using the new estimates of the regression coefficients obtained in the simulation together with the predicted energies with the Bethe approximation from the numbers of contacts actually observed in protein structures.
6. If the regression coefficients do not indicate a sufficiently good match with those values used to estimate contact energies for this iteration, repeat steps from 2 through 6 again by updating the "input" energies. Otherwise the procedure is completed and yields newly estimated energies.

In the Bethe approximation, the intrinsic inter-residue energies are estimated without any adjustable parameter from the ratios of the observed numbers of contacts and their expected numbers in random mixing; see Eq. 23. On the other hand, the partition energies are estimated from the ratios of the observed numbers of residue-to-solvent contacts and their expected numbers in random mixing; see Eq. 24. The numbers of residue-to-solvent contacts are evaluated with Eq. 5 from coordination numbers and the numbers of residue-to-residue contacts. The coordination number for each type of residue was estimated from residue volumes and the volume of the first shell.[2] An improper evaluation of the coordination numbers can affect the correlation for the partition energies. Even though the correct numbers are used for the coordination numbers, the Bethe approximation certainly limits the correlation for the partition energies as seen in the results of lattice simulations. Apart from the physical meaning of coordination numbers, they may be treated here as adjustable parameters to obtain a better correlation for the partition energies. In this case, step 4 is replaced by the following procedure.

4′. Calculate the optimal values for coordination numbers to yield the best correlation for the partition energies, and the regression coefficients of "input" compared with "predicted" values for both types of energies, Δe_{ir} and δe_{ij}. These coordination numbers and regression coefficients are then used in the next iteration.

Also step 6 is replaced by:

6′. Coordination numbers and regression coefficients are both checked to see whether they indicate a sufficiently good match with values used to estimate contact energies for this iteration, and then the whole procedure of steps from 2 through 6′ is repeated until such a convergence is attained.

Now, both coordination numbers and regression coefficients are self-consistently calculated in the iterative procedure, and then contact energies are finally estimated with them from the actual distribution of contacts in proteins.

How to Estimate Contact Energies With Regression Coefficients

The estimates, ϵ_{ij}, of contact energies are calculated from values obtained with the Bethe approximation and the estimates of regression coefficients for "input" versus predicted values. Let us define regression coefficients between "input" and predicted energies with the Bethe approximation for the equilibrium distribution of contacts in Monte Carlo simulations as α, β and η for partition energies, intrinsic inter-residue energies, and relative contact energies, respectively; that is

$$\Delta e_{ir}^{\text{input}} \sim \alpha \cdot \Delta e_{ir}^{\text{pred}} + \text{constant} \qquad (28)$$

$$\delta e_{ij}^{\text{input}} \sim \beta \cdot \delta e_{ij}^{\text{pred}} + \text{constant} \qquad (29)$$

$$\Delta e_{ij}^{\text{input}} \sim \eta \cdot [\alpha \cdot \Delta e_{ir}^{\text{pred}} + \alpha \cdot \Delta e_{rj}^{\text{pred}} + \beta \cdot \delta e_{ij}^{\text{pred}}] + \text{constant} \qquad (30)$$

Then, the estimates of relative contact energies, $\Delta\epsilon_{ij}(\equiv \epsilon_{ij} - \epsilon_{rr})$, are calculated from those regression coefficients and the observed distribution of contacts in all proteins.

$$\Delta\epsilon_{ir} = \alpha' \cdot \Delta e_{ir}^{\text{obs}} \qquad (31)$$

$$\delta\epsilon_{ij} = \beta' \cdot \delta e_{ij}^{\text{obs}} \qquad (32)$$

$$\Delta\epsilon_{ij} = \Delta\epsilon_{ir} + \Delta\epsilon_{rj} + \delta\epsilon_{ij} \qquad (33)$$

$$\alpha' \equiv \eta \cdot \alpha \qquad \beta' \equiv \eta \cdot \beta \qquad (34)$$

where e^{obs} means energy values calculated from the observed numbers of contacts in proteins with Eqs. 23, 24. Here it should be noted that $\Delta e_{ir}^{\text{obs}}$ does depend on the values of coordination numbers q_i that are dealt with as adjustable parameters but $\delta e_{ij}^{\text{obs}}$ does not; see Eqs. 5, 23, 24.

The regression coefficients, α' and β', and coordination numbers if adjusted are self-consistently calculated. That is, until they converge, another iteration is performed by assuming $\Delta\epsilon_{ij}$ as the "input" contact energies e_{ij}^{input}; the collapse energy ϵ_{rr} is unknown, and taken as zero here, because protein structures (n_{r0} and n_{rr}) are fixed and residues are shuffled in a protein, and thus the equilibrium distribution of contacts in a protein does not depend on this constant energy; see Eq. 12.

Coordination Numbers as Adjustable Parameters

Coordination numbers are adjusted to yield the best correlation between "input" and predicted values for the relative partition energies:

$$\frac{1}{2}\, q_i^{\mathrm{adj}} N_i = N_{ir}^{\mathrm{equil}} + N_{i0}^{\mathrm{adj}} \tag{35}$$

$$N_{i0}^{\mathrm{adj}} = \gamma \cdot C_{i0}^{\mathrm{equil}} \exp\,(e_{ir}^{\mathrm{adj}} - e_{rr}^{\mathrm{adj}}). \tag{36}$$

The superscript "equil" means those numbers in the equilibrium distributions of contacts generated in Monte Carlo simulations. The constant γ has been chosen, so that the total number of residue–solvent contacts, N_{r0}, is kept constant; that is, the average value of coordination number is kept constant;

$$N_{r0}^{\mathrm{adj}} = N_{r0} \tag{37}$$

The adjusted values, e_{ir}^{adj}, for partition energies are calculated from the "input" potential, e_{ij}^{input}, assumed for the system:

$$e_{ir}^{\mathrm{adj}} = e_{ir}^{input}/\alpha \tag{38}$$

$$\exp\,(-e_{rr}^{\mathrm{adj}}) = \left[\frac{\sum\limits_{i=1} N_{ir}^{\mathrm{equil}} \exp\,(e_{ir}^{\mathrm{adj}})}{N_{rr}}\right]^{-1} \tag{39}$$

$$\exp\,(-e_{ir}^{\mathrm{input}}) = \left[\frac{\sum\limits_{j=1} N_{ij}^{\mathrm{equil}} \exp\,(e_{ij}^{\mathrm{input}})}{N_{ir}^{\mathrm{equil}}}\right]^{-1} \tag{40}$$

where the constant α is the ratio of "input" relative to predicted values of partition energies. Adjusted coordination numbers yield an improved correlation for the partition energies between "input" and predicted values. Here, coordination numbers are taken as adjustable parameters regardless of their intrinsic physical meaning. A value for α is chosen by minimizing the mean square deviation of the optimum values, q_i^{adj}, from the original values, q_i, of the coordination numbers.

Conformational Energy

The conformational energy of a protein is divided into two terms: secondary structure energies and tertiary structure energies.

$$E^{\mathrm{conf}} \equiv E^{\mathrm{sec}} + E^{\mathrm{tert}} \tag{41}$$

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, and termed long-range interaction energies:[2]

$$E_p^{\mathrm{tert}} = E_p^c + E_p^r \tag{42}$$

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 40 in Miyazawa and Jernigan; even when coordination numbers are treated as adjustable parameters in the iterative procedure, the original values of coordination numbers are used for the calculation of repulsive packing energy.[2] The total contact energies and the total repulsive energies are calculated as the sums of these energies over all residues. Alternatively, the total contact energies can be calculated by simply summing contact energies for all contact residue pairs,$\Sigma_{ij} n_{ij} \epsilon_{ij}$.

The secondary structure energies, which include intrinsic preferences, backbone–backbone interactions, and backbone–side chain interactions, were estimated[13] on the basis of short-range interactions, ignoring the effects of long-range interactions, from the observed frequencies of secondary structures by assuming Boltzmann statistics; interactions between side chains could not be evaluated for these cases because of the limited data. These potentials are used here directly. In this paper, because protein structures are fixed at their native structures and residues are shuffled, intrinsic energies and backbone–backbone interaction energies remain constant and do not affect the results.

RESULTS

First, preliminary calculations have been performed to see how similar the contact frequencies at equilibrium are to the observed ones, when the contact energies estimated in Miyazawa and Jernigan[2] are used as "input" energies. The equilibrium distributions of contacts are simulated by shuffling residues in individual protein structures according to the Metropolis method[36] in a Monte Carlo simulation.

Table I lists relative errors in the equilibrium distribution for the numbers of residues that are within 6.5Å of a central residue and defined to be in contact with the residue, $|\Delta N(i, n^c)|/|N(i, n^c)|$, and the relative error in the equilibrium frequencies of residue pairs in contact, $|\Delta N_{ij}|/|N_{ij}|$, for each type of residue pair, defined by

$$\frac{|\Delta N(i, n^c)|}{|N(i, n^c)|} \equiv \frac{\left[\sum\limits_i \sum\limits_{n^c} (N^{\mathrm{equil}}(i, n^c) - N^{\mathrm{obs}}(i, n^c))^2\right]^{1/2}}{\left[\sum\limits_i \sum\limits_{n^c} N^{\mathrm{obs}}(i, n^c)^2\right]^{1/2}} \tag{43}$$

$$\frac{|\Delta N_{ij}|}{|N_{ij}|} \equiv \frac{\left[\sum\limits_i \sum\limits_{j>i} (2N_{ij}^{\mathrm{equil}} - 2N_{ij}^{\mathrm{obs}})^2 + \sum\limits_i (N_{ii}^{\mathrm{equil}} - N_{ii}^{\mathrm{obs}})^2\right]^{1/2}}{\left[\sum\limits_i \sum\limits_{j>i} (2N_{ij}^{\mathrm{obs}})^2 + \sum\limits_i (N_{ii}^{\mathrm{obs}})^2\right]^{1/2}} \tag{44}$$

where $N(i, n^c)$ represents the number of i type residues in contact with n^c, the number of residues. Superscripts "equil" and "obs" of $N(i, n^c)$ are for the equilibrium and observed distributions, respectively. Proteins[37] employed here are the set of 86 non-homologous, monomeric proteins used previously to assess the contact energies.[2] Calcula-

TABLE I. Characteristics of the Equilibrium Distributions of Contacts in Protein Structures Generated With the Original Contact Energies e_{ij} of Miyazawa and Jernigan (1996)[2] Scaled by Relative Temperature, T_{rel}[†]

	Relative errors		$\delta e_{ij}(\equiv e_{ij} + e_{rr} - e_{ir} - e_{rj})$		$\Delta e_{ir}(\equiv e_{ir} - e_{rr})$		
T_{rel}	$\|\Delta N(i, n^c)\|/\|N(i, n^c)\|$	$\|\Delta N_{ij}\|/\|N_{ij}\|$	Correlation coefficient	Regression coefficient, β	Correlation coefficient	Regression coefficient, α	Number of residue types*
1.0	0.45	0.37	0.83	0.72	0.95	0.52	15
2.0	0.27	0.24	0.79	0.56	0.95	0.30	17
3.0	0.18	0.16	0.81	0.50	0.97	0.29	17
4.0	0.14	0.16	0.85	0.49	0.93	0.20	18
5.0	0.14	0.20	0.88	0.50	0.87	0.12	20

[†]The correlation coefficients and regression coefficients above are those between "input" energies scaled by the relative temperature and values predicted with the Bethe approximation. The original values of coordination numbers are used to predict partition energies (Δe_{ir}). Here 86 non-homologous monomeric proteins used in the original work are employed with equal weights.

*The number of residue types for which Δe_{ir} can be estimated; they are not all estimated because the number of residue-solvent contacts becomes erroneously negative for strongly hydrophobic residues with these interaction strengths.

tions are performed at relative temperatures, T_{rel}, from 1 to 5.

Contrary to expectation, the relative errors are quite large at $T_{rel} = 1$, and their minimum values are located at much higher temperatures; the relative error has its minimum at $T_{rel} \sim 3.5$ for the frequency distribution of residue pairs in contact and at the higher temperature, $T_{rel} \sim 4.5$, for the distribution of the number of residues in contact. This means that there are deficiencies in either the basic assumption of the statistical equilibrium for inter-residue contacts with respect to their contact energies or in the reproducibility of "input" energies by the Bethe approximation. The fact that the relative errors take small values at their minima suggests that this result might be caused by an over-estimation of contact energies rather than the assumptions about the distribution of contacts.

The predicted contact energies from the use of the Bethe approximation are compared to those "input." The comparisons are made for the intrinsic pairwise energies, $\delta e_{ij}(\equiv e_{ij} + e_{rr} - e_{ir} - e_{rj})$, and also for the relative hydrophobic, partition energies, $\Delta e_{ir}(\equiv e_{ir} - e_{rr})$; see the Materials and Methods section for the definitions of these energies. The regression coefficients of the "input" energies divided by T_{rel} compared to the predicted values are also shown in Table I. The "input" values of the intrinsic pairwise energies δe_{ij} are smaller by about a factor 0.49 to 0.72 compared with the predicted values over the temperature range examined. By contrast, the partition energies Δe_{ir} are exaggeratedly over-estimated; the "input" energies are smaller by about a factor 0.12 to 0.52 compared to the predicted values. Here it should be noted that all values of Δe_{ir} cannot be estimated because the number of residue-solvent contacts becomes erroneously negative for strongly hydrophobic residues; the number of residue types for which Δe_{ir} could be estimated are listed in the last column of Table I. On the other hand, even though the absolute values of these predicted energies are quite different from the "input" ones, the correlation coefficients between them are rather good; these are mostly larger than 0.8 for δe_{ij}, and 0.9 for Δe_{ir}.

These facts strongly indicate that the large differences between the predicted and observed distributions of the number of residues in contact and of residue pairs in contact result from an over-estimation of contact energies rather than from the basic assumption of the Boltzmann equilibrium for contact formation.

Reproducibility of Contact Energies by the Bethe Approximation, for Amino Acid Mixtures on Simple Cubic and Face-Centered Cubic Lattices

In order to examine how accurately the Bethe approximation can reproduce "input" contact energies, we consider a mixture of 20 types of amino acid monomers on simple cubic lattices and also on face-centered cubic lattices, in which only nearest-neighbor interactions are assumed. Amino acid pairs occupying nearest-neighbor sites on lattice are assumed to interact with preassumed values of contact energies. First, lattice sites that are occupied by amino acids are determined. The most compact, dense systems, i.e. n by n by n for a simple cubic lattice, are examined. Less dense systems, which are generated by randomly removing amino acids from the most compact systems, were also examined but are not presented here. Second, amino acids are randomly chosen according to the average composition of amino acids in proteins used in Miyazawa and Jernigan.[2] Third, the equilibrium of an amino acid mixture is generated by shuffling amino acids among lattice sites according to the Metropolis method in a Monte Carlo simulation based on the input energies. Finally, contact energies predicted by the Bethe approximation are compared with the true input ones.

Monte Carlo simulations for both types of lattices are carried out for various lengths, from about 60 to 700, of sequences and at relative temperatures ranging from $T_{rel} = 1.0$ to 5.0. When input energies are the contact energies e_{ij} estimated by Miyazawa & Jernigan,[2] the correlation coefficients between predicted and "input" values of the partition energies Δe_{ir} are mostly better than 0.99 in the present temperature range for both types of lattice. Regression coefficients for the partition energies depend strongly on lattice structures, the size of molecules, and temperature, that is, the interaction strength. Those regression coefficients range from 1.87 to 0.87 at $T_{rel} = 1$ and from 1.33 to 1.01 at $T_{rel} = 5$ for shorter to longer sequences, respectively, on simple cubic lattices, and from 0.68 to 0.30 at $T_{rel} = 1$ and from 0.38 to 0.28 at $T_{rel} = 5$ for face-centered cubic lattices. However, there is a general trend that if more amino acids are included, i.e., the smaller the surface-to-volume ratio is, then the more the partition energies Δe_{ir} are over-estimated; the surface-to-volume ratio is measured here as the ratio of the number of residue–solvent contacts to the number of feasible contacts, $n_{r0}/(q_r n_r/2)$; $n_r \equiv \sum_{i=1} n_i$ and $q_r n_r \equiv \sum_{i=1} q_i n_i$. Such a dependence of the partition energies on chain length or the surface-to-volume ratio in the Bethe approximation was pointed out.[33] As stated in our original paper,[1] this is a limitation of the Bethe approximation.

One of the interesting features is that unlike the systems on simple cubic lattices, the partition energies are over-estimated in face-centered cubic lattices to a similar degree as those for protein structures. The coordination numbers for protein structures are estimated as 6.28 on average,[2] and these are used in Table I. Thus, the average coordination number for protein structures is closer to that for a simple cubic lattice, 6, than to the value for a face-centered cubic lattice, 12. However, the number of residues in contact with a residue in protein structures within a sphere of 6.5 Å radius can sometimes reach 10 or 11. It may be reasonable that the estimate of partition energies depends on the feasible range of contact numbers in structures.

For the original contact energies, correlations between predicted and "input" energies are better for the partition energies than for the intrinsic pairwise energies. The correlation coefficients for the partition energies are 0.99 at $T_{rel} = 1$ and 1.00 at $T_{rel} = 5$ for any length of sequence on simple cubic lattices, and range from 0.92 to 0.99 at $T_{rel} = 1$ for shorter to longer sequences and are 1.00 at $T_{rel} = 5$ for face-centered cubic lattices. On the other hand, the correlation coefficients for the intrinsic pairwise energies range from 0.91 to 0.68 at $T_{rel} = 1$ and from 0.97 to 0.98 at $T_{rel} = 5$ for shorter to longer sequences on simple cubic lattices, and from 0.90 to 0.63 at $T_{rel} = 1$ and from 0.89 to 0.82 at $T_{rel} = 5$ for face-centered cubic lattices. This is probably because the "input" energies consist of strong partition energies, Δe_{ir}, compared to the intrinsic pairwise energies, δe_{ij}. The low correlation for the intrinsic pairwise energies is more pronounced for face-centered cubic lattices in which the net interaction is much stronger than in a simple cubic lattice, because the face-centered cubic lattice has twice as many nearest neighbor sites as the simple cubic lattice. It is confirmed for both types of lattice that the correlation improves as the partition energies become weaker in comparison with the intrinsic pairwise energies.

The values of the regression coefficients for the intrinsic pairwise energies range from 0.81 to 0.50 at $T_{rel} = 1$ and from 0.81 to 0.82 at $T_{rel} = 5$ for shorter to longer sequences on simple cubic lattices, and from 0.95 to 0.47 at $T_{rel} = 1$ and from 0.54 to 0.38 at $T_{rel} = 5$ for face-centered cubic lattices. Thus, the values of the regression coefficients for face-centered cubic lattices also resemble those for protein structures listed in Table I, but this feature is not clear here because of the low correlation between predicted and true energies. However, this feature is confirmed with better correlations by employing more realistic values for the contact energies; this is shown later.

Estimation of Real Energies From Energies Calculated With the Bethe Approximation

The results for lattice monomers described in the preceding section indicate that: 1) the correlations between predicted and "input" energies are high for both the partition energies and intrinsic pairwise energies in the actual range of interaction strength, $T_{rel} \sim 4$ for the original contact energies e_{ij},[2] but 2) the regression coefficients between them vary significantly depending on interaction strength and on structures, i.e., the topology of residue positions in space, and on the number of residues. Because the correlation coefficients are high enough, the "input" energies can reliably be estimated from values calculated in the Bethe approximation with regression coefficients, separately for partition energies and intrinsic pairwise energies. Because these regression coefficients significantly depend on interaction strength and on structure, those values for estimating real contact energies from the actual distribution of contacts in proteins must be self-consistently calculated by an iterative procedure in which the equilibrium distribution of contacts in each protein is generated with the previous estimates of contact energies in a Monte Carlo simulation, and better estimates of the regression coefficients and then of contact energies are calculated; see the description of the iterative procedure in the Materials and Methods section.

The row of Method-A in Table II shows the correlation coefficients and regression coefficients for both types of energies calculated using the above procedure; the coordination numbers q_i used in Method-A are the original values estimated from residue volumes in Miyazawa and Jernigan.[2] In this table, all protein structures used in Miyazawa and Jernigan[2] are employed; the effective number of non-homologous proteins is 251. The correlation coefficients are sufficiently high, larger than 0.96, for both types of energies. The correlation for the intrinsic inter-residue energies (δe_{ij}) is better in this case than in Table I, because the "input" partition energies (Δe_{ir}) are much weaker here in comparison with the intrinsic inter-residue energies; the regression coefficient is much smaller for the partition energies than for the intrinsic inter-residue energies. The change in "input" energies also causes differences in the regression coefficients compared to Table I. The relative error in the equilibrium distributions of contacts compared to the actual distributions falls below 0.1. The "input" and predicted energies with the Bethe approximation for partition energies are compared in Figure 1. The solid line in the figure shows a regression line. The relationship between the "input" and predicted values for the partition energies is slightly concave instead of being precisely linear.

Table II also shows the results where the coordination numbers are adjusted to yield the best correlation; both the coordination numbers and regression coefficients are self-consistently calculated in these iterative procedures.

TABLE II. Correlations Between "Input" Contact Energies and Values Calculated With the Bethe Approximation From the Equilibrium Distributions of Contacts, and Relative Errors in the Equilibrium Distributions of Contacts From the Actual Observed Ones[†]

	Relative errors				$\Delta e_{ir}(\equiv e_{ir} - e_{rr})$		$\delta e_{ij}(\equiv e_{ij} + e_{rr} - e_{ir} - e_{rj})$		$\Delta e_{ij}(\equiv e_{ij} - e_{rr})$	
Method	$\lvert\Delta N(i, n^c)\rvert/\lvert N(i, n^c)\rvert$ and χ^2 with number of degrees of freedom		$\lvert\Delta N_{ij}\rvert/\lvert N_{ij}\rvert$ and χ^2 with number of degrees of freedom		Correlation coefficient	Regression coefficient, α	Correlation coefficient	Regression coefficient, β	Correlation coefficient	Regression coefficient, η
Method-A[a]	0.116	1374 (190)	0.0884	1359 (209)	0.970	0.173	0.962	0.687	0.960	0.943
Method-B[b]	0.0757	717 (171)	0.0645	752 (209)	1.000	0.305	0.947	0.643	0.988	0.930
Method-C[c]	0.0707	497 (171)	0.0563	589 (209)	1.000	0.313	0.959	0.709	0.990	0.948
Method-D[d]	0.0646	405 (171)	0.0698	890 (209)	1.000	0.280	0.942	0.612	0.985	0.938

[†] The "input" contact energies are ones estimated self-consistently from energies calculated with the Bethe approximation from the observed numbers of contacts, and from the regression coefficients between "input" and predicted values. The same set of proteins and sampling weights as in Miyazawa and Jernigan[2] are used; the effective numbers of proteins, residues and contacts are 251, 54,617, and 114,350, respectively.

[a] q_i are fixed at the original values in Miyazawa and Jernigan.[2] Interaction energies consist of contact energies only.

[b] q_i are optimized, and interaction energies consist of contact energies only.

[c] q_i are optimized, and repulsive interaction energies estimated in Miyazawa and Jernigan[2] are included in addition to contact energies.

[d] q_i are optimized, and repulsive energies and secondary structure energies, both of which were estimated in Miyazawa and Jernigan,[2,13] are taken into account as well as contact energies.

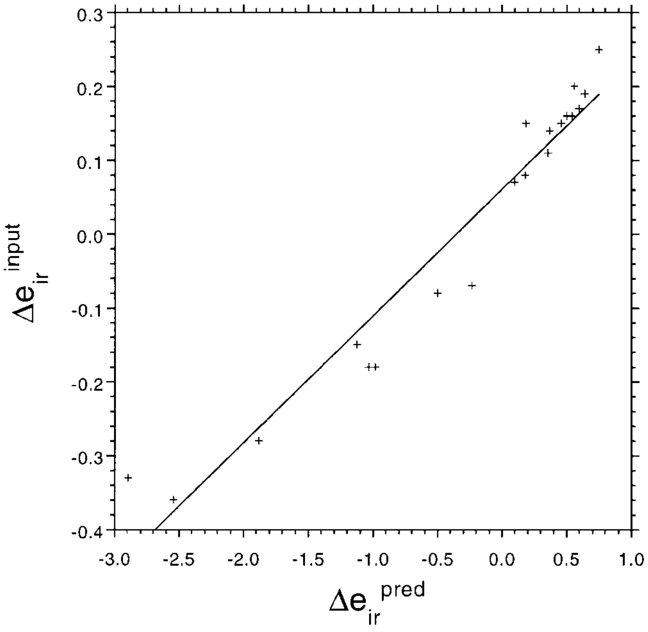

Fig. 1. Comparison of "input" and predicted energies with the Bethe approximation for relative partition energies, Δe_{ir} ($\equiv e_{ir} - e_{rr}$). The original values[2] estimated from residue volumes are used for coordination numbers for 20 types of residues. "Input" energies used here are contact energies estimated self-consistently from the observed numbers of contacts and from the regression coefficient between "input" values and values calculated with the Bethe approximation. Only contact energies are taken into account here; this interaction scheme corresponds to Method-A in Table II. The solid line is the regression line.

The adjusted values for the coordination numbers and the estimated values of the partition energies with the regression coefficient from the observed frequencies of contacts are listed in Table III; the original predictions with the Bethe approximation are listed as "Original" in Table III. Method-B corresponds to a case in which interactions among residues consist of pairwise contact energies only. In Method-C, repulsive packing energies are taken into account in addition to the contact energies. In Method-D, the total interaction energies consist of contact energies, repulsive packing energies, and secondary structure energies. Repulsive packing potentials and secondary structure potentials are estimated from the observed distributions of the numbers of residues in contact with a residue,[2] or from the observed frequencies of secondary structures in protein structures,[13] by assuming the Boltzmann distribution.

In order to compare these characteristics with those for lattice proteins, correlation and regression coefficients between "input" and predicted values are calculated for both simple cubic and face-centered cubic lattices with the contact energies newly estimated in Method-B; their values are listed in Table IV. It should be noted that the value of the coordination number in these calculations is fixed at the actual value, 6 or 12, for simple and face-centered cubic lattices, respectively.

An interesting observation is the fact that adjusting the coordination numbers to obtain the best correlation for partition energies leads to a decrease in the relative errors in the equilibrium frequencies of contacts compared with the observed ones, thereby providing a justification for the adjustment of the coordination numbers; compare the results of Method-A and Method-B in Table II. A general trend in the changes of the coordination numbers is that they increase for non-polar residues and decrease for polar residues, with only a few exceptions. However, it should be noted that the adjustment to the coordination numbers in Methods-B/C/D to increase the correlation between "input" and predicted partition energies are artificial and such changes of the coordination numbers do not have an actual physical basis. This is indicated by the fact that the changes of the coordination numbers from their original values in Method-B correlated well with the adjustments of the coordination numbers to obtain the best correlation in lattice monomers; the correlation coefficient of the deviations of the optimized q_i from their physical values is about 0.85 for the pair of Method-B of Table III-A for

TABLE III. The Self-Consistently Calculated Values of (A) Coordination Numbers q_i and (B) Relative Partition Energies $\Delta\epsilon_{ir}$ ($\equiv\epsilon_{ir} - \epsilon_{rr}$) in RT Units

(A)	CYS	MET	PHE	ILE	LEU	VAL	TRP	TYR	ALA	GLY	THR	SER	GLN	ASN	GLU	ASP	HIS	ARG	LYS	PRO
Original†																				
Method-A	6.65	6.14	5.87	6.04	6.09	6.16	5.79	6.04	6.33	6.28	6.49	6.58	6.47	6.57	6.24	6.49	6.24	6.32	6.57	5.86
Method-B	6.33	6.38	6.42	6.49	6.55	6.43	6.19	6.09	6.16	6.09	5.99	6.06	6.15	6.03	6.42	6.12	5.96	6.00	7.03	6.48
Method-C	6.31	6.43	6.34	6.50	6.62	6.45	5.92	5.93	6.13	6.08	6.11	6.14	6.22	6.06	6.36	6.25	5.99	5.83	7.01	6.37
Method-D	6.34	6.48	6.37	6.57	6.64	6.57	6.00	6.05	6.25	5.35	6.36	6.20	6.50	5.89	6.49	6.45	6.09	5.99	7.11	5.75

(B)	CYS	MET	PHE	ILE	LEU	VAL	TRP	TYR	ALA	GLY	THR	SER	GLN	ASN	GLU	ASP	HIS	ARG	LYS	PRO
Original†	−0.96	−1.36	−2.22	−1.89	−2.29	−1.34	−1.28	−0.88	−0.02	0.38	0.25	0.55	0.56	0.63	0.74	0.72	0.00	0.44	1.01	0.47
Method-A	−0.16	−0.22	−0.36	−0.31	−0.37	−0.22	−0.21	−0.14	−0.00	0.06	0.04	0.09	0.09	0.10	0.12	0.12	0.00	0.07	0.17	0.08
Method-B	−0.36	−0.30	−0.34	−0.33	−0.38	−0.29	−0.24	−0.23	−0.03	0.09	0.01	0.10	0.13	0.13	0.23	0.17	−0.04	0.09	0.32	0.20
Method-C	−0.38	−0.30	−0.38	−0.34	−0.37	−0.29	−0.33	−0.29	−0.04	0.09	0.03	0.12	0.14	0.13	0.23	0.19	−0.04	0.07	0.33	0.19
Method-D	−0.32	−0.25	−0.33	−0.28	−0.32	−0.23	−0.27	−0.23	−0.02	−0.02	0.05	0.11	0.15	0.10	0.21	0.19	−0.02	0.08	0.30	0.11

†"Original" refers to the original method based on the Bethe approximation; for other methods, see the footnotes of Table II. The same set of proteins and sampling weights as in Miyazawa and Jernigan[2] are used.

protein structures and length 172 of Table IV-B for face-centered cubic lattices; note that the deviations of the optimized q_i from the actual value, 12, for face-centered cubic lattices have no physical bases.

For values of the regression coefficients for partition energies, protein structures appear to be more similar to face-centered cubic lattices than to the simple cubic lattices, indirectly confirming the underlying lattice type for protein structures.[38] In the case of intrinsic pairwise energies, regression coefficients exhibit a wide range from 0.37 to 0.84, depending on the protein size, for the face-centered cubic lattice. However, the values, about 0.5 ∼ 0.8, of regression coefficients in Method-B for real proteins lie within nearly this same range.

Figure 2A shows a comparison of "input" and predicted energies with the Bethe approximation for intrinsic pairwise energies in Method-D; the input contact energies are those values self-consistently estimated from the observed frequencies of contacts. Although the relationship between the "input" and predicted energies is slightly convex rather than linear, the "input" values are approximated by a linear regression line. Points that deviate from the regression line in a low energy region correspond to the Cys-Cys pair and residue pairs between positively charged residues, Lys and Arg, and negatively charged ones, Glu and Asp. In addition to the attractive intrinsic pairwise energies for these residue pairs, repulsive interactions for Lys-Lys and Arg-Lys are underestimated. Figure 2B shows the comparison of "input" contact energies and those estimated by assuming a linear relationship between "input" and predicted values for partition energies and intrinsic pairwise energies. Although there are a few residue pairs for which the expected values of the contact energies deviate from their "input" values, the correlation for the contact energies is higher than 0.98; see Table II.

Figure 3A shows the dependences of the regression coefficients of "input" versus predicted energies for partition energies on the protein's surface-to-volume ratios; here only 86 non-homologous monomeric proteins are shown. Because an approximate linearity is clearer for monomeric proteins than for multimeric ones, they depend also on other geometrical factors. The results for lattice monomers show that this linearity is crude, and the regression coefficients depend also on lattice structures and interaction strengths; see Table IV. Generally, the partition energies $\Delta e_{ir}(\equiv e_{ir} - e_{rr})$ tend to be more overestimated for larger proteins with the Bethe approximation, as pointed out by Thomas and Dill.[33]

On the contrary, Figure 3B shows that the estimates of intrinsic pairwise energies depend less on protein size, at least for this set of interaction energies. The deviation of the regression coefficients from the mean tends to be larger for smaller proteins than for larger ones. However, the results for Method-B and Method-C for proteins and Table IV-B for face-centered cubic lattices, in which secondary structure energies are not included, show that the intrinsic pairwise energies tend to increase with the surface-to-volume ratio of structures, although correlations between

TABLE IV. Correlations Between "Input" and Predicted Values for Relative Partition Energies (Δe_{ir}) and Intrinsic Inter-Residue Energies (δe_{ij}) From the Equilibrium Distributions of Amino Acid Mixtures in the Most Compact Configurations on Lattices[†]

A. Simple cubic lattice

		$\Delta e_{ir}(\equiv e_{ir} - e_{rr})$		$\delta e_{ij}(\equiv e_{ij} + e_{rr} - e_{ir} - e_{rj})$	
Number of residues	$n_{r0}/(q_r n_r/2)$*	Correlation coefficient	Regression coefficient	Correlation coefficient	Regression coefficient
64	0.25	0.991	1.13	0.995	0.97
125	0.20	0.987	0.95	0.992	0.89
216	0.17	0.985	0.85	0.990	0.86
343	0.14	0.984	0.80	0.991	0.84
512	0.12	0.983	0.75	0.990	0.83
729	0.11	0.982	0.72	0.990	0.84

B. Face-centered cubic lattice

		$\Delta e_{ir}(\equiv e_{ir} - e_{rr})$		$\delta e_{ij}(\equiv e_{ij} + e_{rr} - e_{ir} - e_{rj})$	
Number of residues	$n_{r0}/(q_r n_r/2)$*	Correlation coefficient	Regression coefficient	Correlation coefficient	Regression coefficient
63	0.37	0.993	0.38	0.971	0.84
108	0.31	0.992	0.35	0.942	0.65
172	0.27	0.994	0.32	0.914	0.54
256	0.23	0.993	0.29	0.890	0.47
365	0.21	0.993	0.28	0.870	0.43
500	0.19	0.993	0.26	0.853	0.39
666	0.17	0.994	0.25	0.838	0.37

[†]In these calculations, the value of the coordination number is fixed at the actual value, 6 for simple cubic lattices or 12 for face-centered cubic lattices. The self-consistently estimated values of contact energies in Method-B are used as "input" nearest neighbor interactions. Relative temperature $T_{\rm rel}$ is taken to be one.

*$n_r \equiv \sum_{i=1} n_i$ and $q_r\, n_r \equiv \sum_{i=1} q_i\, n_i$.

them are quite weak, with correlation coefficients of about 0.6, in Method-B and Method-C.

The Effects of Other Interactions on the Estimation of Contact Energies

Obviously, other interactions besides contact potentials are operative in real proteins. Does the equilibrium distribution of contacts better reproduce the observed one when other interactions are taken into account? Do these significantly change the estimation of contact energies? Here the effects of repulsive packing potentials, secondary structure potentials,[13] and also random noise to simulate other interactions have been examined. As shown in Table II, the repulsive packing potentials reduce the relative error $|\Delta N_{ij}|\ /|N_{ij}|$ as well as $|\Delta N(i,\, n^c)\ |/|N(i,\, n^c)|$. Because the repulsive packing potentials work to reproduce the observed distributions of the number of residues in contact, $N(i,\, n^c)$, it is reasonable that the relative error in their equilibrium distributions, $|\Delta N(i,\, n^{\ c})|/|N(i,\, n^c)|$, is smaller for Method-C than for Method-B. It is noteworthy that the repulsive packing potentials are also effective in reducing the relative error of $|\Delta N_{ij}|/|N_{ij}|$. However, these effects are not detectable in the relative errors, $|\Delta n(i,\, n^c)|/|n(i,\, n^c)|$ and $|\Delta n_{ij}|/|n_{ij}|$, for each protein, which show little or no change upon addition of the repulsive packing energies. On the other hand, correlations between "input" and predicted values for intrinsic pairwise energies are significantly improved for each protein, although there is no such trend for partition energies; Figure 4 shows a comparison of the correlation coefficients for intrinsic pairwise energies for each of 86 non-homologous monomeric proteins. This shows probably because the use of coordination numbers for the maximum number of neighbors becomes more appropriate in this situation. These results indicate that including repulsive packing energies improves on average the reproducibility of the observed frequencies of contacts, and also increases the predictability of the intrinsic inter-residue energies.

On the other hand, secondary structure potentials are not related directly to the distribution of contacts at all, so that they can potentially interfere with favorable pairwise interactions. Correlations between "input" and predicted values for both types of energies, partition and intrinsic pairwise energies, with the Bethe approximation become significantly worse in almost all proteins for Method-D than for the other methods; the correlation coefficients of intrinsic pairwise energies for 86 non-homologous monomeric proteins are scattered mostly within a range of 0.88 to 0.96 in Method-D but 0.94 to 0.98 in Method-C. Thus, secondary structure potentials interfere somewhat with the pairwise interactions. This small incompatibility between secondary structure interactions and pairwise tertiary structure interactions is also reflected in the increase of matched identical residues between equilibrium structures and the native proteins. The average proportion of such identical residues increases almost by a factor of 2 due to secondary structure energies; the ratio of identical residues is mostly within a range of 0.06 to 0.17 for Method-C and 0.12 to 0.35 for Method-D. On the other hand, the decrease in variabilities of residues may reduce relative errors in the predicted distribution of the number of residues in contact. Actually, the relative error is

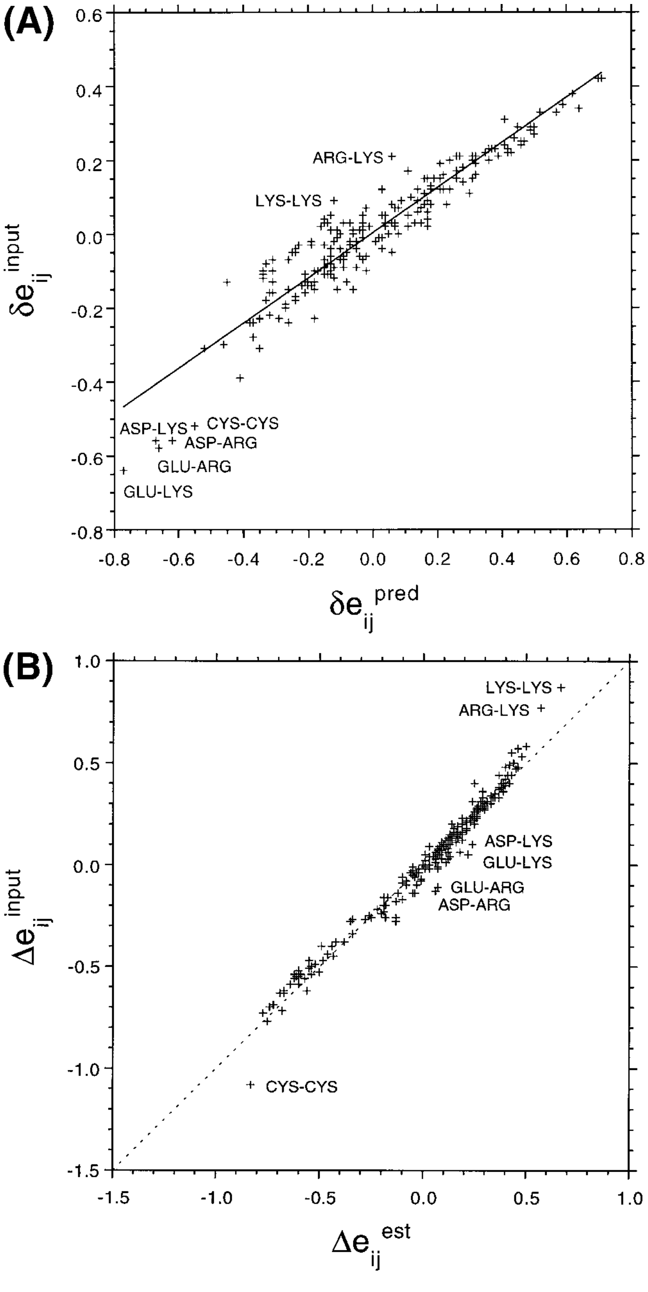

Fig. 2. (**A**) Comparison of "input" energies and energies calculated with the Bethe approximation for intrinsic pairwise energies, $\delta e_{ij}(\equiv e_{ij} + e_{rr} - e_{ir} - e_{rj})$. The solid line shows the regression line. (**B**) Comparison of "input" energies and energies estimated from the regression coefficient between "input" values and values predicted with the Bethe approximation, for $\Delta e_{ij}(\equiv e_{ij} - e_{rr})$; see Eq. 30. The dotted line indicates equal values for both coordinates. The coordination number for each type of residue is adjusted to yield the best correlation for the partition energies. "Input" energies used here are contact energies self-consistently estimated from the observed numbers of contacts and the regression coefficient between "input" values and values calculated with the Bethe approximation. Here interaction energies consist of contact energies, repulsive packing energies, and secondary structure energies; this interaction scheme corresponds to Method-D in Table II.

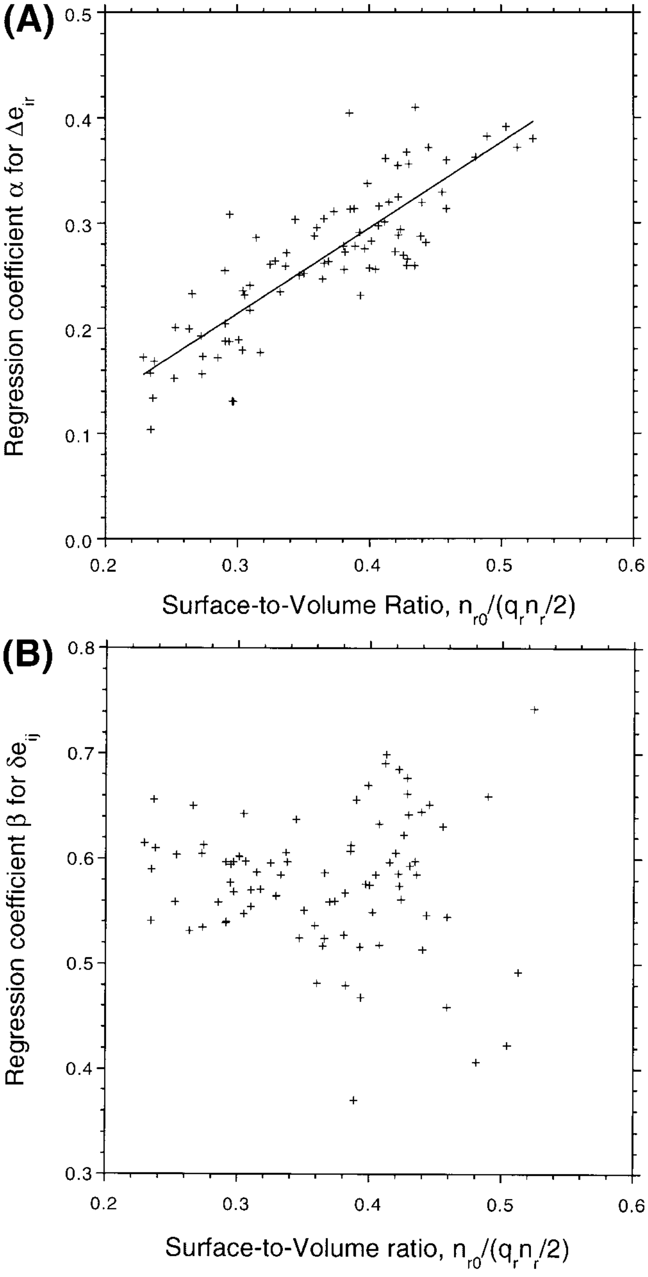

Fig. 3. Dependences of the regression coefficients of "input" versus predicted values, (**A**) α for the partition energies and (**B**) β for the intrinsic pairwise energies, on the surface-volume ratios, $n_{r0}/(q_r n_r/2)$, of monomeric proteins; $n_r \equiv \sum_{i=1} n_i$ and $q_r n_r \equiv \sum_{i=1} q_i n_i$. A solid line in (A) shows the regression line of $0.81 n_{r0}/(q_r n_r/2) - 0.03$. Correlation coefficients are 0.84 for (A) and 0.03 for (B). All interactions, contact energies, repulsive packing energies and secondary structure energies, are included to generate the equilibrium ensemble for each protein; this interaction scheme corresponds to Method-D in Table II. Here, only 86 nonhomologous monomeric proteins[2] are shown.

reduced significantly for all proteins and also for each monomeric protein; see Table II and Figure 5. Even for the frequencies of residue pairs in contact, its relative error for each monomeric protein is slightly improved, although the relative error over all proteins is rather larger for Method-D than for Method-C in Table II.

The effects of repulsive packing interactions and secondary structure interactions on the estimate of contact energies are found in Table II for intrinsic pairwise

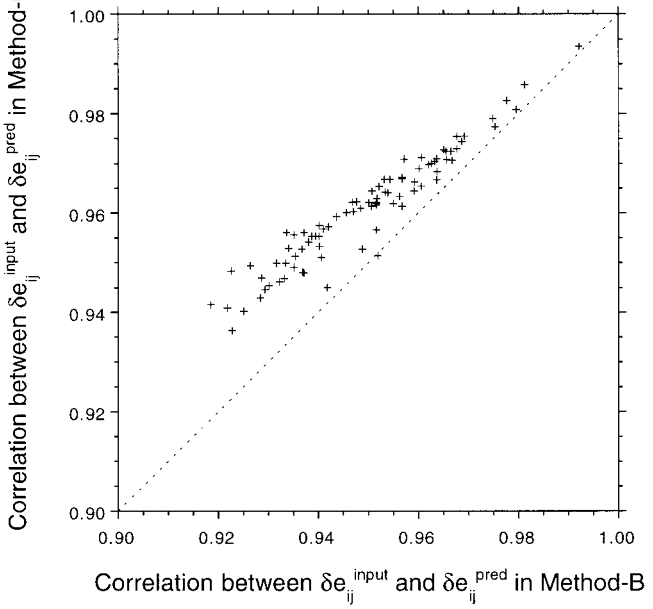

Fig. 4. Effects of repulsive interactions on the correlation coefficients β of "input," $\delta e_{ij}^{\text{input}}$, versus predicted values, $\delta e_{ij}^{\text{pred}}$, for intrinsic pairwise energies. The abscissa and ordinate values correspond to Method-B and Method-C in Table II, respectively. Only 86 non-homologous monomeric proteins[2] are shown here.

energies and in Table III for partition energies. Eqs. 23, 24, 31, and 32 in the Materials and Methods section indicate that partition energies depend on both a regression coefficient and coordination numbers, but intrinsic pairwise energies depend only on a regression coefficient. Table II shows that the regression coefficient for intrinsic pairwise energies does not change much among Method-B, C, and D; its value is 0.64 for Method-B, 0.71 for Method-C, and 0.61 for Method-D. Thus, repulsive packing interactions and secondary structure interactions do not much affect the estimates of intrinsic pairwise energies from the observed distributions of contacts. Likewise, the addition of repulsive packing energies does not change significantly the estimates of the partition energies for the 20 types of amino acids except for Trp. The effects of secondary structure energies are typically reflected in the change of the partition energy of Gly. That is, the addition of secondary structure energies makes the estimate of the partition energy of Gly less positive, because Gly tends to be located on protein surfaces in order to build specific secondary structures such as turns or bends. Generally, the estimates of partition energies become slightly less negative for non-polar amino acids and less positive for polar amino acids. This is interesting, because this feature shows a consistency between secondary structure interactions and tertiary interactions.

Interactions not considered here may be regarded as random noise. A noise with a uniform distribution in the energy range of −0.5 to 0.5 is added to the contact energy for each contact; this energy range is almost equal to the range of Δe_{ij}. In another case, the same strength of noise is added to the secondary structure energies. Both types of noise do not affect at all our estimates of contact energies, and also do not even change the correlation between "input" and predicted energies for each protein. This is probably because equilibrium ensembles rather than minimum energy structures are employed to predict contact energies.

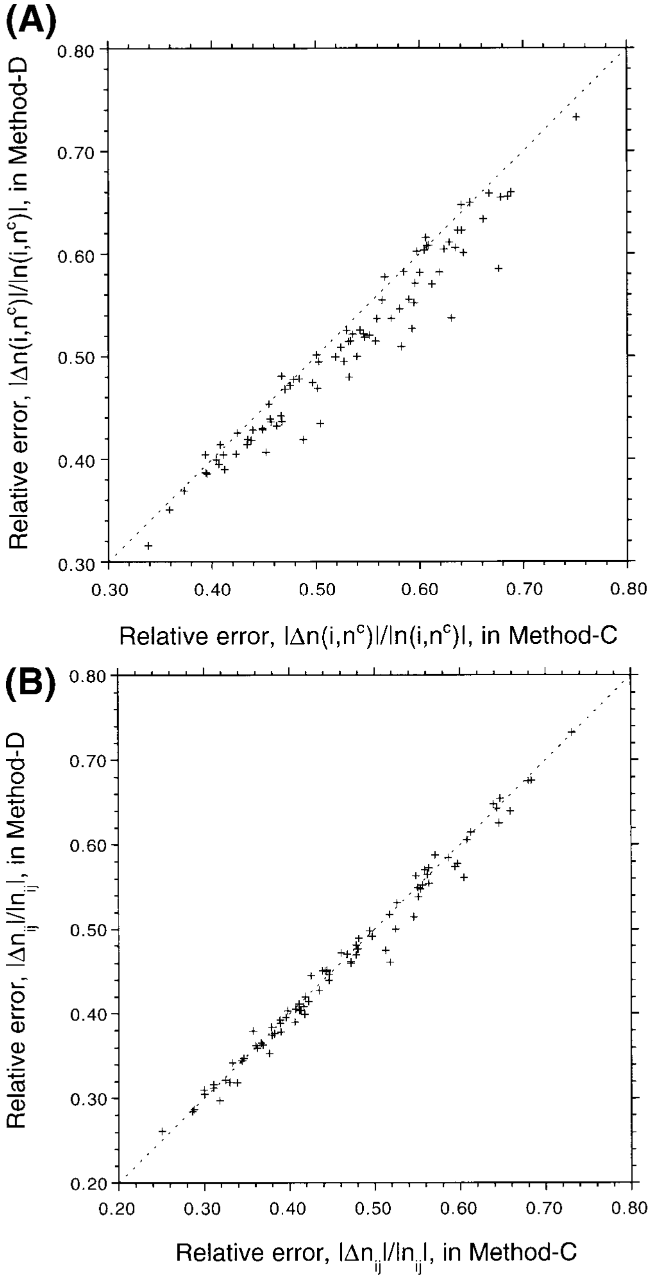

Fig. 5. Effects of secondary structure interactions on (**A**) $|\Delta n(i, n^c)| / |n(i, n^c)|$, the relative errors in the predicted frequencies of the number of contact residues in each protein, and (**B**) $|\Delta n_{ij}| / |n_{ij}|$, the relative errors in the predicted frequencies of inter-residue contacts in each protein; see Eqs. 43–44 for the definitions. The abscissa and ordinate values correspond to Method-C and Method-D in Table II, respectively. Only 86 non-homologous monomeric proteins[2] are shown here.

Can the Distribution of Contacts Be Approximated as the Equilibrium Mixture of Contacts?

In order to assess the assumption of equilibrium mixtures for the distribution of contacts, the equilibrium

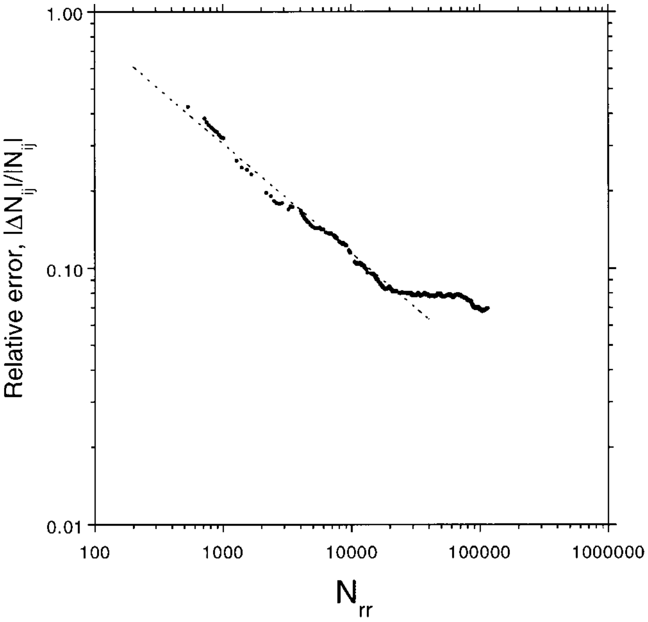

Fig. 6. Dependences of the relative errors in the predicted frequencies of inter-residue contacts, $|\Delta N_{ij}| / |N_{ij}|$, on the total number N_{rr} of contacts accumulated by adding proteins; see Eq. 44 for the definition. The dotted line corresponds to $5.8N_{rr}^{-0.43}$. This predicted distribution is an equlibrium distribution generated with the contact energies estimated by Method-D in Table II, including repulsive packing energies and secondary structure energies.

distribution of contacts is compared with the actual distribution. In Figure 6, the dependences of the relative errors of the equilibrium distributions on the total number N_{rr} of contacts by adding proteins one by one, are shown only for the frequencies of residue pairs in contact; Method-D in Table II is used here to estimate contact energies and to generate the equilibrium distributions of contacts. An interesting fact is that in the range of $N_{rr} < 10^4$ both types of relative errors, $|\Delta N(i, n^c)|/| N(i, n^c)|$ and $|\Delta N_{ij}|/|N_{ij}|$, have power dependences of -0.45 and -0.43 on the total number of contacts N_{rr}, respectively. These power dependences almost coincide with that of random sampling errors, -0.5. However, these relative errors attain a limit, about 0.08 at the sample size of $N_{rr} \sim 10^{4.5}$. The χ^2 value of the equilibrium distributions of contacts for each method in Table II is significantly larger than the 5% limit. Because the χ^2 values should be roughly proportional to $|\Delta N(i, n^c)|^2/|N(i, n^c)|$ or $|\Delta N_{ij}|^2/|N_{ij}|$, the χ^2 values will remain constant in the range of $N_{rr} < 10^4$ but will then increase with N_{rr} beyond the 5% limit. Although the observed distributions of contacts cannot precisely be regarded as the equilibrium distributions with the inter-residue contact energies newly estimated here, they can be approximated with relative errors below 0.1 as the equilibrium mixture of residues in protein structures. The match between the observed and equilibrium frequencies of contacts may be improved with better estimates of contact energies.

Here the same set of proteins is used to estimate contact energies and also to assess the assumption of equilibrium mixtures for the distribution of contacts, because the present purpose is not to examine the dependence of contact energies on data but to test whether the observed distribution of contacts can be regarded as an equilibrium mixture of contacts. It was confirmed[2] that values[2] of contact energies calculated with the Bethe approximation from the present set of proteins do not significantly change from those[1] calculated from the smaller set of proteins, 18,192 contacts. This is consistent with the fact that, as indicated by Figure 6, the estimates of contact energies can hardly be improved by employing more contacts than $N_{rr} \sim 10^{4.5}$.

The dependence of the relative errors in the predicted distribution of the number of contact residues and that in the predicted frequencies of contact residue pairs on the number of samples, i.e., the total number of contacts, n_{rr}, in a protein has also been considered. The slopes of the regression lines in these log-log plots indicate $|\Delta n(i, n^c)|/|n(i, n^c)| \propto n_{rr}^{-0.21}$ and $|\Delta n_{ij}|/|n_{ij}| \propto n_{rr}^{-0.28}$; Method-D in Table II is also used here to estimate contact energies and to generate the equilibrium distributions of contacts. These power dependences are almost half as large as for random sampling errors, probably because of the effects of chain connectivity.

Characteristics of Newly Estimated Contact Energies

The newly estimated values of relative contact energies with Eqs. 31–34 in Method-D are listed in Table V. The most remarkable change in the relative contact energies ($\Delta\epsilon_{ij}$) from the original values directly predicted with the Bethe approximation is that the new estimates are reduced to only about 30% of their original estimates. Also the new estimates of relative partition energies ($\Delta\epsilon_{ir}$) are less than 30% of the original estimates. On the other hand, intrinsic pairwise energies ($\delta\epsilon_{ij}$) are estimated to be about 60% of their original values. Since the original estimates of the partition energies are about twice the hydrophobic energies of Nozaki and Tanford,[39] the present estimates come closer to being about half of their hydrophobic energies, with the energies measured relative to that of glycine. This is a puzzle at present. On the other hand, there are more reasonable features apparent in the present estimate than in the original one. Cys–Cys pairs are the most attractive in the present estimate, but most non-polar residue pairs had lower contact energies than the Cys–Cys pair in the original estimate. This change is caused by two kinds of changes in the energy estimation. First, the proportion of the partition energy in the contact energy is much lower in the present scheme than in the original estimate with the Bethe approximation; in Table II, the regression coefficient α of the partition energy is smaller than that of the intrinsic pairwise energy. Second, the estimate of the partition energy for Cys is about as negative as for other non-polar residues; compare the partition energies for Method-A with those for other methods in Table III-B.

Another remarkable change is that the relative contact energies ($\Delta\epsilon_{ij}$) for residue pairs between charged residues

TABLE V. Contact Energies in RT Units Estimated by Method-D; $\Delta\epsilon_{ij}(\equiv \epsilon_{ij} - \epsilon_{rr})$ for Upper Triangular Half and Diagonal and $\delta\epsilon_{ij}(\equiv \epsilon_{ij} + \epsilon_{rr} - \epsilon_{ir} - \epsilon_{rj})$ for Lower Triangular Half. These Energies Come from Method-D in Table II

	CYS	MET	PHE	ILE	LEU	VAL	TRP	TYR	ALA	GLY	THR	SER	GLN	ASN	GLU	ASP	HIS	ARG	LYS	PRO	
	−1.19	−0.61	−0.67	−0.64	−0.65	−0.59	−0.66	−0.39	−0.33	−0.31	−0.15	−0.13	−0.07	−0.01	0.20	0.12	−0.36	0.08	0.33	−0.18	CYS
CYS	−0.54	−0.70	−0.83	−0.66	−0.70	−0.51	−0.73	−0.56	−0.27	−0.17	−0.11	0.05	−0.06	0.04	0.12	0.30	−0.29	0.03	0.29	−0.13	MET
MET	−0.03	−0.19	−0.88	−0.73	−0.80	−0.67	−0.68	−0.58	−0.36	−0.19	−0.15	−0.12	−0.11	−0.01	0.14	0.18	−0.34	−0.05	0.19	−0.19	PHE
PHE	−0.02	−0.24	−0.22	−0.74	−0.81	−0.67	−0.60	−0.49	−0.37	−0.13	−0.15	0.03	−0.01	0.14	0.17	0.22	−0.13	0.00	0.24	−0.05	ILE
ILE	−0.03	−0.13	−0.11	−0.18	−0.84	−0.74	−0.62	−0.55	−0.38	−0.16	−0.15	−0.02	−0.04	0.04	0.17	0.27	−0.18	−0.04	0.22	−0.12	LEU
LEU	−0.00	−0.12	−0.14	−0.20	−0.19	−0.65	−0.51	−0.38	−0.32	−0.15	−0.07	0.04	0.08	0.12	0.26	0.36	−0.06	0.08	0.29	−0.05	VAL
VAL	−0.03	−0.03	−0.11	−0.16	−0.19	−0.20	−0.64	−0.49	−0.27	−0.25	−0.02	−0.01	−0.02	−0.10	−0.00	0.07	−0.37	−0.21	0.09	−0.37	TRP
TRP	−0.06	−0.21	−0.08	−0.05	−0.03	−0.02	−0.11	−0.45	−0.20	−0.22	−0.09	−0.08	−0.14	−0.11	−0.08	−0.07	−0.30	−0.25	−0.05	−0.25	TYR
TYR	0.16	−0.08	−0.02	0.02	0.00	0.08	0.01	0.00	−0.12	−0.08	0.04	0.10	0.22	0.15	0.38	0.27	0.07	0.24	0.41	0.15	ALA
ALA	0.01	0.00	−0.02	−0.07	−0.04	−0.07	0.01	0.05	−0.09	−0.29	−0.04	−0.01	0.13	−0.01	0.32	0.11	0.00	0.09	0.29	0.02	GLY
GLY	0.03	0.10	0.16	0.17	0.18	0.09	0.03	0.03	−0.05	−0.26	0.03	0.04	0.12	0.04	0.16	0.11	−0.03	0.11	0.33	0.13	THR
THR	0.12	0.09	0.13	0.08	0.13	0.10	0.19	0.08	0.00	−0.08	−0.08	0.05	0.22	0.09	0.18	0.10	0.04	0.16	0.36	0.20	SER
SER	0.08	0.19	0.10	0.20	0.20	0.16	0.15	0.04	0.00	−0.10	−0.13	−0.17	0.20	0.06	0.27	0.24	0.15	0.09	0.28	0.17	GLN
GLN	0.10	0.04	0.07	0.12	0.14	0.15	0.09	−0.06	0.08	−0.01	−0.09	−0.04	−0.10	−0.06	0.12	0.02	0.00	0.10	0.22	0.18	ASN
ASN	0.21	0.19	0.22	0.32	0.26	0.25	0.06	0.02	0.07	−0.10	−0.12	−0.12	−0.20	−0.27	0.46	0.44	0.00	−0.22	−0.06	0.37	GLU
GLU	0.31	0.16	0.26	0.24	0.28	0.27	0.05	−0.07	0.18	0.13	−0.11	−0.14	−0.10	−0.19	0.03	0.29	−0.10	−0.24	−0.01	0.33	ASP
ASP	0.26	0.37	0.33	0.32	0.41	0.41	0.15	−0.03	0.10	−0.06	−0.13	−0.20	−0.10	−0.27	0.04	−0.08	−0.40	0.05	0.38	0.01	HIS
HIS	−0.01	−0.02	0.00	0.18	0.17	0.19	−0.08	−0.05	0.10	0.04	−0.06	−0.06	0.02	−0.08	−0.19	−0.26	−0.36	0.19	0.66	0.17	ARG
ARG	0.33	0.21	0.20	0.21	0.21	0.22	−0.03	−0.10	0.18	0.02	−0.03	−0.04	−0.14	−0.08	−0.52	−0.51	−0.02	0.03	0.76	0.47	LYS
LYS	0.35	0.24	0.22	0.23	0.24	0.22	0.05	−0.12	0.13	0.01	−0.02	−0.05	−0.17	−0.18	−0.58	−0.49	0.10	0.28	0.16	0.11	PRO
PRO	0.03	0.01	0.03	0.12	0.09	0.06	−0.22	−0.13	0.05	−0.08	−0.03	−0.02	−0.10	−0.04	0.04	0.03	−0.08	−0.03	0.05	−0.11	
$e_{ir} - e_{rr}$	−0.32	−0.25	−0.33	−0.28	−0.32	−0.23	−0.27	−0.23	−0.02	−0.02	0.05	0.11	0.15	0.10	0.21	0.19	−0.02	0.08	0.30	0.11	

(Glu, Asp, Arg, and Lys) and non-polar residues (Met, Phe, Ile, Leu, and Val) were negative in the original estimate but become positive in the present estimate. This is caused by the smaller contribution of the partition energies in the present contact energies. Negative energies for the oposite charge pairs Arg-Glu and Arg-Asp in the present estimate are also more reasonable than the positive energies of the original estimate.

Simple Threading With Newly-Estimated Contact Energies

It has been examined how the present estimate of contact energies affects the recognition of the correct pairs of native sequences and structures compared with other non-native sequence–structure pairs, by doing conventional threading and inverse threading simulations as described in Miyazawa and Jernigan.[2,13] A set of proteins each of which represent a different protein fold was prepared. For each protein in this set, we examine whether the pair of native sequence and structure is recognizable over other non-native sequences or structures. Release 1.35 of the SCOP database[40] is used as a classification of protein folds. Another set of proteins is also prepared to provide non-native sequences or structures. This set of proteins consists of single protein representatives from each domain defined in the SCOP database.

These representatives of families or domains are the first entries in the protein lists of each family or each domain in SCOP; if these first proteins in the lists are not appropriate to use for the present purpose, then the second ones are chosen. These families and domains are all those which belong to the protein classes 1 to 5; that is, classes of all α, all β, α/β, α + β, and multi-domain proteins. Classes of membrane and cell surface proteins, small proteins, peptides and designed proteins are not used. Proteins whose structures were determined by NMR or with resolution worse than 2.5 Å are removed. Also, proteins whose coordinate sets either consist of only C^α atoms, include many unknown residues, or lack many atoms or residues, are removed. Proteins shorter than 50 residues are also removed.

In the SCOP database, protein domains whose sequences are highly homologous may be classified into the same domains, and protein domains whose structures are extremely similar may belong to different domains although in the same family. Therefore, protein pairs, which are more similar than 0.9 sequence identity, or whose structures are more similar than 1 Å r.m.s.d. (root mean square deviation), are also removed from the set of domain representatives. As a result, the set of family representatives includes 440 proteins and the set of domain representatives has 988 proteins.

Each protein sequence of family representatives is threaded into the protein structures of domain representatives in the conventional threading case, or the sequences of domain representatives are threaded into each structure of family representatives in the inverse threading case in order to examine if these structures can recognize the native sequences. The total energy score,[2,13] $\Delta E^c(\Delta\epsilon_{ij})$ $(\equiv E^c(\Delta\epsilon_{ij}) - \langle E^c(\Delta\epsilon_{ij})\rangle)$ that is defined as the total relative contact energies $(E^c(\Delta\epsilon_{ij}) = \sum_{ij} n_{ij}\ \Delta\epsilon_{ij})$ relative to the average energy of native proteins ($< E^c(\Delta\epsilon_{ij}) >$), is calculated for protein sequences threaded at all possible positions in all other protein structures, and their means and standard deviations are calculated; no gaps in either the sequences or the structures are allowed. Total energy scores are calculated for multimeric states only if the coordinates of the other bound subunits are given in the PDB file.[37] Then, the positions of the native energies in the distributions of all threadings, i.e., z-scores, are measured in units of standard deviation (s.d.) where larger negative values indicate that the native energies are further below the mean.

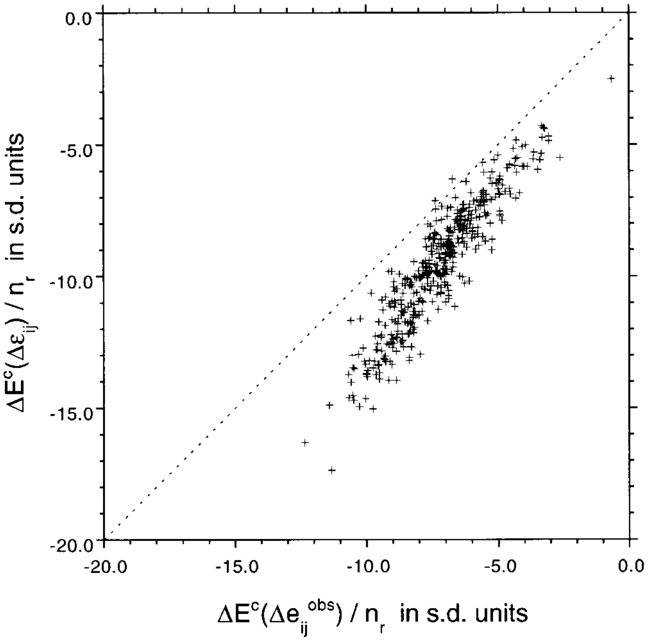

Fig. 7. Comparison of the effects of contact energies on the discrimination of native structures from other non-native folds with a given sequence. Both ordinate and abscissa show z-scores that are defined as the total energy scores per residue $\Delta E^c(\Delta\epsilon_{ij})/\ n_r$ of proteins,[2,13] in standard deviation units from the mean in the energy distribution of random threadings. The previous estimate of contact energies[2] is used on the abscissa and the present estimate with Method-D in Table II is used on the ordinate. 440 proteins each of which represents a protein family in Release 1.35 of SCOP database[40] are threaded into each of 988 proteins each of which represents a protein domain defined in SCOP. See text for details.

Figure 7 shows how the present new estimate of contact energies discriminates native structures from other non-native folds with conventional threading to compare with the previous estimate.[2] The ordinate gives z-scores calculated with the values of contact energies estimated by Method-D in Table II, and the abscissa is for the previous estimate. In the conventional threading case, z-scores for almost all proteins are significantly higher with the present estimate of contact energies than with the previous estimate, although such improvements are not observed in the case of inverse-threading case. This fact supports the present estimates in which the proportion of partition energies present in the contact energies is reduced in comparison with the previous estimates. However, it should be noted here that with this consideration of z-scores one cannot consider the magnitudes of contact energies.

DISCUSSION

It has been pointed out[33,34] that statistical potentials should be tested with respect to their self-consistency to see if “input” potentials can be reproduced. Thomas and Dill[33] found several discrepancies between “input” potentials and extracted potentials using both the methods of Miyazawa and Jernigan[1] and Sippl.[5] Dependences of extracted energies on the surface-to-volume ratio of a protein and/or the differences in apparent temperatures for proteins are among those which are also confirmed here. In other words, the regression coefficient of the partition energies between “input” and predicted energies, which corresponds to an apparent temperature, depends on the surface-to-volume ratio of a protein. Simulations on lattices and on protein structures here indicate that this is certainly one limitation of the Bethe approximation, when it is applied to such small systems; the fact that the Bethe approximation cannot reproduce the surface-to-volume ratio of proteins was pointed out in our original paper.[1] Thomas and Dill[33] asserted from this negative evidence that the statistically extracted potentials are not an accurate approximation of the true potentials. This conclusion contrasts with the result of Mirny and Shakhnovich[34] showing that procedures[1,2,4] based on the Bethe approximation can extract potential with impressive accuracy.

This difference comes from the negative emphasis of Thomas and Dill[33] compared to the more positive attitude of Mirny and Shakhnovich.[34] In fact, the correlations between “input” and predicted energies are mostly remarkably high, at least for a reasonable range of contact energies; see Tables I, II, and IV. Here we have utilized these positive features of the Bethe approximation, in order to statistically estimate contact energies.

Mirny and Shakhnovich[34] showed how to a large extent, applications of pairwise interactions can yield native-like character, not only the lowest energy state for native structures but also large energy gaps, i.e., for the foldability of native structures. Our purpose here is to reproduce actual pairwise interactions between residues in protein structures, but not to optimize pairwise interactions to design native structures. If these are not sufficient to design foldable proteins as they pointed out, then other interactions that are physically present and responsible for protein folding must be sought.

Energy gaps, which can be a measure of protein foldability, were defined by Mirny and Shakhnovich[34] as gaps between the lowest energy and the average energy over all compact structures. Because the average energy of all feasible compact structures depends only slightly on sequence but mostly only on amino acid composition, lower energy structures correspond to structures with larger energy gaps. Therefore, sequences in equilibrium consist of those with relatively large energy gaps and thus have relatively high foldabilities.

A theory was proposed[41,42] to explain why the statistics of globular structures are Boltzmann-like. It is phenomenologically known that the distributions of secondary structures in protein structures are Boltzmann-like. However, the present results on the reproducibility of contact energies by the Bethe approximation show that even in residue mixtures interacting with contact energies, the equilibrium distribution of contacts is Boltzmann-like but the relative temperatures of the Boltzmann factors are different for partition energies and intrinsic pairwise energies. This temperature also depends, among other things, on the surface-to-volume ratio of a protein.

Here a more basic question, i.e., whether or not the observed distributions of contacts are well approximated by the equilibrium distributions of residue mixtures interacting with pairwise contact energies, has been examined by comparing their distributions to one another. The relative errors of the equilibrium distributions depend on the number of samples roughly in the form of $N_{rr}^{-0.43}$, which is close to the form of $N_{rr}^{-0.5}$ expected for random sampling; see Figure 6. However, the relative errors become almost constant at a sample size of $N_{rr} \sim 10^{4.5}$ and cannot be improved by utilizing larger samples, indicating a limitation to the estimation of contact energies and/or other higher order interactions not considered. The effects of chain connectivity on the distribution of contacts are also indicated by the power dependence of the relative errors on the number of contacts for individual proteins which is -0.28, much smaller than the value -0.43 for a large set of proteins.

In this paper, a linear regression is used to correct intrinsic inter-residue energies calculated with the Bethe approximation and then to derive contact energies, which are calculated from corrected partition energies and intrinsic pairwise energies; see Eqs. 29, 30. Estimating intrinsic pairwise energies or contact energies by higher order polynomials rather than with a linear equation could improve the estimation of contact energies, especially for Cys–Cys and charged residue pairs; see Figure 2. Cys and charged residues (Glu, Asp, Arg, and Lys) have relatively large χ^2 values for the agreement of the frequencies of residue pairs in contact, N_{ij}. Such an improvement might yield a better match between the observed and equilibrium frequencies of contacts especially for these residues.

Also, the reference state for inter-residue contacts, i.e., C_{ij} in Eqs. 23, 24, is always assumed to be a random mixture of residues, but it should be set up for Method-D to include secondary structure interactions in the reference state. This might also improve correlations between “input” and predicted values with the Bethe approximation in Method-D.

Some contributions to the frequencies of contacts ignored here are chain connectivity[43] and interactions such as long-range electrostatic interactions, hydrogen bonding interactions, and interactions of higher order[44] than two-body. The present analyses indicate that the observed distributions of contacts could be approximated with a relative error less than 0.1 by their equilibrium distributions if distributions for many proteins rather than for single proteins are considered.

It was suggested[35] that the over-emphasis of hydrophobic energies is responsible for the (relatively) poor performance of the contact potentials of Miyazawa and Jernigan[1] for the recognition of near-native folds. Actually the present estimate of contact energies, in which partition energies and intrinsic pairwise energies are better balanced, is shown to increase the capability of discriminating native folds from other non-native folds; see Figure 7. However, energy differences among residue pairs in the present estimate of contact energies are only about half of the hydrophobic energy changes estimated from the experimental values of hydrophobic energies for residues.

CONCLUSION

Here pairwise contact energies for 20 types of residues have been self-consistently estimated in the approximation of equilibrium mixtures of residues for proteins. First, by comparing "input" and energies predicted with the Bethe approximation (quasi-chemical approximation) for the equilibrium mixtures of residues interacting with "input" contact energies, the reproducibilities of "input" contact energies with this approximation are examined. In this system, the intrinsic pairwise energies and the (relative) partition, hydrophobic energies can be predicted by the Bethe approximation. Calculations for amino acid mixtures on lattices and in protein structures indicate that correlations between "input" and values predicted for both types of energies are sufficiently high, better than 0.9 for most proteins, but regression coefficients themselves depend on lattice or protein structures, as well as interaction strengths. Because of these high correlations, "input" contact energies can be estimated well from the predicted values and the regression coefficient. Because of the dependences of the regression coefficients on interaction strengths, contact energies are self-consistently estimated from the actual observed frequencies of contacts with regression coefficients obtained by comparing "input" and predicted values for the equilibrium mixtures of residues generated with the contact energies taken as the true ones. Coordination numbers are optimized to obtain the best correlation between "input" and predicted values for partition energies. Other interactions such as repulsive packing energies, secondary structure energies, and random noise are added to generate equilibrium mixtures of residues, and their effects on the estimation of contact energies are examined.

The contact energies self-consistently estimated indicate that the partition energies predicted with the Bethe approximation may be reduced by a factor of about 0.3 and the intrinsic pairwise energies by a factor of about 0.6, decreasing the contribution of the partition energies. This new estimate of contact energies, in which the proportion of partition energies is much less than in the predicted values with the Bethe approximation, increases the capability for discriminating native structures from other non-native folds.

The equilibrium mixture approximation of residues for proteins is supported at least to the extent that the observed frequencies of residue pairs in contact can be approximated with a relative error of about 0.08, if many proteins are employed to collect more than 20,000 contacts. The inclusion of repulsive packing interactions[2] and secondary structure interactions[13] further reduces the relative errors.

APPENDIX

Differences Between the Present Contact Energies and Sippl's Type of Pairwise Potentials of Mean Force

Sippl[5] defined a net potential, $\Delta E_k^{ij}(d)$, between i and j types of amino acids as the potential, $E_k^{ij}(d)$ of mean force relative to the overall potential, and $E_k(d)$, of mean force:

$$\Delta E_k^{ij}(d) \equiv E_k^{ij}(d) - E_k(d) = -\ln \frac{f_k^{ij}(d)}{f_k(d)} \tag{45}$$

where $f_k^{ij}(d)$ is the probability that i and j types of amino acids, separated by k residues in protein sequence, are located at the distance d in protein structures. $f_k(d)$ is such a probability over all types of amino acid pairs. The pairwise potentials devised by Kocher et al.[12] and Nishikawa and Matsuo[16] are essentially the same as Sippl's.

From this definition, the following equation can be derived; in the following the subscript "k" that represents the length of separation in the amino acid sequence is omitted for simplicity:

$$\begin{aligned}
&\Delta E^{ij}(d) + \Delta E^{rr}(d) - \Delta E^{ir}(d) - \Delta E^{rj}(d) \\
&= -\ln\left[\left(\frac{2\overline{n_{ij}(d)}}{n_i n_j}\right)\Big/\left(\frac{2\overline{n_{rr}(d)}}{n_r(n_r-1)}\right)\right] + \ln\left[\left(\frac{2\overline{n_{ir}(d)}}{n_i(n_r-1)}\right)\Big/\left(\frac{2\overline{n_{rr}(d)}}{n_r(n_r-1)}\right)\right] \\
&\quad + \ln\left[\left(\frac{2\overline{n_{rj}(d)}}{(n_r-1)_{nj}}\right)\Big/\left(\frac{2\overline{n_{rr}(d)}}{n_r(n_r-1)}\right)\right] \simeq -\ln\left[\frac{\overline{n_{ij}(d)n_{rr}(d)}}{\overline{n_{ir}(d)n_{rj}(d)}}\right] \\
&= e_{ij}(d) + e_{rr}(d) - e_{ir}(d) - e_{rj}(d).
\end{aligned} \tag{46}$$

Here the numbers of contacts and contact energies are expressed as functions of distance. That is, $2n_{ij}(d)$ is the number of amino acid pairs of i and j types at distance d. $n_{ir}(d)$ is the sum of $n_{ij}(d)$ over all amino acid types, j. The bar designates the statistical average. n_i is the number of i type amino acids, and n_r is the total number of amino acids. $e_{ij}(d)$ is an interaction energy between i and j types of amino acids at distance d, and is defined in the same way as Eq. 9a of Miyazawa and Jernigan:[1]

$$e_{ij}(d) \equiv -\ln\left(\frac{\overline{n_{ij}(d)n_{00}(d)}}{\overline{n_{i0}(d)n_{0j}(d)}}\right) \tag{47}$$

where the subscript 0 means solvent molecules (water). Here it should be noted that the equation above is the definition for $e_{ij}(d)$, but is no longer in accord with the Bethe approximation for the contact energies, e_{ij}.

Another useful expression that clarifies the relationships between Sippl's potential $\Delta E_k^{ij}(d)$ and contact energies e_{ij} is as follows:

$$e_{ij}(d) - e_{rr}(d) \simeq \Delta E^{ij}(d) + \Delta E^{i}(d) + \Delta E^{j}(d) \quad (48)$$

$$e_{ir}(d) - e_{rr}(d) = \Delta E^{ir}(d) + \Delta E^{i}(d) \quad (49)$$

where $\Delta E^i(d)$ corresponds to the relative hydrophobic energy for a pair of amino acids one of which is i type and the other of which is located at distance d, and is defined as

$$\Delta E^{i}(d) \equiv -\ln\left[\left(\frac{\overline{n_{r0}(d)}}{n_r}\right)\Big/\left(\frac{\overline{n_{i0}(d)}}{n_i}\right)\right]. \quad (50)$$

Thus, in comparison with $e_{ij}(d)$, Sippl's potential $\Delta E_k^{ij}(d)$ does not include the hydrophobic energies $\Delta E^i(d)$ as well as the collapse energy $e_{rr}(d)$.

When the distance d in Sippl's potential is coarse-grained and is categorized into two classes of contact and non-contact, $e_{ij}(d \le R_c)$ must be equal to e_{ij} that is the contact energy defined by Eq. 5a of Miyazawa and Jernigan[1]; here R_c is the maximum distance defining residue pairs as pairs in contact:

$$e_{ij}(d \le R_c) = e_{ij} \quad (51)$$

$$\equiv E_{ij} + E_{00} - E_{i0} - E_{0j} \quad (52)$$

where E_{ij} is the absolute contact energy between i and j types of amino acids. The total energy is represented in the scheme of contact energies by

$$E^{total} = \sum_i \sum_j E_{ij} n_{ij} = \sum_i (2E_{i0} - E_{00}) q_i n_i / 2 + e_{rr} n_{rr} + \sum_i \sum_j (e_{ij} - e_{rr}) n_{ij} \quad (53)$$

where q_i is the coordination number for a i type amino acid. The second and third terms depend on protein conformations, but the first term does not.

These relations, specifically Eq. 48, indicate that the conformational energy of a protein could be represented in the scheme of Sippl's potential as

$$E^{\text{conf}} = \sum_d e_{rr}(d) n_{rr}(d) + \sum_d \sum_i \sum_j (e_{ij}(d) - e_{rr}(d)) n_{ij}(d) \quad (54)$$

$$= \sum_d e_{rr}(d) n_{rr}(d) + \sum_d \sum_i \sum_j (\Delta E^{ij}(d) + \Delta E^{i}(d) + \Delta E^{j}(d)) n_{ij}(d). \quad (55)$$

Thus, at least including the hydrophobic energies $\Delta E^i(d)$ is required to estimate correctly conformational energies of proteins, and even for fold recognition. Energies corresponding to $\Delta E^i(d)$ were not taken into account either in the identification of native protein folds[46] or in the detection of native-like folds.[9] A missing term in Sippl's potential was also pointed out by Sippl[45] himself.

Jones et al.[10] used the following type of formula for relative hydrophobicities of the 20 types of residues:

$$\Delta E^{i}(h) \equiv -\ln\left(\frac{f^{i}(h)}{f(h)}\right) \quad (56)$$

where $f^i(h)$ is the fraction of i type residues with residue accessibility h, and $f(h)$ is that fraction for all residue types. Kocher et al.[12] also used the same formula but employed the accessible surface area of a residue as the variable h. On the other hand, Nishikawa and Matsuo[16] employed as h the number of residues within a certain size shell surrounding a given residue. These expressions for relative hydrophobicity, which were used together with Sippl's type of pairwise potentials for fold recognition, do not exactly correspond to the present expression for relative hydrophobic energy that is required together with Sippl's type of pairwise potentials.

Finally, it should be noted that the Bethe approximation supports Eq. 47 and Eq. 52 only for nearest-neighbor interactions but does not assure the rationality of the equation, $e_{ij}(d) = E_{ij}(d) + E_{00}(d) - E_{i0}(d) - E_{0j}(d)$, for any distance beyond nearest neighbor contacts. How well Eq. 55 can approximate actual interaction energies between residues in proteins is unknown. The potentials of mean force, $\Delta E^{ij}(d)$ and $\Delta E^i(d)$, include the effects of many body interactions in proteins. High frequencies for certain amino acid pairs of i and j types at distances corresponding to a next-neighbor shell can result from favorable nearest-neighbor interactions between i and k types of amino acids and between k and j types. As a result, including the potentials of mean force for distant amino acid pairs beyond the nearest neighbors might improve the estimates of conformational energies but also could cause them to become worse by including many body effects. All interactions between amino acids in proteins have short range components including hydrophobic interactions, hydrogen bonding, van der Waals interactions, and electrostatic interactions; whereas only electrostatic interactions can have long range effects. As a result of the large number of short range interactions compared to long range ones, the short range terms are likely to be dominant.

REFERENCES

1. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
2. 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.
3. Hinds DA, Levitt M. A lattice model for protein structure prediction at low resolution. Proc Natl Acad Sci USA 1992;89:2536–2540.
4. Hinds DA, Levitt M. Exploring conformational space with a simple lattice model for protein structure. J Mol Biol 1994;243:668–682.

5. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. J Mol Biol 1990;213:859–883.
6. Bryant SH, Lawrence CE. An empirical energy function for threading protein sequence through the folding motif. Proteins 1993;16:92–112.
7. Miyazawa S, Jernigan RL. Protein stability for single substitution mutants and the extent of local compactness in the denatured state. Protein Eng 1994;7:1209–1220.
8. 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.
9. 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.
10. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
11. 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.
12. 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.
13. Miyazawa S, Jernigan RL. Empirical energy potentials with reference states for protein fold and sequence recognition. Submitted.
14. Godzik A, Koliński A, Skolnick J. Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. Protein Sci 1995;4:2107–2117.
15. 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.
16. Nishikawa K, Matsuo Y. Development of pseudoenergy potentials for assessing protein 3-D–1-D compatibility and detecting weak homologies. Protein Eng 1993;6:811–820.
17. 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 (Tokyo) 1995;118:137–148.
18. Lüthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature 1992;356:83–85.
19. Pellegrini M, Doniach S. Computer simulation of antibody binding specificity. Proteins 1993;15:436–444.
20. 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.
21. Wallqvist A, Jernigan RL, Covell DG. A preference-based free-energy parameterization of enzyme-inhibitor binding. Applications to HIV-1-protease inhibitor design. Protein Sci 1995;4:1881–1903.
22. Wilson C, Doniach S. A computer model to dynamically simulate protein folding: Studies with crambin. Proteins 1989;6:193–209.
23. Skolnick J, Kolinski A. Simulations of folding of a globular proteins. Science 1990;250:1121–1125.
24. Sun S. Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms. Protein Sci 1993;2:762–785.
25. Koliński A, Skolnick J. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. Proteins 1994;18:338–352.
26. Miller S, Janin J, Lesk AM, Chothia C. Interior and surface of monomeric proteins. J Mol Biol 1987;196:641–656.
27. Pohl FM. Empirical protein energy maps. Nature New Biol 1971;234:277–279.
28. Pohl FM. Statistical analysis of protein structures. In: Jaenicke R, editor. Protein Folding. Amsterdam: Elsevier/North-Holland Biomedical Press; 1980. p 183–196.
29. Némethy G, Scheraga HA. Protein folding. Quart Rev Biophys 1977;10:239–352.
30. Bryant S, Lawrence CE. The frequency of ion-pair substructures in proteins is quantitatively related to electrostatic potential: A statistical model for nonbonded interactions. Proteins 1991;9:108–119.
31. MacArthur MW, Thornton JM. Influence of proline residues on protein formation. J Mol Biol 1991;218:397–412.
32. Rashin AA, Ionif M, Honig B. Internal cavities and buried waters in globular proteins. Biochemistry 1986;25:3619–3625.
33. Thomas PD, Dill KA. Statistical potentials extracted from protein structures: How accurate are they? J Mol Biol 1996;257:457–469.
34. 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.
35. 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.
36. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH. Equation of state calculations by fast computing machines. J Chem Phys 1953;21:1087–1092.
37. Bernstein FC, Koetzle TF, Williams GJB et al. The protein data bank: A computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
38. Raghunathan G, Jernigan RL. Ideal architecture of residues packing and its observation in protein structures. Protein Sci 1997;6:2072–2083.
39. Nozaki Y, Tanford C. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions; establishment of hydrophobicity scale. J Biol Chem 1971;246:2211–2217.
40. 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.
41. Gutin AM, Badretdinov AYa, Finkelstein AV. Why are the statistics of globular protein structure Boltzmann-like? Mol Biol (Russia), Engl Transl 1992;26:94–102.
42. Finkelstein AV, Badretdinov AYa, Gutin AM. Why do protein architectures have Boltzmann-like statistics? Proteins 1995;23:142–150.
43. Skolnick J, Jaroszewski L, Kolinski A, Godzik A. Derivation and testing of pair potentials for protein folding. When is the quasi-chemical approximation correct? Protein Sci 1997;6:676–688.
44. 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.
45. Sippl MJ. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. J Comput Aided Mol Des. 1993;7:473–501.
46. Hendlich M, P Lackner S, Witckus H et al. 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.