Protein Science

Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes

MR Betancourt and D Thirumalai

Protein Sci. 1999 8: 361-369
Access the most recent version at doi:10.1110/ps.8.2.361

Protein Science (1999), *8*:361–369. Cambridge University Press. Printed in the USA.

Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes

MARCOS R. BETANCOURT[1] AND D. THIRUMALAI[1,2]

[1] Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742
[2] Department of Chemistry and Biochemistry, University of Maryland, College Park, Maryland 20742

(RECEIVED July 1, 1998; ACCEPTED October 19, 1998)

Abstract

We examine the similarities and differences between two widely used knowledge-based potentials, which are expressed as contact matrices (consisting of 210 elements) that gives a scale for interaction energies between the naturally occurring amino acid residues. These are the Miyazawa–Jernigan contact interaction matrix **M** and the potential matrix **S** derived by Skolnick J et al., 1997, *Protein Sci 6*:676–688. Although the correlation between the two matrices is good, there is a relatively large dispersion between the elements. We show that when Thr is chosen as a reference solvent within the Miyazawa and Jernigan scheme, the dispersion between the **M** and **S** matrices is reduced. The resulting interaction matrix **B** gives hydrophobicities that are in very good agreement with experiment. The small dispersion between the **S** and **B** matrices, which arises due to differing reference states, is shown to have dramatic effect on the predicted native states of lattice models of proteins. These findings and other arguments are used to suggest that for reliable predictions of protein structures, pairwise additive potentials are not sufficient. We also establish that optimized protein sequences can tolerate relatively large random errors in the pair potentials. We conjecture that three body interaction may be needed to predict the folds of proteins in a reliable manner.

Keywords: inter-residue pair potentials; native state sensitivity to perturbations; protein folding; protein stability; quasi-chemical approximation

The prediction of the three-dimensional structure of proteins starting from the primary sequence requires fairly accurate estimates of energy functions, which describe the interactions between the amino acid residues. The effective interactions between the amino acids are assumed to be the same as the potential of mean force between the moieties, so that the effects due to the solvent degrees of freedom are only implicitly included. It is clear that even the computation of the potential of mean force starting from atomic detailed description is difficult. The inherent complexity in computing the residue–residue interactions has prompted a search for a simple, but realistic, representations of the free energy potentials that implicitly include solvent effects (Levitt, 1976; Miyazawa & Jernigan, 1985, 1996; Godzik et al., 1995, 1996; Skolnick et al., 1997).

A key assumption in most of the schemes used in the literature is that the "exact" free energy potential for any protein may be written as a sum of pairwise interactions involving all the residues. It is not a priori clear that such an approximation can adequately describe cooperative effects, which are known to be important in the folding of proteins. Even if this approximation is valid, the calculation of the pair potentials requires additional simplifications. The strategy for computing pair potentials occurs in two steps. First, a protein is represented using a coarse grained description, which usually consists of α-carbons representation with the side chains being confined to their centers of mass. This provides low resolution folds of proteins. With this simpler representation of proteins, the calculation of pairwise interaction potentials between the residues boils down to deciphering the interactions that are consistent with a given set of folds.

The above stated inverse problem is approximately solved using as much knowledge about a given database of folds as possible. The resulting interactions potentials are referred to as statistical potentials. In most of the statistical potentials, the interactions between residues are presumed to exist only if the distance between the centers of mass of the side chains are less than a certain cut-off distance. This method of obtaining contact potentials from a database of proteins was pioneered by Tanaka and Scheraga (1976), and later refined by Miyazawa and Jernigan (1985, 1996). There are several variations on these basic principles that have lead to many different interaction schemes. The statistical potentials have now been used in many different contexts (Sun, 1993; Sippl, 1995; Jernigan & Bahar, 1996; Mirny & Shakhnovich, 1996; Moult, 1997).

Reprint requests to: D. Thirumalai, Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742; e-mail: thirum@glue.umd.edu.

The basic assumption of the statistical potentials is that an empirical knowledge-based potential that estimates the effective residue–residue interaction may be made from the contact frequencies of the residues in the database of folds. Arguably, the best known and widely used statistical potentials are due to Miyazawa and Jernigan (1985, 1996), Godzik et al. (1995, 1996), and Skolnick et al. (1997). To extract the contact potentials, Miyazawa and Jernigan (MJ) employed the quasi-chemical approximation, i.e., the residues are assumed to be in equilibrium with the solvent. The interaction between residue type *i* and *j* can be thought of as occurring in two steps. The first step involves the desolvation that gives the reversible work required to expel a solvent molecule in contact with a given residue. In the second step, mixing between residue *i* and *j* takes place. This procedure, which uses different reference states in the two steps, takes into account the effects of excluded volume due to solvent molecules. The main approximation in this method is that it ignores chain connectivity. The justification for this is that if we consider a large ensemble of proteins in the database then the bias introduced by chain connectivity is averaged out. Schematically, one can represent the procedure leading to the MJ potentials as a set of chemical equilibria, namely (Miyazawa & Jernigan, 1985; Thomas & Dill, 1996),

$$2(i)(s) \rightleftarrows (i)(i) + (s)(s) \tag{1A}$$

$$2(j)(s) \rightleftarrows (j)(j) + (s)(s) \tag{1B}$$

$$(i)(i) + (j)(j) \rightleftarrows 2(i)(j) \tag{1C}$$

where *i* and *j* refer to residue types, *s* refers to the solvent and $(\alpha)(\beta)$ means that species α and β are in contact, i.e., the distance between their respective center of mass is less than a cut-off value. From the charging procedure given in Equation 1, the excess energy due to contact (the reversible work required to bring *i* and *j* into contact) is given by

$$M_{ij} = E_{ij} + E_{00} - E_{i0} - E_{j0}, \tag{2}$$

where $E_{\alpha\beta}$ is the contact energy between the moieties labeled α and β and the subscript 0 refers to the solvent. The various elements of the matrix **M** are calculated using the Bethe approximation in which the reference system is one in which there is equilibrium between the solvent and the residues, i.e., the random mixing approximation is used.

One of the obvious limitations of the MJ procedure, which was already noted in the original paper (Miyazawa & Jernigan, 1985), is that it ignores entropy losses due to chain connectivity. More recently, Skolnick, Jaroszewski, Kolinski, and Godzik (SJKG) (Skolnick et al., 1997) introduced forms of knowledge-based pair-potentials for amino acids that explicitly includes effects due to chain connectivity, compactness of the native state and the effects of secondary structures. Surprisingly, they concluded that the quasi-chemical approximation, which ignores all these effects, is in general sufficient for extracting pair-potentials. This is further corroborated by comparing the MJ and the SJKG potentials.

Let the matrix **S** represent the potentials due to SJKG. The elements of **S** correspond to Table 3A of SJKG. We find that the matrices **M** (upper half of Table 3, Miyazawa & Jernigan, 1996) and **S** are well correlated, with the correlation coefficient being 0.82. However, when the elements of the two matrices are compared in detail, we find remarkable differences in the magnitude and signs of several terms. Recall that in both instances the value of the interaction potential is in terms of *RT* units where *R* is the gas constant and *T* is the absolute temperature. The values of M_{ij} are mostly negative even for interactions between like-charged residues. Such a discrepancy cannot be resolved by merely adding a scalar quantity because it would be inconsistent with the underlying approximations described by the equilibrium relations given by Equation 1. The values of the elements of the **S** matrix, on the other hand, seem to have more physically reasonable values.

These observations lead to two questions: (1) Can one choose a reference system within the MJ scheme so that the dispersion between the resulting potential and the SJKG potential is considerably reduced? (2) How sensitive are the predicted native structures to variations in the interaction potentials? In this paper we address these two questions. The answer to the first question is affirmative. To address the second question, we use lattice models to establish that variations in the interaction schemes can lead to substantial changes in the energies and topology of the native states.

Results

Rescaled Miyazawa–Jernigan potential

The overall correlation between the matrix **M** (MJ potential) and the matrix **S** (SJKG potential) is very good (see Fig. 1). The optimal linear relation between the two is

$$S_{ij} \simeq 1.26 + 0.37 M_{ij}. \tag{3}$$

The quality of the fit can be assessed by computing the dispersion between the fit and the matrix **S** and the correlation coefficient, *r*. We find the dispersion to be 0.39 and $r = 0.82$. In explicitly using these contact matrix elements in the prediction of structures, the fluctuations between the two matrices are relevant. An assessment of such fluctuations requires comparing the two matrices. A direct

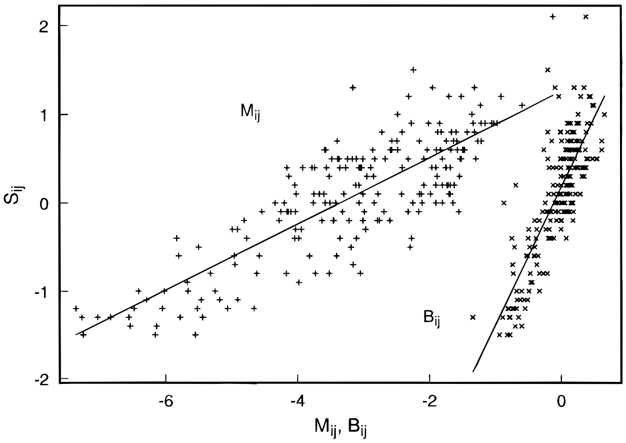

Fig. 1. Correlation between the **M** and **S** and the **S** and **B** matrix elements. For the **M** matrix, the correlation coefficient is $r = 0.82$ and the relative dispersion δ (see Equation 4) between the two models is $\delta = 3.40$. For **B**, $r = 0.82$ and the dispersion δ is only 0.45.

measure of the relative difference between **S** and **M** may be obtained by computing

$$\delta = \sqrt{\frac{1}{210}\sum_{ij}(S_{ij} - M_{ij})^2} \qquad (4)$$

which is found to have a relatively large value of 3.4. The large dispersion arises because contact interactions between certain residues are substantially different in the two schemes. As has been noted already by Godzik et al. (1996), the principal difference between the numerous parameter sets is due to the variations in the reference states in computing the interaction schemes. As described in Equation 1, MJ use the random mixing approximation to calculate the elements M_{ij}. A key difficulty in computing M_{ij} is in estimating E_{00} and E_{i0}, which require the average number of solvent-solvent contacts and residue-solvent contacts, respectively. In contrast, SJKG obtained their parameter set using native reference states, which consists of a library of real structures whose compactness is comparable to the overall native fold. Following the spirit of Skolnick et al. (1997), we chose a different reference state within the MJ procedure, which in effect brings the two parameter sets into closer agreement with respect to each other in the sense of reducing δ (see Equation 4).

We chose the reference state in which the solvent molecule (see Equation 1) is replaced by one of the nonpolar amino acid residues. The reference residue type that replaces the solvent molecule is chosen so that the nature of interaction between this residue and others is "similar" to the interaction between the solvent and the other residues. The reference residue is obtained by requiring that the hydrophobicities of the residues, predicted by the resulting potential, correlate as closely as possible with the experimental values. Let t be such a reference residue. Then, we can define a matrix **B** whose elements are

$$B_{ij} = M_{ij} + M_{tt} - M_{ti} - M_{tj}, \qquad (5)$$

for $i, j = 1, 2, 3, \ldots, 20$. If 0 represents the solvent, then it follows from Equation 5 that $M_{i0} = 0.0$ for $i = 0, 1, 2, \ldots, 20$. Since t is the reference state with respect to which all interactions are measured, we set $B_{i0} = B_{0i} = 0$ for $i = 0, 1, 2, \ldots, 20$. However, we get from Equation 5 that $B_{i0} = M_{tt} - M_{ti}$. It is hoped that a reference residue type t may be found so that our assumption that $B_{i0} = 0$ is followed as well as possible, and that also leads to good correlations with experimental hydrophobicities. The advantage of the parameter set B_{ij} is that it eliminates the need for evaluating E_{00} and E_{i0} while utilizing the desirable features of the successful MJ potential.

We first show that a residue type t is chosen so that the hydrophobicity associated with residue i

$$h_i = B_{ii}/2 \qquad (6)$$

correlates well with the experimental estimates. The experimental hydrophobicities for the various residues are typically inferred from transfer experiments. We compare h_i, obtained using different reference residues, to three different hydrophobicity scales and their averages. The correlation coefficients are presented in Table 1. For comparison we also give the results for the **S** and **M** matrices. The correlation coefficients were calculated by excluding Cys. It is clear from Table 1 that the highest correlation is obtained when t = Thr for all the hydrophobicity scales. The maximum correlation results between h_i and the hydrophobicities reported by Fauchere and Pliska (1983).

Table 1. *Correlation coefficients with experimental side-chain hydrophobicities for different effective potential models*[a]

	a	b	c	d
Thr	0.86	0.92	0.88	0.90
Asn	0.76	0.85	0.84	0.82
Ser	0.79	0.86	0.84	0.84
Gln	0.69	0.73	0.73	0.73
Gly	0.47	0.62	0.52	0.53
His	0.41	0.49	0.53	0.49
S	0.73	0.80	0.65	0.73
M	0.79	0.91	0.84	0.85

[a] The first column gives the reference residue used in the calculation of the rescaled MJ matrix (Equation 5). The last two rows are the correlation coefficients for the **S** and **M** matrices. Column a is for amino acids hydrophobicities some of which were measured by Nozaki and Tanford (1971) and others by Levitt (1976). Column b is for N-acetil amino acid amide hydrophobicities measured by Fauchere and Pliska (1983). Column c correspond to hydrophobicities calculated from individual groups that make up each side chain, obtained by Roseman (1988). Column d is obtained from the averages of the hydrophobicities in (a), (b), and (c). The highest correlations are obtained for the rescaled MJ parameters using Thr as the reference residue (defined as matrix **B**).

In Figure 2 we plot the experimental hydrophobicities (corresponding to column b and t = Thr in Table 1) vs. h_i. With the exception of Cys, the correlation between the computed values and experiments is very good when Thr is chosen as the reference state. From the effective hydrophobicities it is clear that the amino acids divide themselves into three classes. They are: (1) strongly hydrophobic (Cys, Phe, Leu, Trp, Val, Ile, Met); (2) mildly hydrophobic (His, Tyr, Ala, Gly, Pro); (3) hydrophilic (Asn, Thr, Ser, Arg, Gln, Asp, Lys, Glu). It is, perhaps, useful to further classify class (2) residues as weakly hydrophobic (Tyr, Ala, Pro) or weakly hydrophilic (His, Gly).

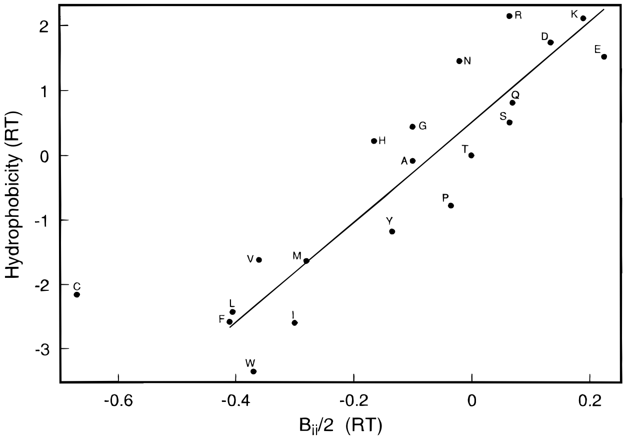

Fig. 2. Correlation of hydrophobicies between the calculated values using the **B** matrix and the experimental values given by Fauchere and Pliska (1983). The solid line is obtained from a linear regression (excluding Cys) with slope = 7.75, intercept = 0.51, and correlation coefficient = 0.92.

The parameters for the potential introduced here (see Equation 5) is given in Table 2. In addition to the contact energies B_{ij}, Table 2 shows the demixing terms in the lower triangular portion, and the average values of the interactions and their standard deviations. The demixing term (referred to as excess pair interaction term by Skolnick et al.) is

$$X_{ij} = B_{ij} - \tfrac{1}{2}(B_{ii} + B_{jj}) = M_{ij} - \tfrac{1}{2}(M_{ii} + M_{jj}). \qquad (7)$$

The demixing terms X_{ij} measure the tendency of a pair of residues to disassociate when positive, or to associate when negative. By inspecting B, the expected qualitative features of the interactions are evident. With the exceptions of minor fluctuations for few matrix elements, in particular those of Trp with hydrophilic residues, the signs of the hydrophobic interactions are physically reasonable. So are the signs of charge group interactions, i.e., repulsive between charges of equal sign and attractive between between charges of opposite sign. Thus, the elements of the B matrix introduced here provide estimates for contact interactions that are consistent with known classification of the nature of the amino acids.

Nature of contact interactions: Justification of the HP model

To get additional insights into the nature of the residue–residue interactions, we have divided most of the residues into four groups: hydrophobic (H) [Phe, Leu, Trp, Val, Ile, Met, Tyr, Ala]; hydrophilic (P) [Asn, Thr, Ser, Gln]; negatively charged (−) [Asp, Glu], and positively charged (+) [Arg, Lys]. The rest of the residues [Cys, His, Gly, Pro] cannot be easily classified into the previous four groups, at least according to the **B** matrix elements, and are left out for simplicity. The average effective interactions and demixing energies for these groups are shown in Table 3. There is a strong attraction between hydrophobic residues HH, a weak repulsion between hydrophobic and hydrophilic residues HP, and an almost zero interaction between hydrophilic residues PP. This result is in agreement with other hydrophobic models in which hydrophobic residues are effectively attractive and hydrophilic residues are neutral as the solvent. This classification naturally shows that the HP model, advocated by Chan and Dill (1990), is a reasonable starting point for understanding general aspects of protein folding. Li et al. (1997) have also arrived at a similar conclusion based on their analysis of the MJ potential. Inspection of the matrix elements B_{ij} shows that there are large fluctuations between hydrophobic and hydrophilic residues (especially those with intermediate hydrophobicity). In the folding structures, the residues with intermediate hydrophobicity could be at the interface between the hydrophobic core and the hydrophilic residues that are in contact with water. The Coulomb interactions between residues with opposite charges are considerably stronger than the repulsion between like charges. The demixing energies (lower triangular partition of Table 3) are also physically meaningful. Residues of similar hydrophobicity are mixing on an average ($X_{HH} \approx X_{PP} \approx 0$), while residues of different hydrophobicity types tend to segregate ($X_{HP} > 0$). For charged groups, like charged residues are slightly demixing ($X_{++} \approx X_{--} \gtrsim 0$) while mixing is strongly favored by opposite charged groups. These classification of residue–residue interactions shows that while the simple HP model is indeed a good starting point for certain purposes, the inclusion of diverse interactions that occur in natural amino acids is necessary for a semi-quantitative description of protein folding.

Table 2. *The elements of the* **B** *matrix, which were obtained by rescaling the MJ potential*[a]

	Cys	Phe	Leu	Trp	Val	Ile	Met	His	Tyr	Ala	Gly	Pro	Asn	Thr	Ser	Arg	Gln	Asp	Lys	Glu
Cys	$\underline{-1.34}$	−0.53	−0.50	−0.74	−0.51	−0.48	−0.49	−0.19	−0.16	−0.26	−0.09	−0.18	0.28	0.00	0.09	0.32	0.04	0.38	0.35	0.46
Phe	0.55	$\underline{-0.82}$	−0.78	−0.78	−0.67	−0.65	−0.89	−0.19	−0.49	−0.33	0.11	−0.19	0.29	0.00	0.10	0.08	−0.04	0.48	0.11	0.34
Leu	0.57	0.04	$\underline{-0.81}$	−0.70	−0.80	−0.79	−0.68	0.10	−0.44	−0.37	0.14	−0.08	0.36	0.00	0.26	0.09	0.08	0.62	0.16	0.37
Trp	0.30	−0.00	0.08	$\underline{-0.74}$	−0.62	−0.65	−0.94	−0.46	−0.55	−0.40	−0.24	−0.73	−0.09	0.00	0.07	−0.41	−0.11	0.06	−0.28	−0.15
Val	0.52	0.10	−0.04	0.11	$\underline{-0.72}$	−0.68	−0.47	0.18	−0.27	−0.38	0.04	−0.08	0.39	0.00	0.25	0.17	0.17	0.66	0.16	0.41
Ile	0.49	0.06	−0.08	0.02	−0.02	$\underline{-0.60}$	−0.60	0.19	−0.33	−0.35	0.21	0.05	0.55	0.00	0.35	0.18	0.14	0.54	0.21	0.38
Met	0.46	−0.20	0.00	−0.29	0.17	−0.02	$\underline{-0.56}$	−0.17	−0.51	−0.23	0.08	−0.16	0.32	0.00	0.32	0.17	−0.01	0.62	0.22	0.24
His	0.64	0.38	0.67	0.08	0.70	0.66	0.28	$\underline{-0.33}$	−0.21	0.21	0.23	−0.05	0.10	0.00	0.15	0.04	0.22	−0.22	0.26	−0.11
Tyr	0.64	0.05	0.10	−0.04	0.22	0.10	−0.10	0.09	$\underline{-0.27}$	−0.15	−0.04	−0.40	0.01	0.00	0.07	−0.37	−0.18	−0.07	−0.40	−0.16
Ala	0.51	0.18	0.14	0.07	0.08	0.05	0.15	0.48	0.08	$\underline{-0.20}$	−0.03	0.07	0.24	0.00	0.15	0.27	0.21	0.30	0.20	0.43
Gly	0.68	0.62	0.65	0.23	0.50	0.61	0.46	0.50	0.20	0.17	$\underline{-0.20}$	−0.01	0.10	0.00	0.10	0.14	0.20	0.17	0.12	0.48
Pro	0.53	0.25	0.36	−0.32	0.31	0.38	0.16	0.15	−0.23	0.20	0.12	$\underline{-0.07}$	0.13	0.00	0.17	−0.02	−0.05	0.25	0.12	0.26
Asn	0.97	0.72	0.78	0.30	0.77	0.87	0.62	0.28	0.17	0.36	0.22	0.18	$\underline{-0.04}$	0.00	0.14	0.02	−0.05	−0.12	−0.14	−0.01
Thr	0.67	0.41	0.41	0.37	0.36	0.30	0.28	0.17	0.14	0.10	0.10	0.04	0.02	$\underline{0.00}$	0.00	0.00	0.00	0.00	0.00	0.00
Ser	0.70	0.44	0.60	0.37	0.54	0.58	0.53	0.25	0.14	0.18	0.13	0.14	0.09	−0.06	$\underline{0.13}$	0.12	0.25	0.01	0.10	0.10
Arg	0.93	0.42	0.43	−0.11	0.46	0.41	0.39	0.14	−0.30	0.30	0.18	−0.05	−0.02	−0.06	−0.01	$\underline{0.13}$	−0.12	−0.71	0.50	−0.75
Gln	0.64	0.30	0.42	0.19	0.46	0.37	0.20	0.32	−0.11	0.24	0.23	−0.08	−0.10	−0.07	0.11	−0.25	$\underline{0.14}$	0.12	−0.20	0.10
Asp	0.92	0.75	0.89	0.30	0.89	0.70	0.77	−0.19	−0.07	0.26	0.14	0.15	−0.24	−0.14	−0.19	−0.91	−0.08	$\underline{0.27}$	−0.69	0.40
Lys	0.83	0.33	0.38	−0.10	0.33	0.32	0.31	0.23	−0.46	0.11	0.03	−0.03	−0.31	−0.19	−0.15	0.25	−0.46	−1.01	$\underline{0.38}$	−0.87
Glu	0.90	0.53	0.55	−0.01	0.55	0.45	0.30	−0.17	−0.25	0.30	0.36	0.07	−0.21	−0.22	−0.19	−1.04	−0.20	0.04	−1.28	$\underline{0.45}$
ave	−0.18	−0.24	−0.19	−0.42	−0.14	−0.12	−0.19	−0.01	−0.25	−0.03	0.08	−0.05	0.12	0.00	0.15	−0.01	0.05	0.15	0.02	0.12
dev	0.44	0.42	0.46	0.31	0.43	0.44	0.43	0.21	0.18	0.26	0.16	0.22	0.19	0.00	0.09	0.31	0.13	0.38	0.34	0.37

[a] The diagonal and upper elements are the transformed terms, i.e., $B_{ij} = M_{ij} + M_{tt} - M_{ti} - M_{tj}$. The reference residue is t = Thr. Below the diagonal are the demixing terms, i.e., $X_{ij} = B_{ij} - (B_{ii} + B_{jj})/2$, which are identical to the ones obtained from the **M** matrix. The interactions with the solvent, i.e., B_{i0}, are zero for $i = 0,1,\ldots,20$ where 0 represents the solvent. The last two lines are the averages and the standard deviations of the **B** elements for each residue with respect to all others.

Table 3. *Average values of **B** obtained by dividing the amino acids into hydrophobic H, hydrophilic P, negatively charged − and positively charged + groups*[a]

	H	P	−	+	
	$\underline{-0.56}$	0.12	0.32	0.03	H
H	0.03	$\underline{0.06}$	0.02	−0.03	P
P	0.39	−0.00	$\underline{0.38}$	−0.76	−
−	0.43	−0.18	0.02	$\underline{0.38}$	+
+	0.20	−0.18	−1.06	0.12	
	H	P	−	+	

[a] The hydrophobic residues are composed of (Phe, Leu, Trp, Val, Ile, Met, Tyr, Ala); the hydrophilic of (Asn, Thr, Ser, Gln); the negative charged groups of (Asp, Glu) and the positive charged groups of (Arg, Lys). The other residues are not included because of their marginal or ambiguous properties.

In summary, it is clear that a viable contact interaction parameter set, without explicitly referring to the solvent state, can be constructed within the MJ scheme. The resulting interaction potential has a high degree of correlation with the experimental hydrophobicities. There is an obvious discrepancy between h_i and the experimental values, namely, the computed h_i are nearly a factor eight smaller. There are three related reasons for this discrepancy: (1) In principle, hydrophobicity refers to the reversible work required to transfer a residue i in a solvent made up of the residue itself or an inert solvent to water. The virtue of the MJ interaction scheme is that $M_{ii}/2$ describes the result of this process (see Equation 1). Here we have compared h_i to data obtained from experiments that involve transfer of residues from a reference solvent to water. Our calculated values of hydrophobicities refers to a situation in which residues initially in equilibrium with a solvent made up of Thr are transferred to water. In the transfer experiment the reference state corresponds to a solvent different from water. So we do not expect absolute correspondence with experimental measurements. (2) The database of folded structures, from which the frequencies of pairing of various residues are computed, correspond to compact structures. In such structures certain residues may occur with higher probability due to chain connectivity than would normally be the case. Therefore, choosing random polymer reference state could decrease the interaction energy of hydrophilic residues and increase that of the hydrophobic ones. Both, the **B** and the **S** matrix calculation scheme reduce the probability of such contacts by eliminating the use of the random polymer reference state. However, the magnitudes of the interactions could be underestimated, which affects the scale of the interactions but not their character (sign) with relation to the solvent. (3) In our scheme, the transformation in Equation 1 effectively eliminates explicit reference to solvent. The matrix **B** is computed using only the frequencies of contact between the residues in the folded native state whose average structure does not fluctuate much below the folding transition temperature. The fundamental assumption of the statistical potential is that the frequencies of occurrence of various pairs of residues obeys Boltzmann statistics at a constant temperature. A plausible rationale for this has been proposed by Finkelstein et al. (1995) using the random energy model. More recently, Zhang and Skolnick (1998) have also given conditions when "true" potentials may be derived from a database of structures. However, a more precise test of this assumption using a database of 346 PDB structures (Thomas & Dill, 1996) shows that the Boltzmann distribution is not obeyed. In fact, the extracted temperatures can vary considerably (Thomas & Dill, 1996). Therefore, it is not surprising that the slope of the correlation between h_i and the experimental hydrophobicities can deviate from unity.

Approximations of the interaction matrices

The natural classification of the amino acid residues into a few subclasses based on the interaction schemes suggest that the 20 × 20 matrix consisting of 210 elements may be adequately described by a smaller number of parameters. Based on eigenvalue decomposition of the MJ matrix, Li et al. (1997) proposed that the matrix **M** can be defined from the hydrophobicity alone. They found that **M** has two dominant eigenvalues, and the corresponding eigenvectors were strongly correlated. Thus, one can express one of the eigenvectors in terms of the other. These observations suggest that a suitable HP model suffices to describe the main features of residue–residue interactions. Based on this, Li et al. (1997) showed that the **M** matrix can be approximated by

$$M'_{ij} = h'_i + h'_j - C_2\,(u_i - u_j)^2/2 \tag{8A}$$

and

$$h'_i = C_0/2 + C_1 u_i + (C_2/2)u_i^2, \tag{8B}$$

where the u_i's are the components of the eigenvector corresponding to one of the dominant eigenvalues and C_0, C_1, and C_2 are constants which are obtained from the dominant eigenvalues and the linear relation between the corresponding eigenvectors. In Equation 8, $\mathbf{h}'$ are the approximate effective hydrophobicities $h'_i = M'_{ii}/2$. The last term in Equation 8A is a model for the demixing term, i.e., $X'_{ij} = -(C_2/2)(u_i - u_j)^2$. When compared to the mixing energy obtained from Hildebrand's solubility theory, the u_i can be related to the vaporization energies (Li et al., 1997). The question that arises is whether the same conclusions can be drawn from the eigenvalue decomposition of the **B** and **S** matrices. We find that for the the **B** and **S** matrices, the spectrum of eigenvalues does not significantly separate as in **M**. Furthermore, the eigenvectors of the two largest eigenvalues are not correlated. This implies that a simple reduction of the 20 × 20 matrix into a generalized HP model is not always possible. This conclusion is in accord with the recent calculations of Du et al. (1998) and R.L. Jernigan (pers. comm.).

It follows from Equations 5 and 8 that a corresponding matrix $\mathbf{B}'$, which should be a reasonable approximation to **B**, is $B'_{ij} = M'_{ij} + M'_{tt} - M'_{ti} - M'_{tj}$. We have eliminated the cysteine and charged group residues from the analysis because they are not well represented by the demixing term in Equation 8. The elements of **B** can be written as

$$B'_{ij} = h'_i + h'_j - K_2\,(u_i - u_j)^2/2 \tag{9A}$$

and

$$\begin{aligned} h'_i &= \frac{K_0}{2} + K_1 u_i + \frac{K_2}{2}\,u_i^2 \\ &= \frac{K_2}{2}\,(u_i - u_t)^2, \end{aligned} \tag{9B}$$

where h' is redefined as $h'_i = B'_{ii}/2$ and $K_0 = -0.193$, $K_1 = 1.48$ and $K_2 = -11.13$. The correlation between the elements of **B** and **B**$'$ ($r = 0.85$) is significantly lower than found between **M** and **M**$'$. In addition, the distributions of points is somewhat asymmetric with respect to the **B**$'$ = **B** line indicating that Equation 9 is not fully adequate to represent the interaction matrix. This is also apparent from Equation 9B, which always gives a zero or negative hydrophobicity because $K_2 = C_2 < 0$. Therefore, residues with positive hydrophobicity cannot be faithfully represented by this equation. A better fit may be obtained by expressing the equation for B' as a higher order expansion in u_i.

To understand the decrease in correlation between **B** and **B**$'$ compared to **M** and **M**$'$, we decompose the **B** matrix into the hydrophobic components $H_{ij} = (B_{ii} + B_{jj})/2$ and a demixing contribution $X_{ij} = B_{ij} - H_{ij}$. Both the **H** and **X** matrices can be approximated as $H'_{ij} = h'_i + h'_j$ and $X'_{ij} = -(K_2/2)(u_i - u_j)^2$. Such a decomposition can also be done for the **M** matrix. The correlation between H_{ij} and H'_{ij} is found to be somewhat worse than the corresponding relationship between the hydrophobic components extracted from the **M** matrix. Furthermore, the correlation between X_{ij} and X'_{ij} is even smaller than between H_{ij} and H'_{ij}. While the **X** matrix remains the same for both **B** and **M**, the average magnitude of the **H** matrix elements obtained for the **B** matrix is much smaller than the one obtained for the **M** matrix. Therefore, because the demixing terms in the **B** matrix play a more important role than in the **M** matrix, the correlation between **B** and **B**$'$ is decreased. The high correlation seen between **M** and **M**$'$, which allows the description of the 20 × 20 matrix into 23 parameters, is due to the reference state used by Miyazawa and Jernigan (1985, 1996).

Reduced representation of the **B** *matrix*

It appears that the approximation to the **M** matrix found by Li et al. (1997) is due to the fact that in the MJ scheme the hydrophobic interactions are overemphasized. In the other schemes, such as the **S** matrix and the **B** matrix introduced here, there is no separation in the eigenvalues. Nevertheless, it is possible to obtain a representation for the contact matrix that also takes into account charged residues and Cys. The charged residues can be included using the Coulomb potential proportional to q_iq_j, where $q_i = \{0, \pm 1\}$ is the charge of residue i. The repulsion between like-charged groups can be absorbed in the hydrophobic potential yielding an effective Coulombic potential proportional to $(q_i - q_j)^2$. The Cys-Cys interactions are modeled by adding an additional constant parameter, which is defined as $D = E_{CC} - 2h'_C$, where C stand for cysteine. If we let $v_i = u_i - u_t$ (u_t is the value of the eigenvector **u** for the residue Thr), the expansion of **B**$'$ and **h**$'$ to third order in q_i can be written as

$$B'_{ij} = h'_i + h'_j - A(v_i - v_j)^2 + Q(q_i - q_j)^2 + D\delta_{i,C}\delta_{j,C}, \qquad (10A)$$

where the contribution of the term proportional to $(v_i + v_j)(v_i - v_j)^2$ is negligible, and

$$h'_i = R_1 v_i + R_2 v_i^2 + R_3 v_i^3. \qquad (10B)$$

The new coefficients along with the vector components v_i are obtained by fitting the parameters to the matrix values $\mathbf{B}_{ij}$. The fitting is carried out by a steepest descent method. Figure 3 shows the correlation between **B**$'$ and **B**. The model is a fair representation of the potential although the parameters obtained from the steepest descent are perhaps not optimal. The values for the parameters are $A = -2.661$, $R_1 = -0.961$, $R_2 = -2.023$, $R_3 = 5.594$, $Q = -0.195$, and $D = -0.535$. If we set $A = R_2$ (in analogy to Equations 8 and 9) the result remain practically the same. The vector **v** is **v**(Cys, Phe, Leu, Trp, Val, Ile, Met, His, Tyr, Ala, Gly, Pro, Asn, Thr, Ser, Arg, Gln, Asp, Lys, Glu) = (0.37, 0.45, 0.46, 0.28, 0.31, 0.40, 0.32, 0.06, 0.20, 0.08, −0.02, −0.00, −0.08, 0.00, −0.06, −0.10, −0.04, −0.24, −0.12, −0.17). The correlation coefficient is $r = 0.870$ and without considering charged residues or cysteine it is $r = 0.888$, which is higher than the one obtained by Equation 9 ($r = 0.854$). It is possible that even better correlation between **B**$'$ and **B** can be obtained by optimizing the charges q_i. Thus, it is clear that the reduced representation of the pair potential cannot always be accomplished using Equation 9, which considers only the dominant eigenvalues.

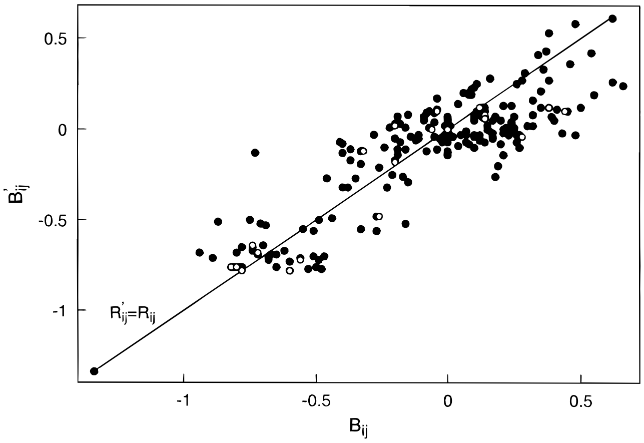

Fig. 3. Correlation between the **B**$'$ (see Equation 10) and the **B** matrix elements. The correlation coefficient is $r = 0.870$ and the dispersion between the **B** and **B**$'$ is $\delta = 0.184$. The white circles are for the diagonal elements and the black ones for all others. All 20 amino acid interactions are included. This figure shows that a reduced representation of **B** may be computed by systematic expansion using the eigenvectors of **B**.

Sensitivity of native state to variations in the interaction scheme

The preceding sections show that, despite the differences in the interaction matrices, the hydrophobicities extracted from them show very good agreement with experiments. Despite the overall similarity between **B** and **S** matrices (they correlate well with each other and the relative dispersion is small), there are differences in the magnitude of several matrix elements describing contact interactions between the residues. These differences are systematic because there is correlation between the two matrices. The correlation arise because in both the schemes, the database of folded proteins is utilized to compute the interaction matrices. It is a priori difficult to assess the effect the differences in the various matrix elements have on the structure and energy of the native states for a given sequence. More generally, we can ask the following question. If the interaction parameters given by the **B** matrix were exact, then, what effect would substituting the **S** matrix have on the predicted native state? This question is related to the issue of how accurate should interaction potentials be so that the folded structure can be predicted to the desired accuracy (Bryngelson, 1994; Pande et al., 1995).

The question posed above can be precisely answered using lattice models of proteins. The strategy we adopt is the following: We have generated five optimized sequences using the **B** matrix interaction parameters. Once the native states for each sequence are determined, we switch the interaction matrix to **S** keeping the sequences identical. The native states are redetermined, and a comparison of the various native states allows us to assess the dependence of topology and energy on the interaction scheme used.

We model proteins as chains of N $(= 36)$ successively connected beads on the sites of a cubic lattice. The energy of a conformation specified by the sites on the lattice $r_i (i = 1, 2, \dots, N)$ is

$$E(\{r_i\}) = \sum \Delta_{ij}\,\delta(r_{ij} - a) \tag{11}$$

where the contact interactions Δ_{ij} is B_{ij} or S_{ij}, $r_{ij} = |r_i - r_j|$, a is the lattice spacing, and $\delta(0) = 1$ and 0 otherwise. Using $\Delta_{ij} = B_{ij}$ we generated five optimized sequences. We determined the thermodynamics of these sequences using the multi-histogram technique (Ferrenberg & Swendsen, 1989). In particular, the collapse (T_θ) and folding transition (T_f) temperatures were determined using standard methods (Camacho & Thirumalai, 1993). For all five sequences $\sigma_T = (T_\theta - T_f)/T_\theta \approx 0$, and hence these are expected to fold thermodynamically and kinetically in a two state manner (Camacho & Thirumalai, 1993). The five sequences were obtained by a design algorithm that efficiently leads to a smooth landscape so that the values of σ are minimized (M. Betancourt & D. Thirumalai, unpubl. results). They were generated in the course of unraveling thermodynamic factors that also determine the kinetic accessibility of the native states. The sequences and associated properties are displayed in Table 4.

The native state for one of these sequences is shown in Figure 4. This structure is maximally compact and is confined to 3 × 3 × 4 sites on the cubic lattice. We now keep the identity of the sequence, and let $\Delta_{ij} = S_{ij}$. The topology of the resulting native state is shown in Figure 5. A comparison of the two native states shows that the topology of the native state has been considerably altered. The native conformation for the **S** matrix is no longer maximally compact. Approximately 25% of the tertiary contacts have been altered. Similar results are found for other sequences as can be seen from Table 4. In particular, the differences in the fraction of native tertiary contacts for other optimized sequences varies from 0.25 to about 0.60. It can be seen from Table 4 that the actual ground state energies for the five sequences using the S_{ij} matrix elements are lower than that computed from the structures generated with B_{ij} interaction scheme. This shows that, at least within pairwise interaction schemes, relatively small (nonrandom) differences in the contact energies can have profound effects on the topology of the native conformation. We should emphasize that this depends on the sequence and the topology of the native state.

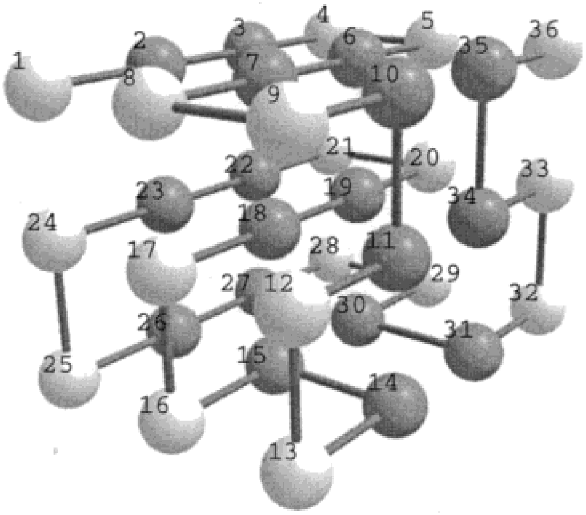

Fig. 4. The native structure of designed sequence #1 using the rescaled MJ potential **B**. Hydrophobic residues are drawn darker than hydrophilic residues. The native state is unique, and has an energy of $E_{ns} = -28.45$ (in units of RT).

We should hasten to point out that the extreme sensitivity of the designed native state to alterations in the interaction potentials may arise because of the underlying lattice. It possible that if one uses a high coordination lattice, the response of the native state upon switching the potentials form **B** to **S** may be less severe.

Tolerance of optimized sequences to random perturbations

It is clear that even if pair potentials are very accurate for structure prediction they cannot be determined with very high precision.

Table 4. *Designed sequences for five 36-mers using the **B** matrix with some of their thermodynamic properties*[a]

#	Sequence	E_{ns}^B	E_{ts}^S	E_{ns}^S	Q
1	KMIKDVIERACDHCMHKFVKDVMEHMIKDVCKDCAK	−28.45	−24.3	−26.1	0.25
2	KKLPMHLRKDEILKKDDVCCIRKDEICPMKKDEIWC	−30.51	−20.6	−22.2	0.28
3	EICGHERYDKLWCEKHGCVGHEKWLKDYRREWVKQL	−26.03	−20.8	−22.7	0.22
4	CCDDDDDIFKKKRKCLEKVIAMPMDEDDDPPCIWYK	−29.12	−17.7	−19.2	0.39
5	DMVPADKIFREYKKGDIGEYIRGACPCDKCLEKIYI	−25.12	−19.8	−22.6	0.60

[a] We have used a one letter representation for amino acids. E_{ns}^B is the native state energy of the target (and native) structure using the **B** matrix. All designed sequences with the B matrix produce stable and nondegenerate native states. E_{ts}^S and E_{ns}^S are, respectively, the target and native energies when the **S** matrix is substituted for the **B** matrix, using the same sequences and target structures. Q is the fraction of native contacts, as defined by the target structure, in the native structure obtained using the **S** matrix.

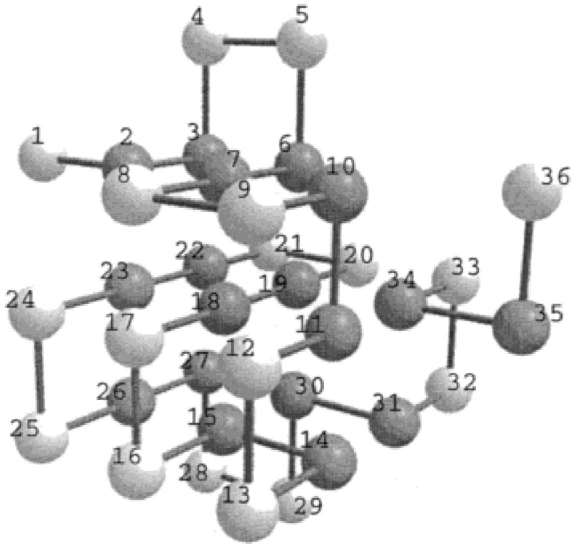

Fig. 5. The native state for the same sequence shown in Figure 4 computed using the **S** matrix. The dark residues are hydrophobic while the white are hydrophilic. This native state is (at least) twofold degenerate. The degeneracy arises because trivial rearrangement of residue 36, for example, does not alter the energy of the native state, $E_{ns} = -26.10$. Most of the contacts (7 of 11) in this structure, which differ from the structure in Figure 4, are Lys-Asp contacts. These contacts have negative energy in the **B** matrix, as expected for oppositely charged residues, but is slightly repulsive in the **S** matrix. A comparison of the structures with the one shown in Figure 4 shows that errors in matrix elements can cause substantial changes in topology.

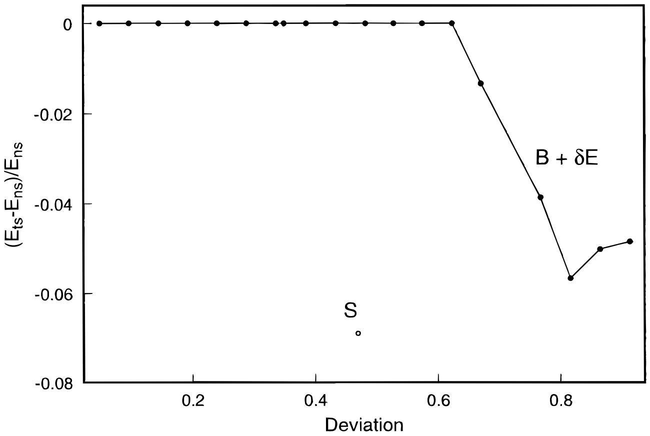

Fig. 6. Fractional difference between the energy of the target native structure shown in Figure 4 and the actual native state, where both energies are obtained by adding Gaussian noise to the matrix **B**. This figure shows that optimized sequences can tolerate considerable random errors before becoming unstable. For comparison, we also show the energy difference between the target native state and the actual native state computed using the **S** matrix.

There are inherent errors even in the most sophisticated interaction schemes. Some of these errors could be systematic, while others can be random. Thus, in the course of predicting the structures of folded states the effect of random errors on the predicted native states have to be examined. This issue is related to question of the accuracy needed in the interaction potentials to determine the native conformations, and to the problem of thermodynamic stability of folded proteins to mutations (Bussemaker et al., 1997).

We used the designed 36-mer sequences with $\Delta_{ij} = B_{ij}$ to examine the effect of random perturbations on the native states. The random perturbations are modeled by adding a term δE_{ij} to each contact term B_{ij}. The random terms are assumed to be distributed as

$$P(\delta E_{ij}) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\delta E_{ij}^2}{2\sigma^2}\right) \qquad (12)$$

where σ is the width of the distribution that gives an estimate of the nonsystematic error in determining the interaction parameters. For each sequence with the new interaction matrix $B_{ij} + \delta E_{ij}$, we recalculated the energy of the native state as a function of σ. In Figure 6 we plot $(E_{ts} - E_{ns})/E_{ns}$ as a function of σ for the sequence shown in Figure 4. Here E_{ns} is the energy of the native state with $\Delta_{ij} = B_{ij} + \delta E_{ij}$, and E_{ts} is the energy obtained for the target structure in Figure 4 using the same interaction matrix. This figure shows that this optimized sequence can tolerate significant errors in the interaction potential while leaving the energy and structure unaltered.

The sensitivity of the structures of the native states to inaccuracies in the interaction matrix was first addressed by Bryngelson (1994). He showed that, for random heteropolymer models, the relative errors in the potentials have to be less than $1/\sqrt{N}$ (N = number of amino acid residues) for reliable predictions. This gives a very stringent requirement in determining pair potentials. It has been suggested by Pande et al. (1995) that optimized sequences can tolerate considerably larger error than random sequences. Furthermore, Bussemaker et al. (1997) have argued that sequences that are optimized become unstable to perturbations (or inaccuracies in the potentials) only if N is relatively large. The numerical results obtained here are consistent with the earlier theoretical studies of Pande et al. (1995) and Bussemaker et al. (1997).

Discussion

In this paper we have examined the schemes used to devise pairwise contact potentials for structure prediction and design of proteins. The popular versions of this method, namely, the Miyazawa–Jernigan statistical potential and those devised by Skolnick and coworkers emphasize different aspects of the residue–residue potentials. While the correlation between the two schemes is good, it is clear that when the elements are compared in detail there are differences. Since there is no mathematically unique way of comparing the similarity of nonrandom matrices, we used the dispersion between the matrix elements as a criterion for assessing the closeness of two matrices. The dispersion between the two matrices (**M** and **S)** is large. We showed that a reference residue could be chosen so that not only does the resulting interaction parameter set, given by the matrix **B**, yields hydrophobicities in accord with experiments but it also reduces the dispersion between the matrix **M** and **S**.

We also addressed the issue of how sensitive are the predicted topologies of the native state to variations in the interaction schemes. Even though the dispersion between **S** and **B** is small we showed, using lattice models as testing ground, that they can have rather dramatic effect on the topology of the native states. Presumably, the inaccuracies in the interaction schemes are not totally random because at some level all of them use the database of proteins for counting the frequencies of occurrence of various contacts. Thus, the systematic errors can lead to substantial differences in the predicted native states. On the other hand, it appears that random errors, which would arise in any interaction scheme, are to a large extent not that important for optimized sequences. This is in accord with general theoretical arguments.

Our results allow us to make general comments regarding pair potentials obtained by knowledge-based schemes. In an illuminating paper, Thomas and Dill (1996) investigated the assumptions in the schemes used to extract statistical potentials by considering a "model PDB" consisting of structures obtained using the HP model. They showed that the extracted energy differed from the true interaction energies. In addition to the interdependence of the HH, HP, and PP energies, the excluded volume interactions alone induces long-range correlations. It is known that even though the range of excluded volume interactions are short, the chain connectivity gives an effective potential (when integrated over the connected residues) that is long ranged (Edwards, 1965). Thus, one cannot expect the extracted energies to coincide with the true energies. Since the energies will be renormalized due to sequence dependence, connectivity, and excluded volume effects, it is useful to wonder if there is an effective 2×2 matrix that has the same structure as the original matrix. The procedure that we have used here (see Equation 5) suggest that there must be a "solvent-like residue" in the HP model with respect to which the renormalized 2×2 interactions has the same structure as the true HP model. To show that this is possible, we consider the case of the 18mer studied by Thomas and Dill (1996). The extracted energies are $e_{HH} \approx -5$, $e_{HP} = e_{PH} \approx -1$, and $e_{PP} \approx 0$ (Thomas & Dill, 1996). Following Equation 5, we considered the rescaled values $e^R_{ij} = e_{ij} + e_{pp} - e_{ip} - e_{jp}$ where the reference residue is P. With this transformation, we obtain $e^R = -3$, and all other elements are zero. This choice of reference state, which accounts for the aforementioned renormalization of the bare potential, reproduces the original HP model up to a multiplicative factor. We, therefore, conclude that the knowledge-based method of computing pair potential, which is not very accurate, may be qualitatively useful. However, the degree of accuracy obtained using this procedure may not be sufficient for structure prediction—a conclusion that is consistent with that reached by Thomas and Dill (1996).

The rather different results for native structures obtained by related contact potentials suggests that it is unlikely that purely pairwise potentials are sufficient for structure prediction. The inclusion of higher order correlations could be important in reducing the sensitivity of ground state topologies to variations in the potentials. It is known from polymer physics that a description of the stable globular shape of isolated polymers require the addition of three body terms (de Gennes, 1985). The changes in the topology of the lattice structures given in Figure 5 can be viewed as the instability of the native state to alterations in (not unrelated) potentials. By analogy, with the stability of globules we conjecture that the inclusion of three body terms might minimize the sensitivity of the topology of the native state to various interaction schemes. For hydrophobic residues a generalized Axilrod–Teller potential (Axilrod & Teller, 1943), which is based on dipole interactions, may be adequate. It is not yet clear whether one requires accurate three body interactions for reliably predicting the folds of proteins.

Acknowledgments

M.R. Betancourt is a National Science Foundation Minority Postdoctoral Fellow. One of us (D.T.) is grateful to R. Mountain, J.D. Weeks, and R. Zwanzig for useful discussions. We would like to thanks R.L. Jernigan for sending us copies of manuscripts prior to publication. This work was supported in part by a grant from the National Science Foundation (through grant number CHE-96-29845).

References

Axilrod RM, Teller E. 1943. Interaction of the van der Waals type between three atoms. *J Chem Phys 11*:299–300.

Bryngelson J. 1994. When is a potential accurate enough for structure prediction: Theory and application to a random heteropolymer model of protein folding. *J Chem Phys 100*:6038–6045.

Bussemaker HJ, Thirumalai D, Bhattacharjee JK. 1997. Thermodynamic stability of folded proteins against mutations. *Phys Rev Lett 79*:3530–3533.

Camacho J, Thirumalai D. 1993. Kinetics and thermodynamics of folding in model proteins. *Proc Natl Acad Sci USA 90*:6369–6372.

Chan HS, Dill KA. 1990. On the origins of structure in globular proteins. *Proc Natl Acad Sci USA 87*:6388–6392.

de Gennes PG. 1985. *Scaling concepts in polymer physics*. Ithaca, NY: Cornell University Press.

Du R, Grosberg AY, Tanaka T. 1998. Models of protein interactions: How to choose one. *Fold Des 3*:203–211.

Edwards SF. 1965. Statistical mechanics of polymers with excluded volume. *Proc Phys Soc (London) 85*:613–624.

Fauchere JL, Pliska V. 1983. Hydrophobic parameters π of amino acid side chains from the partitioning of N-acetyl-amino acid amides. *Eur J Med Chem 18*:639–375.

Ferrenberg AM, Swendsen RH. 1989. Optimized Monte Carlo data analysis. *Phys Rev Lett 63*:1195–1198.

Finkelstein AV, Badretdinov AY, Gutin AM. 1995. Why do protein architecture have Boltzmann-like statistics? *Proteins Struct Funct Genet 23*:142–150.

Godzik A, Kolinski A, Skolnick J. 1995. Are proteins ideal mixtures of amino acids? Analysis of energy parameters sets. *Protein Sci 4*:2107–2117.

Godzik A, Kolinski A, Skolnick J. 1996. Knowledge-based potentials for protein folding: What can we learn from protein structures? *Structure 4*:363–366.

Jernigan RL, Bahar J. 1996. Structure-derived potentials and protein simulations. *Curr Opin Struct Biol 6*:195–209.

Levitt M. 1976. A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol 104*:59–107.

Li H, Tang C, Wingreen NS. 1997. Nature of driving force for protein folding: A result from analyzing the statistical potential. *Phys Rev Lett 79*:765–768.

Mirny LA, Shakhnovich EI. 1996. How to derive a protein folding potential? A new approach to an old problem. *J Mol Biol 264*:1164–1179.

Miyazawa S, Jernigan RL. 1985. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules 18*:534–552.

Miyazawa S, Jernigan RL. 1996. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol 256*:623–644.

Moult J. 1997. Comparison of database potentials and molecular mechanics for fields. *Curr Opin Struct Biol 7*:194–199.

Nozaki Y, Tanford C. 1971. A simplified representation of protein conformations from rapid simulations of protein folding. *J Biol Chem 246*:2211–2217.

Pande VS, Grosberg AY, Tanaka T. 1995. How accurate must potentials be for successful modeling of protein folding. *J Chem Phys 103*:9482–9491.

Roseman MA. 1988. Hydrophilicity of polar amino acid side-chains is markedly reduced by flanking peptide bonds. *J Mol Biol 200*:513–522.

Sippl MJ. 1995. Knowledge-based potentials for proteins. *Curr Opin Struct Biol 5*:229–235.

Skolnick J, Jaroszewski L, Kolinski A, Godzik A. 1997. Derivation and testing of pair potentials for protein folding: When is the quasichemical approximation correct? *Protein Sci 6*:676–688.

Sun S. 1993. Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms. *Protein Sci 2*:762–785.

Tanaka S, Scheraga HA. 1976. Medium and long range interaction parameters between amino acids for predicting three dimensional structures of proteins. *Macromolecules 9*:945–950.

Thomas PD, Dill KA. 1996. Statistical potentials extracted from protein structures: How accurate are they? *J Mol Biol 257*:457–469.

Zhang L, Skolnick J. 1998. How do potentials derived from structural databases relate to "true" potentials? *Protein Sci 7*:112–122.