WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Dissecting contact potentials for proteins: Relative contributions of individual amino acids

N.-V. Buchete,[1]* J. E. Straub,[2] and D. Thirumalai[3]

[1] Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892

[2] Department of Chemistry, Boston University, Boston, Massachusetts 02215

[3] Biophysics Program, Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742

ABSTRACT

Knowledge-based contact potentials are routinely used in fold recognition, binding of peptides to proteins, structure prediction, and coarse-grained models to probe protein folding kinetics. The dominant physical forces embodied in the contact potentials are revealed by eigenvalue analysis of the matrices, whose elements describe the strengths of interaction between amino acid side chains. We propose a general method to rank quantitatively the importance of various inter-residue interactions represented in the currently popular pair contact potentials. Eigenvalue analysis and correlation diagrams are used to rank the inter-residue pair interactions with respect to the magnitude of their relative contributions to the contact potentials. The amino acid ranking is shown to be consistent with a mean field approximation that is used to reconstruct the original contact potentials from the most relevant amino acids for several contact potentials. By providing a general, relative ranking score for amino acids, this method permits a detailed, quantitative comparison of various contact interaction schemes. For most contact potentials, between 7 and 9 amino acids of varying chemical character are needed to accurately reconstruct the full matrix. By correlating the identified important amino acid residues in contact potentials and analysis of about 7800 structural domains in the CATH database we predict that it is important to model accurately interactions between small hydrophobic residues. In addition, only potentials that take interactions involving the protein backbone into account can predict dense packing in protein structures.

Proteins 2008; 70:119–130.

Key words: *amino acid ranking; protein folding; contact interactions; amino acid substitution; minimal alphabet for proteins; protein binding; protein design; eigenvalue analysis.*

INTRODUCTION

The number of resolved protein structures and sequences deposited in protein data banks increases every year by thousands.[1] Nevertheless, the majority of protein structures for which sequences are known, remain unresolved. In recent years, atomistic approaches to simulating and predicting protein structures have evolved rapidly, taking advantage of advances in both algorithmic and computational hardware capabilities. However, it is still not feasible to apply atomistic methods to large scale protein structure prediction or to studies of protein–protein interactions or binding of small molecules and peptides to proteins. The difficulty in simulating in detail the folding or binding of even modest sized proteins and peptides has led to the development of minimalistic coarse-grained models.[2] The need to model, at least qualitatively, interactions between proteins, or ligand driven allosteric transitions in biological nanomachines, has lead to the development of a number of novel coarse-grained models. Although the level of detail in these models varies, the energy functions in many of these are often derived from databases of known structures.[3–5] Because of the increasing popularity of coarse-grained models in the context of structural biology,[2,6,7] it is useful to assess the extent to which they include chemical diversity of amino acids. The purpose of this article is to dissect the relative contributions of individual amino acids to commonly used pair potentials derived to identify fold recognition.

Pairwise contact potentials are the most simple and widely used representations of inter-residue interactions. Since their introduction,[3,8] contact potentials have been successfully used in many applications ranging from protein structure prediction to protein design and docking.

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Grant sponsor: National Science Foundation; Grant numbers: NSF-CHE-05-14056, NSF-CHE-03-16551; Grant sponsor: NIDDK, NIH (Intramural Research Program).

*Correspondence to: Nicolae-Viorel Buchete, Laboratory of Chemical Physics, NIDDK, National Institutes of Health, 9000 Rockville Pike, Bldg. 5, Rm. 137A, Bethesda, Maryland, 20892-0520.
E-mail: buchete@nih.gov

Received 31 January 2007; Accepted 20 March 2007
Published online 19 July 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21538

The contact potentials describe the interactions between the 20 side chains by a 20 × 20 matrix, the elements of which give the interaction strength between a pair of amino acids at contact. Two amino acid residues are in contact if the distance between them is less than a cutoff distance, R_c. Typically, the contact potentials are derived from known protein structures, and hence R_c is chosen to reflect the value in the X-ray or NMR structures.

A strong interest in analyzing contact potentials comes from the need to understand the effects of amino acid sequence complexity on the nature of the protein structural fold and their stability.[9–13] Efforts have been made to classify amino acids[14–17] with the goal of identifying the minimal number of amino acid types that is needed for protein design and protein folding.[18–21] Rapid methods to assess binding of ligands and peptides to proteins require knowledge of the overall contributions that different amino acids make to the various potentials. For example, it has been shown[22] that binding of antigenic peptides to major histocompatibility complex (MHC), which is a prerequisite for recognition by cytotoxic T-cells, is better predicted by the BT[23] potential that treats hydrophilic interactions more adequately than the MJ-96 potential,[5] which places emphasis on hydrophobic interactions.

Previous studies[18,23–25] of the 20 × 20 contact potential matrices suggest that eigenvalue analysis are useful for investigating their specific features, and for characterizing the underlying physical driving forces involved in protein folding. In Figure 1 we illustrate, using a gray scale representation, six contact potential matrices that are further analyzed in this article. They were developed by Miyazawa and Jernigan (MJ-96,[5] and MJ-99[26]), Betancourt and Thirumalai (BT[23]), Skolnick *et al.* (SJKG,[4] and Sko-1a and Sko-1b from Tables 1a and 1b in Ref. 27), Hinds and Levitt (HL[28]), Tobi *et al.*[25] (TSLE-5a and TSLE-5b from Tables 5a and 5b in Ref. 25), and Buchete *et al.* (BST[29]). The BST matrices were derived from orientational and distance-dependent interactions. To reduce them to contact form, the full potentials were integrated over distance and angles: BST-fu (forward-up, $\theta \in [0, \pi/2]$, $\phi \in [0, \pi]$), BST-bd (backwards-down, $\theta \in [\pi/2, \pi]$, $\phi \in [\pi, 2\pi]$), and BST (all angles, i.e., $\theta \in [0, \pi]$, $\phi \in [0, 2\pi]$). All contact matrices were rearranged such that the amino acid order is the same as in the Miyazawa-Jernigan[5] matrix (MJ-96). We also subtracted, the corresponding mean values from all the analyzed matrices to prevent an extremely big largest eigenvalue.[24] In the gray scale representation (Fig. 1), lighter shades correspond to more attractive interactions while darker shades correspond to stronger repulsions.

Li *et al.*[24] showed that the popular Miyazawa-Jernigan[5] potential matrix has only two dominant eigenvalues (Fig. 2, MJ-96), and that their corresponding eigenvectors are strongly correlated to each other and to a hydrophobicity scale.[30] The presence of the two dominant eigenvalues implies that only two types of residues (hydrophobic (H) and polar (P)) are needed to describe the major forces that determine the nature of protein folds. More recently, Wang and Lee[18] deepened the analysis of the MJ-96 potentials, by showing that the origin of the strong HP character of the interactions is due to important correlations between the elements of the leading eigenvector (q_i) and the dipolar moments (Q_i) of the side chains.[31] These observations support the widely held notion that the most relevant characteristic of a given residue's interactions is how a residue interacts with water.[32] The relationship between hydrophobicity and the principal eigenvector of contact potential matrices was recently used to study the structure, stability and evolution of proteins.[33–37] Pokarowski *et al.*[38] have analyzed a large set of contact potentials and have shown that they can be largely classified in two classes, both having strong correlations with hydrophobic transfer energies. However, only one class is significantly correlated to amino acid isoelectric points.

During the last decade, details related to chain connectivity, compactness of the native state, and the effects of secondary structure have been incorporated in contact potentials.[4,27] One example is the newer Miyazawa-Jernigan (MJ-99) potentials, parameterized using an improved self-consistent procedure that leds to enhanced ability to discriminate native structures from non-native folds.[26] Such improvements, which account for a variety of characteristics beyond the HP classification, result in a more complex potential with a weaker eigenvalue separation than in the MJ-96 case (Figs. 1–3).

In this article we introduce a general amino acid ranking method based on an eigenvalue analysis for pairwise contact potential matrices. Eigenvalue analysis is a general tool that may be employed to study any contact potentials, and permits the ranking of the relative contributions of each interacting amino acid. Our ranking method allows us to reconstruct the contact potentials using the most important residues. Such a "mean field" reconstruction is indicative of the importance of amino acids of different chemical character in the contact potentials. The objective ranking of the amino acid interactions makes possible the direct, quantitative comparison of various contact potentials and it may be applied to protein structure and design, to protein–protein interactions, and to the interpretation of amino acid mutation studies.

METHODS (THEORY)

The pairwise contact potential matrices are symmetric and self-adjoint. Thus all the eigenvalues are real, and the corresponding eigenvectors can be constructed as a complete orthonormal set.[39] The eigenvalue equation for matrix **M** is

DOI 10.1002/prot

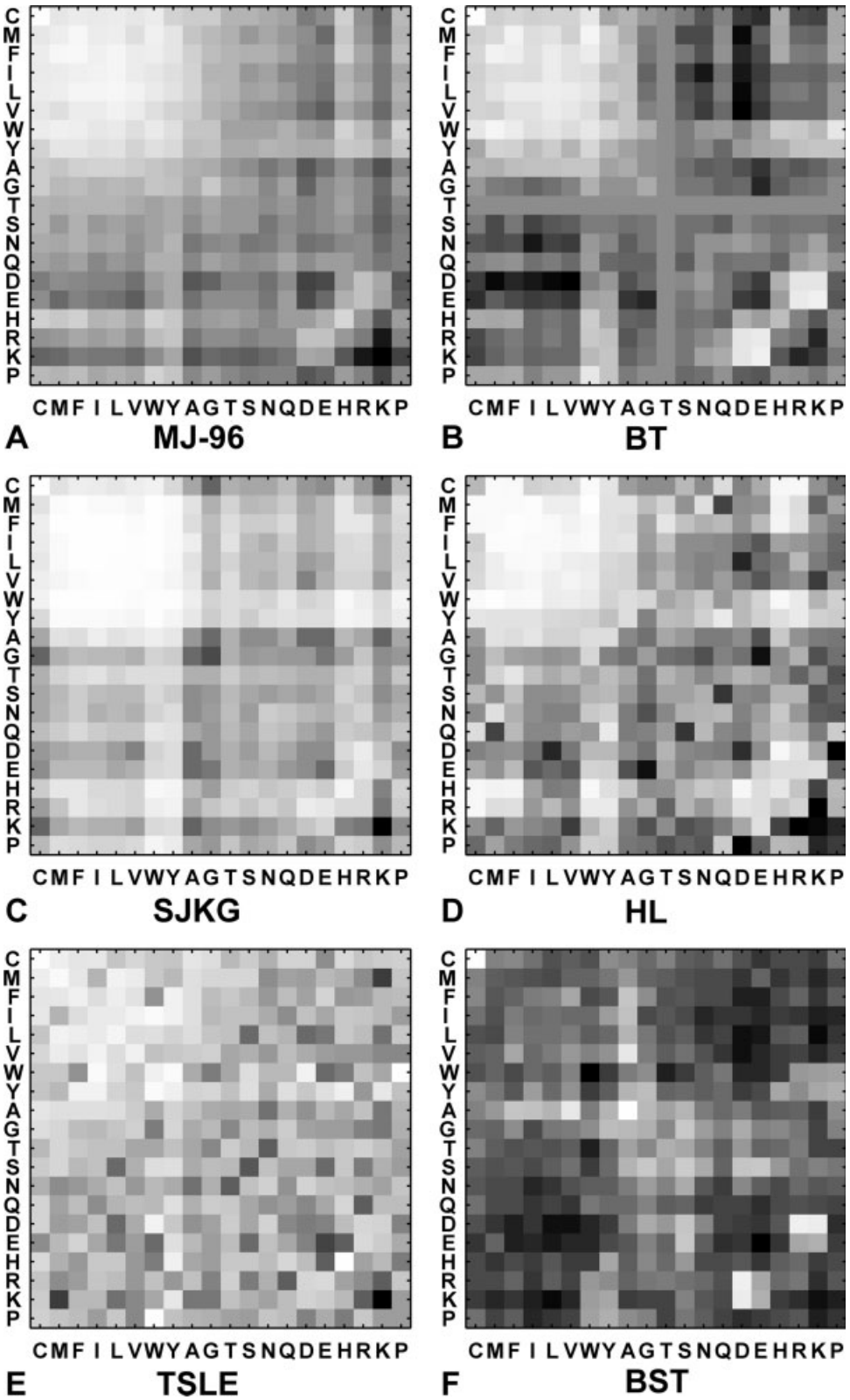

Figure 1

*Gray scale representation of some of the contact potential matrices. They are (**a**) Miyazawa and Jernigan[5] (MJ-96), (**b**) Betancourt and Thirumalai[23] (BT), (**c**) Skolnick* et al.*[4] (SJKG), (**d**) Hinds and Levitt[28] (HL), (**e**) Tobi* et al.*[25] (TSLE, from Table 5a in Ref. 25), and Buchete* et al.*[29] (BST).*

$$\mathbf{M}|\mathbf{v}^i\rangle = \lambda_i|\mathbf{v}^i\rangle \quad (1)$$

where λ_i are the real eigenvalues and $\langle \mathbf{v}^i \mid \mathbf{v}^j \rangle = \delta_{ij}$ is the orthonormality relation of the $i \in \{1, 2, \ldots, 20\}$ eigenvectors.

Figure 2(a,b) show the leading λ_i values calculated for contact potentials such as the ones depicted in Figure 1. If the complete set of real eigenvalues and eigenvectors are known, the original matrix can be reconstructed exactly using

$$\tilde{\mathbf{M}} = \sum_{n=1}^{N=20} |\mathbf{v}^n\rangle \lambda_n \langle \mathbf{v}^n| \quad (2)$$

where $\langle \mathbf{v}^n|$ is the transpose of the eigenvector $|\mathbf{v}^n\rangle$, and $\mathbf{v}_j^n$ is the j-th element. In cases where there are only a few ($N_{\min} < 20$) dominant eigenvalues (e.g., as for MJ-96), the following approximate reconstruction formula can be employed with good accuracy

$$\tilde{\mathbf{M}}_{ij} = \sum_{n=1}^{N_{\min}} \lambda_n \mathbf{v}_i^n \mathbf{v}_j^n \quad (3)$$

This eigenvalue-based reconstruction procedure is illustrated in Figure 3 for the newer MJ-99 matrix and the corresponding reconstructed matrices ($\tilde{\mathbf{M}}$) using only the first [Fig. 3(b)], the first two [Fig. 3(c)] and the first three [Fig. 3(d)] largest eigenvalues.

To facilitate the comparison of the contact potentials on equal footing, all the matrices $\mathbf{M}$ were first scaled to the [0, 1] range, and the mean value was subtracted.[24] All contact matrices were also rearranged such that the amino acid order is the same as for the Miyazawa-Jernigan[5] matrix. On the basis of the analysis, we conclude that for most contact potential matrices the separation of the leading eigenvalues is not as strong as for MJ-96. Figure 2 shows the relative magnitude of the eigenvalues for the contact potential matrices depicted in Figure 1.

A quantitative measure of the accuracy of reconstruction is the linear correlation coefficient r which is defined for any two matrices $\mathbf{M}$ and $\tilde{\mathbf{M}}$ as

$$r = \frac{\langle \mathbf{M} \times \tilde{\mathbf{M}} \rangle - \langle \mathbf{M} \rangle \langle \tilde{\mathbf{M}} \rangle}{\sqrt{[\langle \mathbf{M} \times \mathbf{M} \rangle - \langle \mathbf{M} \rangle^2][\langle \tilde{\mathbf{M}} \times \tilde{\mathbf{M}} \rangle - \langle \tilde{\mathbf{M}} \rangle^2]}}, \quad (4)$$

where the average $\langle \mathbf{M} \times \tilde{\mathbf{M}} \rangle$ is calculated for the products between the corresponding individual elements $\mathbf{M}_{ij}$ and $\tilde{\mathbf{M}}_{ij}$ and not over the matrix product. Figure 4 shows the correlation coefficients of the elements of the original $\mathbf{M}$ matrices and their reconstructed values ($\tilde{\mathbf{M}}$). Using this analysis we can answer the question: How many eigenvalues are necessary and sufficient to obtain a reconstructed matrix that has a correlation coefficient with the original matrix of r_c or better? Here, r_c is a critical threshold

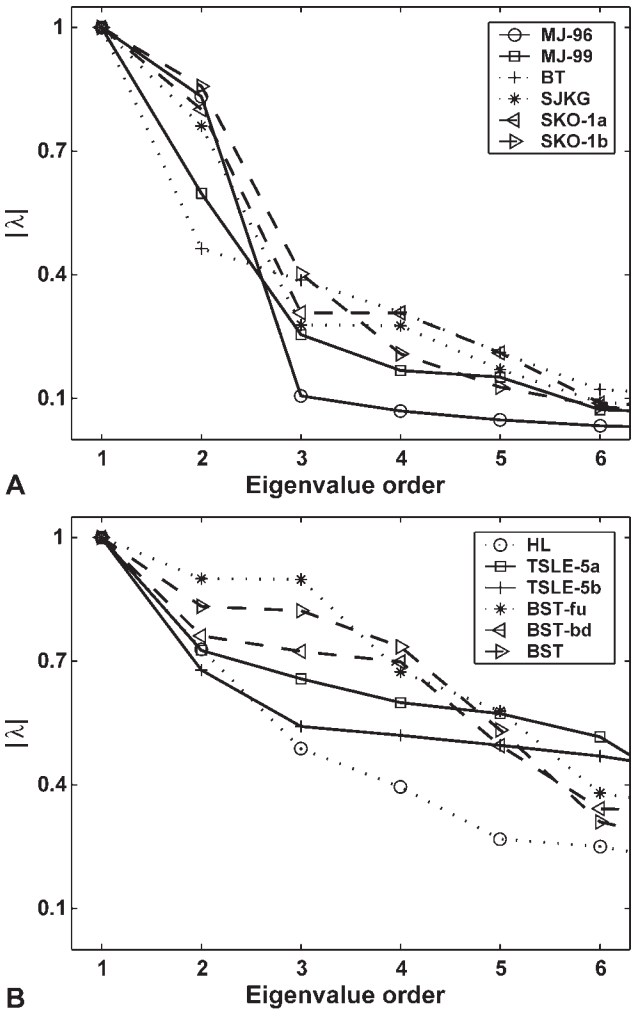

Figure 2

The largest eigenvalues of several statistical contact potential matrices. The eigenvalues are ranked according to their absolute magnitude. These contact potentials were developed by Miyazawa and Jernigan (MJ-96,[5] and MJ-99[26]), Betancourt and Thirumalai (BT[23]), Skolnick et al. (SJKG,[4] and Sko-1a and Sko-1b from Tables 1a and 1b in Ref. 27), Hinds and Levitt (HL[28]), Tobi et al.[25] (TSLE-5a and TSLE-5b from Tables 5a and 5b in Ref. 25), and Buchete et al. (BST,[29] see Fig. 1 and the text for details).

value of the correlation coefficient. For example, if $r_c = 0.9$ (i.e. a very strong correlation) we see from Figure 4 that only two largest eigenvalues are sufficient in the case of the MJ-96 matrix, while three eigenvalues are necessary for the BT interaction matrix. For most contact potentials (Fig. 4) only a few eigenvalues are required to reconstruct the original matrix.

RESULTS AND DISCUSSION

The relative contribution of each amino acid

The eigenvalue analysis of the MJ-96 matrix revealed[18,24] strong correlations between the elements of the eigenvector

DOI 10.1002/prot

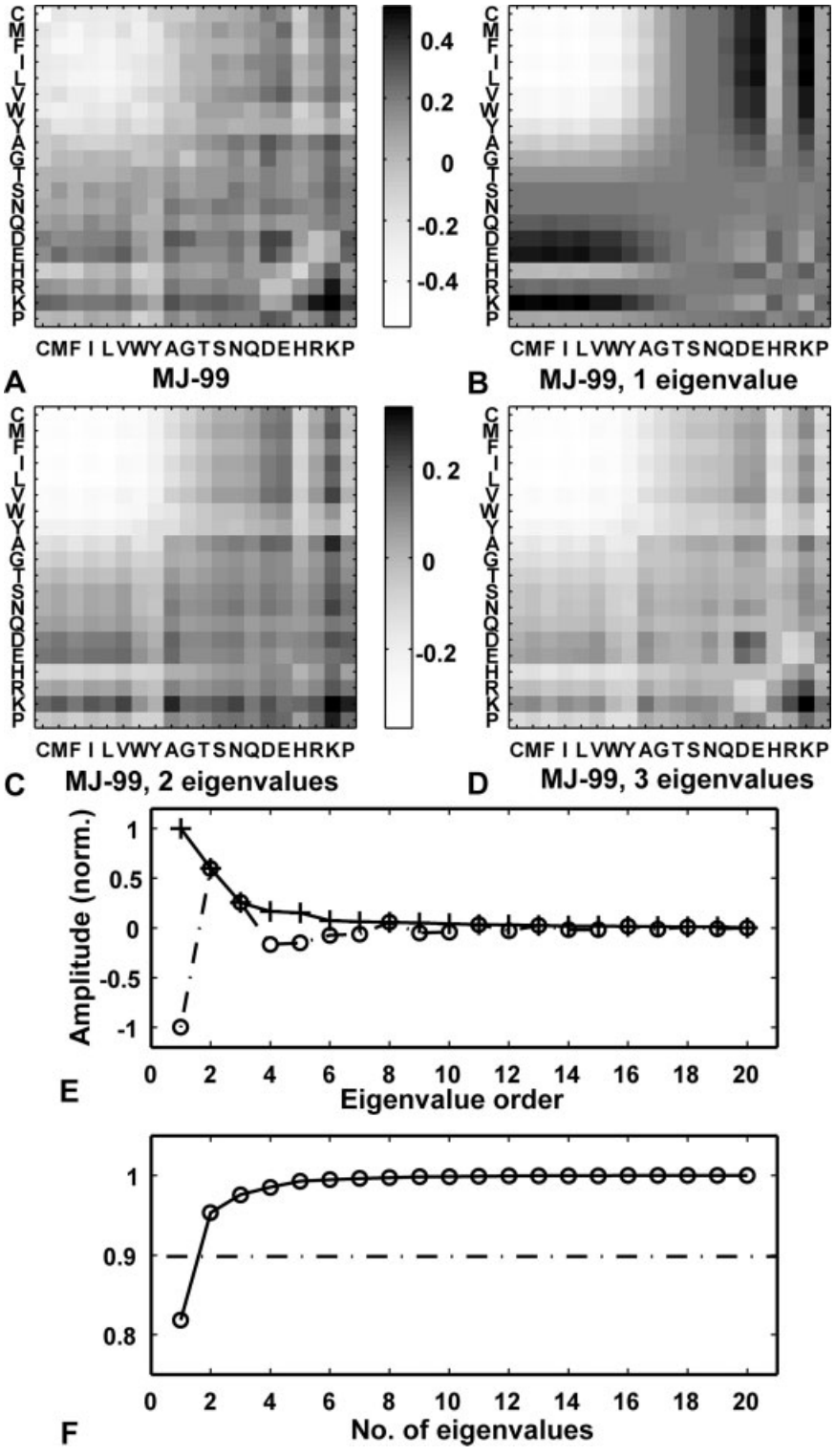

Figure 3

*The MJ-99 matrix (**a**) and the corresponding matrices reconstructed by using (**b**) one largest eigenvalue, (**c**) 2 largest eigenvalues, and (**d**) 3 largest eigenvalues and their corresponding eigenvectors. (**e**) The eigen value spectrum (circles, normalized to the largest eigenvalue) and its absolute values (plus sign). (**f**) The correlation coefficient between the original MJ-99 matrix and its approximate reconstructions using only a few largest eigenvalues up to the full (i.e., 32 values) spectrum.*

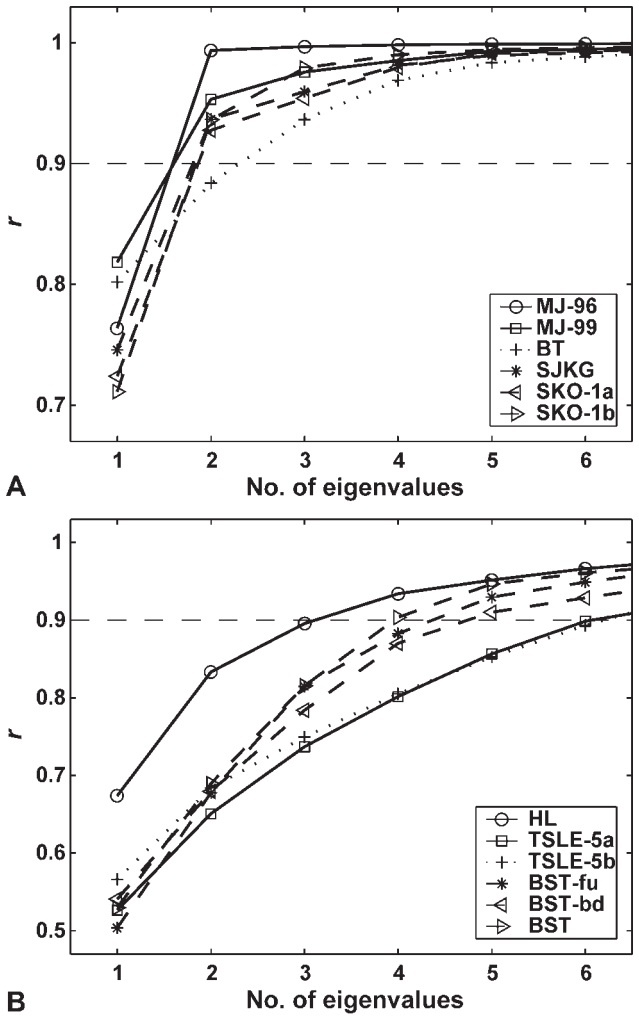

Figure 4

Correlation coefficients (r) calculated between several potential matrices (see text) and their approximate reconstructions using only a few largest eigenvalues.

corresponding to the largest eigenvalue of the MJ-96 matrix, and physical properties of the individual amino acids such as the hydrophobicities [see Eq. (4) and Fig. 2 in Ref. 24] and the electric dipole moment [see Eq. (3) and Fig. 1 in Ref. 18]. These observations suggest that the amplitudes of the elements of the eigenvectors corresponding to dominant eigenvalues are directly proportional to the magnitude of the physical interactions between the corresponding amino acids. Based on this observation, we define an *importance* vector **I** with components

$$\mathbf{I}_i = \sum_{j=1}^{N_{\min}} |\lambda_j \mathbf{v}_i^j| \qquad (5)$$

The elements of **I** are proportional to the relative magnitudes of the interactions that each residue makes to $\mathbf{I}_i$. To facilitate the comparison of **I** vectors obtained for different contact potentials, it is better to map the elements of each vector **I** to the [0, 1] range by using the scaling relation $\mathbf{I}_i \rightarrow (\mathbf{I}_i - \min(\mathbf{I}))/(\max(\mathbf{I}) - \min(\mathbf{I}))$.

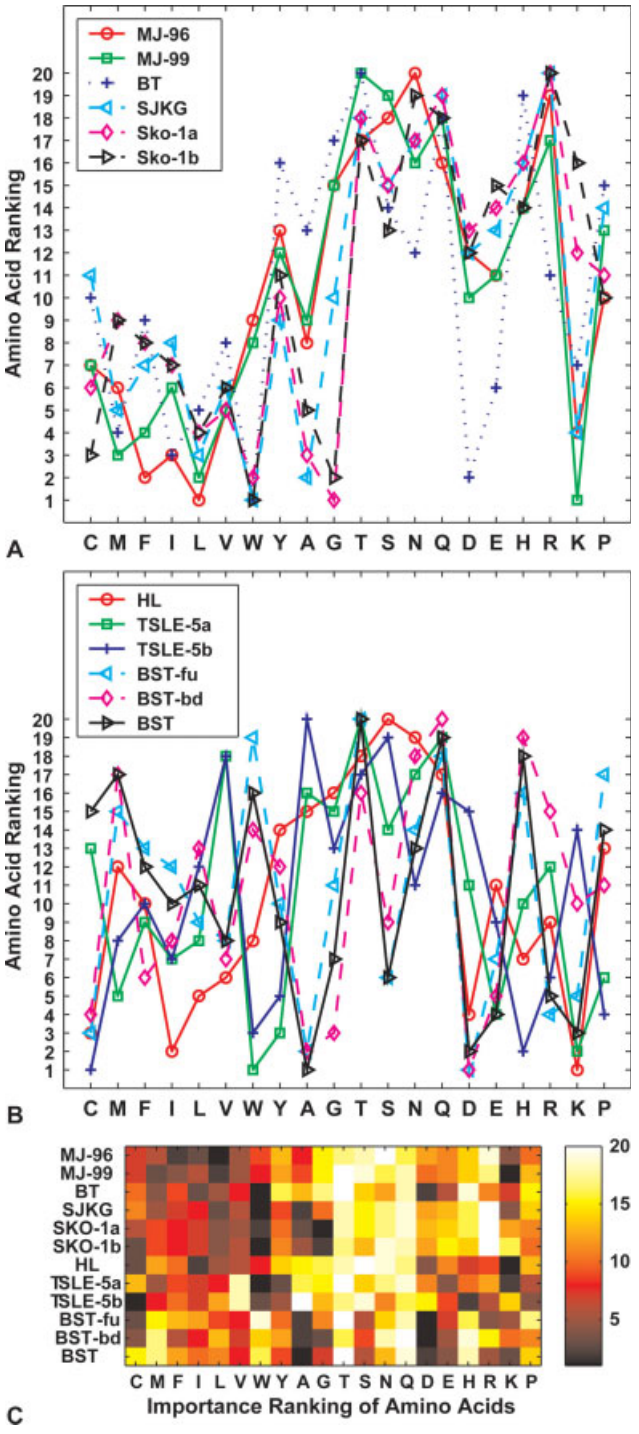

Figure 5

*(**a**, **b**) The importance ranking of specific amino acids (i.e., the $\mathbf{I}_{rank}$ vectors, 1 being the most important) for several contact potential matrices. (**c**) Representation of the $\mathbf{I}_{rank}$ values calculated for several, commonly used contact potentials. Important amino acids are dark red and black, as shown in the color scale.*

DOI 10.1002/prot

We show in Figure 5 the ranking values obtained for the vectors **I** for the various contact potentials. Amino acids such as *Thr*, *Asn*, and *Gln* have low **I** values for most contact potentials, while interactions involving hydrophobic or charged amino acids have higher values (Fig. 5). Since in some cases different amino acids have similar $\mathbf{I}_i$ values, it is useful to analyze the ranking of the various amino acids (i.e., 1st, 2nd, etc.) corresponding to each **I** vector.

Although the amino acid ranking is relatively similar for the contact matrices analyzed in Figure 5(a) (MJ-96, MJ-99, BT, SJKG, Sko-1a and Sko-1b), it is different for the other potentials [Fig. 5(b,c)]. As a confirmation of the validity of the amino acid ranking method proposed here, we note that *Thr* is ranked as the "least important" amino acid for the BT potential, which justifies its choice as the optimal reference state.[23]

Mean field reconstruction of contact interactions

Another argument in favor of the amino acid ranking proposed above (Eq. 5) comes from analyzing the correlation coefficients between the full, original potential matrices, and the matrices reconstructed using the mean field approximation. If only "important" amino acid interactions are maintained from the original matrix and all other elements are replaced by the corresponding mean values for each potential, one would expect that matrices reconstructed using "less important" amino acids should be consistently less correlated with the original matrix. In Figure 6(a) are shown several mean field reconstructed matrices for the MJ-96 potential, to illustrate this method.

The results of the correlation calculations between the original MJ-96 matrix and its corresponding mean field reconstructed matrices, using different combinations of more or less important amino acids, are presented in Figure 6(b). The data points on the bottom correspond to r values computed when only one single amino acid (corresponding to the nearby letter) is used. The second set of points from the bottom, corresponds to cases when two amino acids are used, and so on. For example, the data point labeled "LFI" corresponds to a mean field matrix $\tilde{\mathbf{M}}$ that was reconstructed by using only *Leu*, *Phe*, and *Ile*. The continuous straight lines represent linear fits for each series of data points. All fits have negative slopes, indicating that the amino acid ranking defined above is consistent with the mean field representation. We have calculated this type of correlation diagrams for all the potentials mentioned above (Figs. S1 and S2 in Supplementary Material), and the results are shown in Figure 7. For all contact potentials studied, the matrices reconstructed using less important amino acids are consistently less correlated with the full, original matrices, than matrices corresponding to important amino acids. These results, summarized in Figure 7 and Table S-I (Supplementary Material), answer the question: How many and which specific amino acids are necessary and sufficient for building a mean field reconstructed potential that has a correlation with the original potential of r_c or better? (here $r_c = 0.9$).

The reduced sets of amino acids extracted for the contact interaction potentials listed in Table S-I are shown in Table S-II, together with their side chain size, charged, or hydrophobic properties, respectively.[40] The same reduced sets of amino acids are shown in Table S-III, with emphasis on the character of their packing in the interior of proteins. We note that both MJ-96 and MJ-99 potentials are strongly dominated by interactions between predominantly small hydrophobic residues, together with strong contributions from *Lys* and *Cys*. The acidic and polar residues appear to have an average role in the MJ-96 and MJ-99 interaction schemes, as well as the amino acids with large side chains. The interactions with large side chains such as Trp, His and Tyr are more relevant for the HL, SJKG, BT, and TSLE contact potentials than for the MJ and BST interactions. The most important MJ-96 and MJ-99 (Table S-III) residues are typically found in the interior of protein structures, with the exception of *Lys* that is predominantly exposed to the solvent, and *Cys* that has a strong affinity for forming *Cys–Cys* contacts. Comparatively, the other contact potentials have a less hydrophobic character, with amino acid classes represented almost uniformly in their interaction schemes. An interesting general observation is that the polar, uncharged *Thr*, *Asn*, and *Gln* amino acids are assigned the weakest interactions by all contact potentials investigated here.

Randomly generated contact potentials

As one more test of the proposed method for ranking the 20 amino acids based on their contribution to contact interactions, we estimate the probability of obtaining a similar ranking by generating random contact potential matrices. By extracting parameters for the best fitting Gaussian distributions of the elements of the potentials analyzed in this paper (Fig. 8), we can generate new random contact potentials.

We analyzed data obtained for random matrices that correspond to Gaussian distributions similar to the original contact potential matrices (Figs. S3–S6). Ten thousand such matrices were generated for each contact potential analyzed in this work, and their amino acid rankings were compared to the original reference matrices. The results show clearly that the probability of obtaining amino acid rankings similar to the original, reference interaction matrix is extremely small (e.g., as shown in Fig. S4) for all the types of contact potentials. The probability to obtain an amino acid ranking from a randomly generated matrix that has a correlation coefficient of 0.6 or better with the ranking obtained for the original matrix, is in the [0.004, 0.006] range. However,

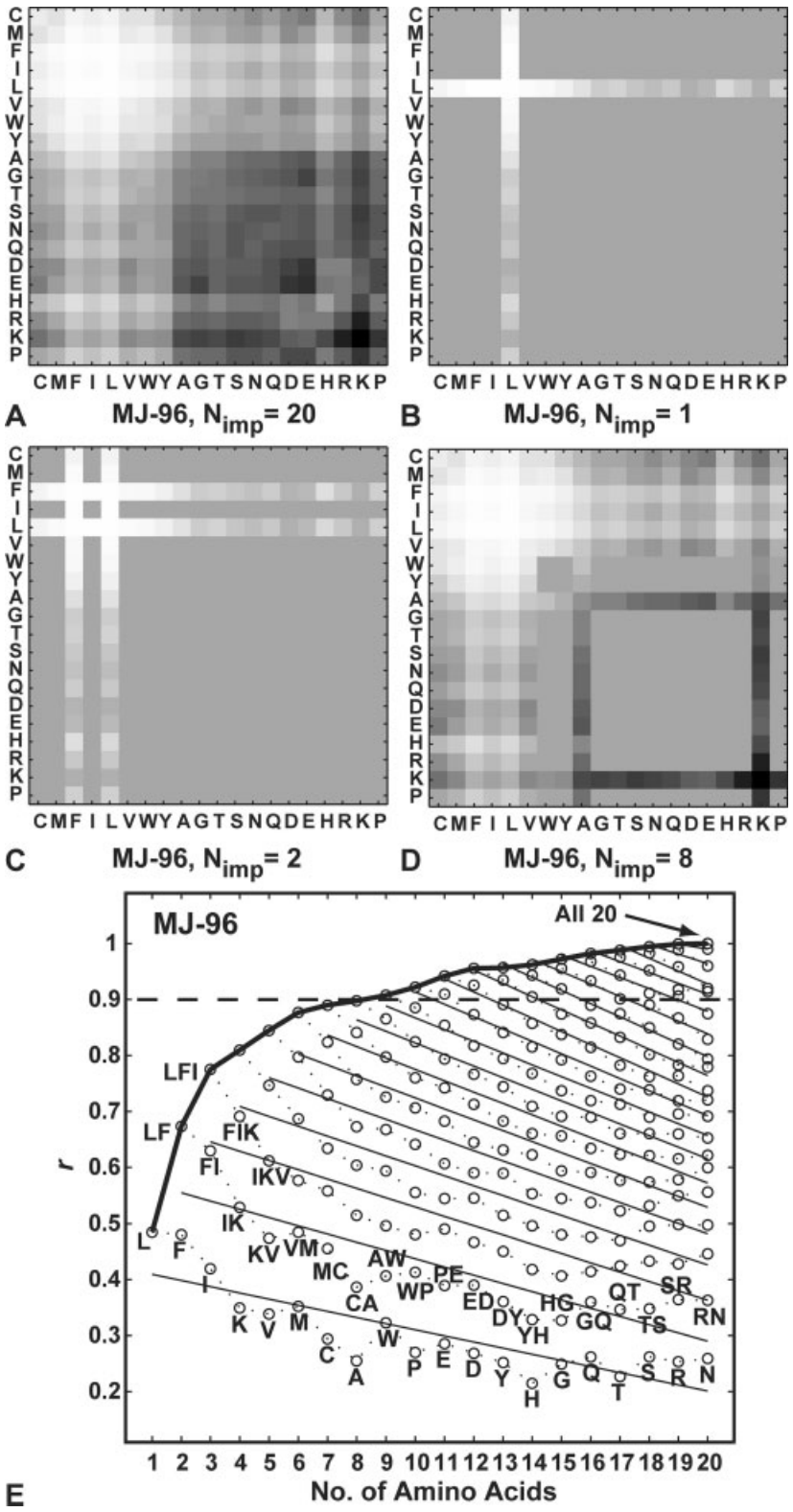

Figure 6

Illustration of the mean-field reconstruction procedure of the Miyazawa-Jernigan (MJ-96) potential.[5] *The original values (**a**) are reconstructed using only one (**b**), two (**c**) or eight (**d**) most important amino acids, while all others are replaced by the mean value. (**e**) The importance ranking is tested by computing the correlation coefficient between the original MJ-96 potential matrix and the matrices reconstructed using the "mean-field" procedure. When less important amino acids are used, the correlation is consistently smaller. Note that at least eight amino acids are needed for* $r_c = 0.9$.

DOI 10.1002/prot

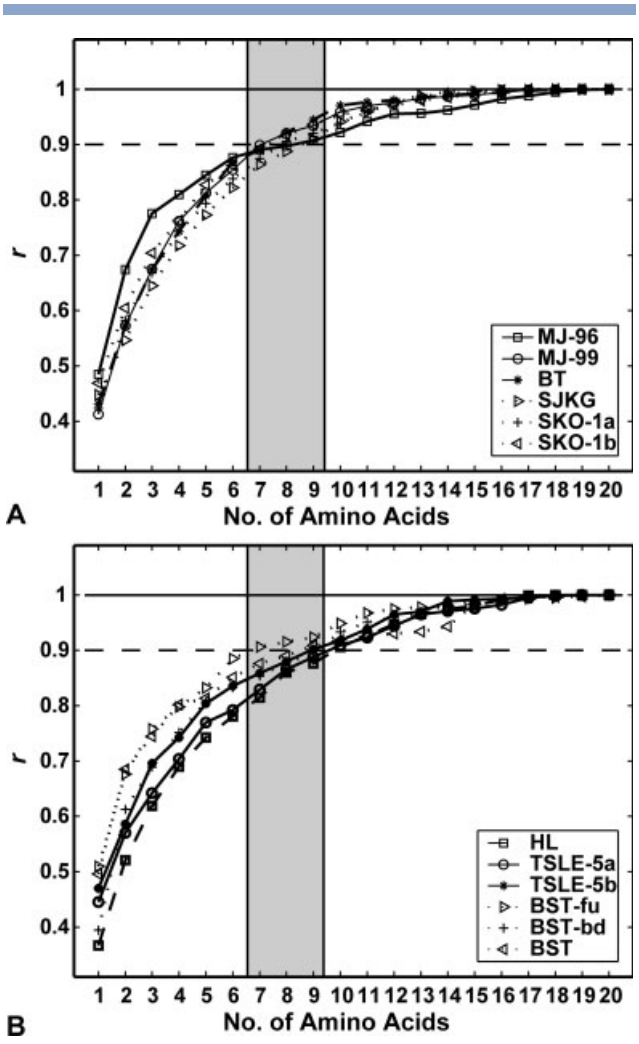

Figure 7

*(**a**, **b**) Correlation curves constructed for mean field reconstructed values for several contact potentials. At least $N_{AA} = 7$ amino acids are necessary for any contact potential matrices to be reconstructed with a correlation $r > 0.9$ to the original matrix. For all matrices, only 7–9 important amino acids (gray zone) are sufficient for reconstructing the full contact potential with $r > 0.9$.*

this probability drops dramatically to the [0.0004, 0.0006] range if a correlation coefficient of 0.7 or better is sought for the amino acid ranking. Our amino acid ranking method seems therefore to be robust against randomly generated data.

We conclude that the most commonly used contact matrices reflect the nature of the forces that stabilize protein folds. Thus, the quasi-chemical approximation, inherent in these potentials, is a reasonable approximation for describing interactions in proteins.

Contacts potentials and classes of protein structures

Most of the available contact potentials suggest that about 7–9 amino acid residues are required to capture the chemical diversity of proteins (Fig. 7). It is likely the case that the most effective contact potential will depend on the application, as was shown in the context of ligand binding to MHC complexes.[22] We can get further insight into the appropriateness of the contact potentials by considering packing in proteins, which is important in the context of structure prediction. Since our analysis permits the ranking of all SC–SC interactions for any type of contact potential, we can use it to predict the appropriateness of using a certain interaction scheme to modeling proteins with different secondary structures.

To relate the contact potentials to protein secondary structures, we calculate the preponderance of interactions that are present in a variety of protein structures. For this purpose, we use the CATH (version 3.0.0, May 2006) database[41] of representative protein classes (i.e., class (1) mainly-α, (2) mainly-β, (3) α + β, and (4) a class that contains miscellaneous protein domains with low secondary structure content) to assess the fraction of side chain contacts that are typically present in proteins. We use nine classes for grouping the 20 residues types[40] as: "sH" for the small-hydrophobic (A,V,I,L,M), "LH" for large-hydrophobic (Y,W,F), "sP" for small-polar (S,T), "LP" for large-polar (N,Q,H), "pos" for the positive (R,K), "neg" for negative (D,E) and single letter codes for "G", "P", and "C." The values in Table I are mean values obtained for each structural class by dividing the sets of representative domains in the CATH database into 9 subsets. The corresponding standard error for each value is given in brackets.

The results in Table I show that most contacts occur between sH residues, with about 4.5% higher frequencies for Class 1 (mainly-α) protein domains than for Class 2 (mainly-β) (i.e., 23.9 vs. 19.4%, Table IA). When considering the fraction of side chain-backbone (SC-BB) contacts (i.e., by using an extra interaction site "BB" located on the backbone, as in our previous work[42]) these results show a very high SC-BB fraction of contacts (Table IB and IC) in all cases. However, mainly-α structures have a 20% higher fraction of SC–SC contacts (Table IC) as compared to mainly-β structures, at the expense of BB–BB interactions. Together with data in Table IB, it appears that in mainly-β structures, many sH-sH and sH-LH contacts (which are more common in mainly-α structures), are being replaced by BB–BB backbone contacts.

On the basis of the above observations and on the interaction ranking resulting from our method (e.g., see Tables S-II and S-III), we predict that the MJ, BT, and SJKG contact potentials will perform better than other potentials in modeling secondary structure of typical proteins because they have a good balance between contacts of sH, LH, and charged residues. However, the MJ types of contact potentials may be more adequate to model proteins that are classified as CATH Class I, mainly-α (and, accordingly, BT and SJKG may perform

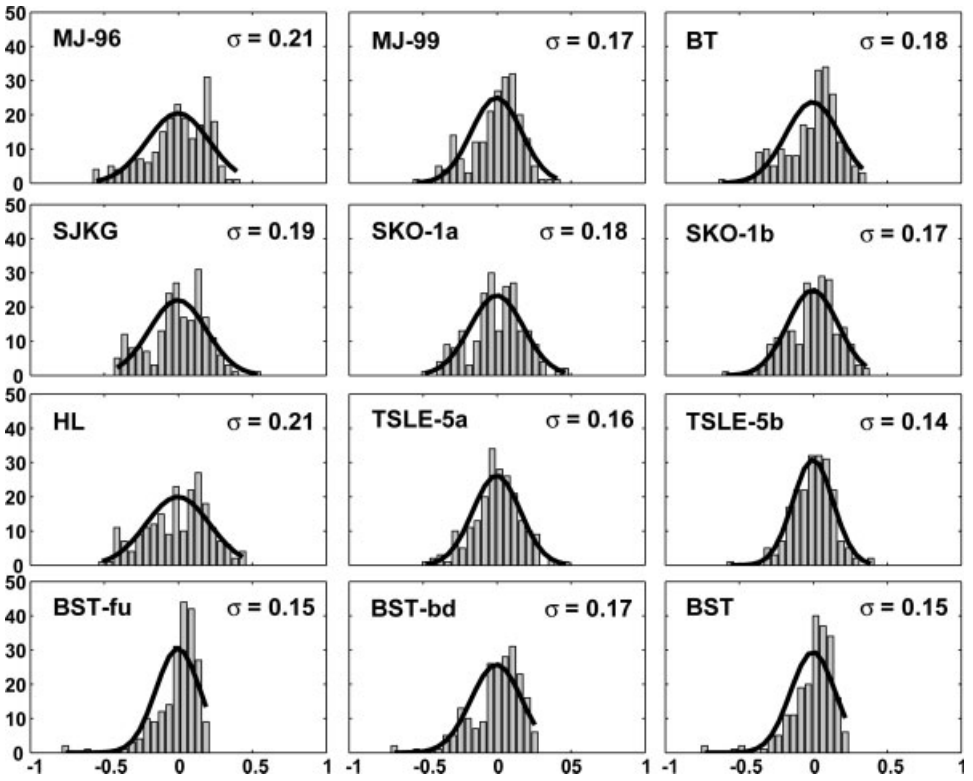

Figure 8

Extracting Gaussian distribution parameters for contact potentials. Since the matrices are symmetric, only 210 interaction values are used for building the histograms. For an unbiased comparison, all interactions are first scaled to the [0, 1] interval and the mean values are subtracted. Note that some contact matrices appear to be less normally distributed than others.

better for modeling Class 2 and Class 3 structures) because MJ contacts appear to give higher weights to interactions between small-hydrophobic amino acids. As suggested for scoring functions used in protein docking,[43] the direct correlation between contact potentials suggests that a variety of interaction schemes may be needed to predict the structure of proteins. The present analysis clearly shows the need to develop potentials that also include the shapes and size of amino acid residues.

CONCLUSIONS

We have presented a general method for the analysis of pairwise contact potential matrices, which permits the ranking of each inter-residue interaction component according to its contribution to the global features of contact potentials. The method is used to analyze several widely used contact potential interaction matrices for proteins. We show that the new ranking method (see Eq. 5) is consistent with the mean field reconstruction technique, and with the selection of reference states used in previous studies (e.g., *Thr* for the BT potential[23]). This new method offers a theoretical basis for protein design using a minimum number of amino acids. In particular, our results support the findings that stable and unique designs can be achieved using only a subset of suitably chosen amino acids.[44–47] The present analysis identifies the precise minimum subset of residues that globally correlate with each contact potential.

Quantitatively, our analysis shows that only 7–9 residues are sufficient for a very good approximation of the most widely used 20 × 20 inter-residue contact interaction schemes (i.e., such that the reconstructed interaction matrix has a correlation coefficient of at least 0.9 with the full 20 × 20 matrix). The amino acid importance ranking, resulting from the fast growing variety of contact interaction potentials, was applied to study the relationship between the different types of contact potentials and their efficacy to model specific classes of protein secondary structures as defined by the CATH database. The correlation between contact potentials and the analysis of the CATH database shows that the preponderance of interactions between small hydrophobic residues must be considered for accurately predicting protein structures. Moreover, interactions involving backbone atoms must also be modeled for

 DOI 10.1002/prot

Table I
Fractions of Side Chain Contacts in Protein Structural Classes (i.e., as Defined for the CATH Database, v.3.0.0, May 2006)

	Class 1	Class 2	Class 3	Class 4
	α[1877]	β[1839]	α + β[3956]	misc. [162]
A. Percentage of side chain contacts using the 9 amino acid (AA) groups (see text for notation)				
sH-sH	23.92 (0.24)	19.4 (0.13)	24.55 (0.25)	11.29 (1.17)
sH-LH	10.87 (0.09)	9.35 (0.09)	9.46 (0.07)	6.68 (0.49)
sH-sP	6.60 (0.07)	7.00 (0.07)	6.94 (0.09)	4.91 (0.60)
sH-LP	5.93 (0.09)	4.72 (0.07)	4.93 (0.08)	5.27 (0.45)
sH-pos	5.33 (0.09)	–	–	5.67 (0.43)
sH-G	–	4.54 (0.09)	4.60 (0.05)	–
B. Percentage of contacts using 10 AA groups (9 + a ``BB´´ side on backbone)				
sH-BB	15.06 (0.13)	14.06 (0.06)	15.47 (0.10)	10.55 (0.64)
sH-sH	12.72 (0.17)	6.23 (0.05)	10.08 (0.12)	–
BB-BB	10.13 (0.17)	28.07 (0.09)	19.79 (0.07)	19.61 (0.66)
sH-LH	5.78 (0.06)	–	3.91 (0.05)	–
BB-LH	4.10 (0.06)	3.83 (0.04)	–	4.48 (0.38)
BB-sP	–	6.05 (0.07)	4.80 (0.05)	–
BB-C	–	–	–	6.49 (0.88)
BB-LP	–	–	–	4.78 (0.52)
C. Percentage of side chain - backbone (``BB´´) contacts				
SC-SC	53.14 (0.26)	32.10 (0.07)	41.05 (0.07)	37.44 (1.09)
SC-BB	36.73 (0.12)	39.83 (0.07)	39.17 (0.08)	42.95 (0.68)
BB-BB	10.13 (0.18)	28.07 (0.09)	19.78 (0.07)	19.61 (0.66)

The number of representative protein structural domains used are given in square brackets. The standard errors estimated for each type of fraction of contacts are shown in brackets. Only the largest five fractions of contact types are shown for each class.

describing the folded structures of proteins, especially those involving β-sheets. Our ranking method can be used as a guide in the development and evaluation of new potentials for the study of protein folding, for protein structure prediction and design, or for the development of novel residue substitution matrices for protein sequence analysis.[48–50]

ACKNOWLEDGMENTS

NVB is thankful to Dr. Gerhard Hummer for helpful discussions and support during the preparation of this manuscript. This work was supported in part by the National Science Foundation through the grants CHE-05-14056 (DT) and CHE-03-16551 (JES).

REFERENCES

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
2. Buchete NV, Straub JE, Thirumalai D. Development of novel statistical potentials for protein fold recognition. Curr Opin Struct Biol 2004;14:225–232.
3. Tanaka S, Scheraga HA. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. Macromolecules 1976;9:945–950.
4. 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.
5. 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.
6. Best RB, Chen Y-G, Hummer G. Slow protein conformational dynamics from multiple experimental structures: the helix/sheet transition of Arc repressor. Structure 2005;13:1755–1763.
7. Bahar I, Rader A. Coarse-grained normal mode analysis in structural biology. Curr Opin Struct Biol 2005;15:586–592.
8. Levitt M, Warshel A. Computer simulation of protein folding. Nature 1975;253:694–698.
9. Wolynes PG. As simple as can be? Nat Struct Biol 1997;4:871–874.
10. Doi N, Kakukawa K, Oishi Y, Yanagawa H. High solubility of random-sequence proteins consisting of five kinds of primitive amino acids. Protein Eng Des Sel 2005;18:279–284.
11. Murphy LR, Wallqvist A, Levy RM. Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Eng 2000;13:149–152.
12. Li T, Fan K, Wang J, Wang W. Reduction of protein sequence complexity by residue grouping. Protein Eng 2003;16:323–330.
13. Khatun J, Khare SD, Dokholyan NV. Can contact potentials reliably predict stability of proteins? J Mol Biol 2004;336:1223–1238.
14. Esteve JG, Falceto F. Classification of amino acids induced by their associated matrices. Biophys Chem 2005;115(2–3, Special Issue):177–180.
15. Du R, Grosberg AY, Tanaka T. Models of protein interactions: how to choose one. Fold Des 1998;3:203–211.
16. Loose C, Klepeis JL, Floudas CA. A new pairwise folding potential based on improved decoy generation and side-chain packing. Proteins 2004;54:303–314.
17. Kosiol C, Goldman N, Buttimore NH. A new criterion and method for amino acid classification. J Theor Biol 2004;228:97–106.
18. Wang Z-H, Lee HC. Origin of the native driving force for protein folding. Phys Rev Lett 2000;84:574–577.
19. Wang J, Wang W. Grouping of residues based on their contact interactions. Phys Rev E 2002;65:419111–419115.
20. Williams G, Doherty P. Inter-residue distances derived from fold contact propensities correlate with evolutionary substitution costs. BMC Bioinformatics 2004;5:153.
21. Fan K, Wang W. What is the minimum number of letters required to fold a protein? J Mol Biol 2003;328:921–926.
22. Schueler-Furman O, Altuvia Y, Sette A, Margalit H. Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. Protein Sci 2000;9:1838–1846.
23. Betancourt MR, Thirumalai D. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein Sci 1999;8:361–369.
24. Li H, Tang C, Wingreen NS. Nature of driving force for protein folding: a result from analyzing the statistical potential. Phys Rev Lett 1997;79:765–768.
25. Tobi D, Shafran G, Linial N, Elber R. On the design and analysis of protein folding potentials. Proteins 2000;40:71–85.
26. Miyazawa S, Jernigan RL. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. Proteins 1999;34:49–68.
27. Skolnick J, Kolinski A, Ortiz A. Derivation of protein-specific pair potentials based on weak sequence fragment similarity. Proteins 2000;38:3–16.
28. Hinds DA, Levitt M. A lattice model for protein structure prediction at low resolution. Proc Natl Acad Sci USA 1992;89:2536–2540.
29. Buchete NV, Straub JE, Thirumalai D. Anisotropic coarse-grained statistical potentials improve the ability to identify nativelike protein structures. J Chem Phys 2003;118:7658–7671.

30. Levitt M. A simplified representation of protein conformations for rapid stimulation of protein folding. J Mol Biol 1976;104:59–107.
31. Chipot C, Maigret B, Rivail JL, Scheraga HA. Modeling amino-acid side-chains. I. Determination of net atomic charges from ab initio self-consistent-field molecular electrostatic properties. J Phys Chem 1992;96:10276–10284.
32. Chan HS, Dill KA. Origins of structure in globular proteins. Proc Natl Acad Sci USA 1990;87:6388–6392.
33. Bastolla U, Porto M, Roman HE, Vendruscolo M. Looking at structure, stability, and evolution of proteins through the principal eigenvector of contact matrices and hydrophobicity profiles. Gene 2005;347(2, Special Issue):219–230.
34. Bastolla U, Porto M, Roman HE, Vendruscolo M. Prinicipal eigenvector of contact matrices and hydrophobicity profiles in proteins. Proteins 2005;58:22–30.
35. Esteve JG, Falceto F. A general clustering approach with application to the Miyazawa-Jernigan potentials for amino acids. Proteins 2004;55:999–1004.
36. Rivas E. Evolutionary models for insertions and deletions in a probabilistic modeling framework. BMC bioinformatics 2005;6:63.
37. Wiederstein M, Sippl MJ. Protein sequence randomization: efficient estimation of protein stability using knowledge-based potentials. J Mol Biol 2005;345:1199–1212.
38. Pokarowski P, Kloczkowski A, Jernigan RL, Kothari NS, Pokarowska M, Kolinski A. Inferring ideal amino acid interaction forms from statistical protein contact potentials. Proteins 2005;59:49–57.
39. Arfken GB, Weber H-J. Mathematical methods for physicists. Boston: Elsevier; 2005. p 1182.
40. Dima RI, Thirumalai D. Asymmetry in the shapes of folded and denatured states of proteins. J Phys Chem B 2004;108:6564–6570.
41. Pearl F, Todd A, Sillitoe I, Dibley M, Redfern O, Lewis T, Bennett C, Marsden R, Grant A, Lee D, Akpor A, Maibaum M, Harrison A, Dallman T, Reeves G, Diboun I, Addou S, Lise S, Johnston C, Sillero A, Thornton J, Orengo C. The CATH domain structure database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Res 2005;33 (Database issue):D247–D251.
42. Buchete NV, Straub JE, Thirumalai D. Orientational potentials extracted from protein structures improve native fold recognition. Protein Sci 2004;13:862–874.
43. Murphy J, Gatchell DW, Prasad C, Vajda S. Combination of scoring functions improves discrimination in protein–protein docking. Proteins 2003;53:840–854.
44. Riddle DS, Santiago JV, Bray-Hall ST, Doshi N, Grantcharova VP, Yi Q, Baker D. Functional rapidly folding proteins from simplified amino acid sequences. Nat Struct Biol 1997;4:805–809.
45. Chan HS. Folding alphabets. Nat Struct Biol 1999;6:994–996.
46. Wang J, Wang W. A computational approach to simplifying the protein folding alphabet. Nat Struct Biol 1999;6:1033–1038.
47. Cieplak M, Holter NS, Maritan A, Banavar JR. Amino acid classes and the protein folding problem. J Chem Phys 2001;114:1420–1423.
48. Miyazawa S, Jernigan RL. A new substitution matrix for protein sequence searches based on contact frequencies in protein structures. Protein Eng 1993;6:267–278.
49. Tan YH, Huang H, Kihara D. Statistical potential-based amino acid similarity matrices for aligning distantly related protein sequences. Proteins 2006;64:587–600.
50. Prlic A, Domingues FS, Sippl MJ. Structure-derived substitution matrices for alignment of distantly related sequences. Protein Eng 2000;13:545–550.

 DOI 10.1002/prot