Nature of Driving Force for Protein Folding: A Result From Analyzing the Statistical Potential

Hao Li, Chao Tang, and Ned S. Wingreen
NEC Research Institute, 4 Independence Way, Princeton, New Jersey 08540

(Received 5 December 1996)

In a statistical approach to protein structure analysis, Miyazawa and Jernigan derived a 20 × 20 matrix of inter-residue contact energies between different types of amino acids. Using the method of eigenvalue decomposition, we find that the Miyazawa-Jernigan matrix can be accurately reconstructed from its first two principal component vectors as $M_{ij} = C_0 + C_1(q_i + q_j) + C_2 q_i q_j$, with constant C's, and 20 q values associated with the 20 amino acids. This regularity is due to hydrophobic interactions and a force of demixing, the latter obeying Hildebrand's solubility theory of simple liquids. [S0031-9007(97)03600-4]

PACS numbers: 87.15.By, 02.10.Sp, 64.75.+g

Proteins fold into specific three dimensional structures to perform their diverse biological functions. It is now well established that for small proteins the information contained in the amino acid sequence is sufficient to determine the folded structure, which is the structure with minimum free energy [1]. Thus the native structure is dictated by the physical interactions between amino acids in the sequence, and understanding the nature of such interactions is crucial for protein structure prediction.

As a protein contains thousands of atoms and interacts with huge number of water molecules, it is not feasible to calculate the free energy function from first principles. An often adapted practical approach is to derive a coarse grained potential (often on the level of amino acids) using the known structures in the existing protein data banks. In such an approach, the energy of a particular substructure in proteins is derived from the number of its appearances in the structure data bank via a Boltzmann factor [2–4]. A classic example of such a statistical potential is the Miyazawa-Jernigan (MJ) matrix, a 20 × 20 inter-residue contact-energy matrix derived by Miyazawa and Jernigan [2,5]. This matrix tabulates the interaction strength between any two types of amino acids in proteins, and has been widely applied in protein design and folding simulations [6,7].

In this Letter, we apply a general method of matrix analysis, namely, eigenvalue decomposition, to the MJ matrix. The analysis reveals an intrinsic regularity of the MJ matrix, which yields basic information about the nature of the driving force for protein folding. We show that despite the complicated interactions in proteins, the major driving force is hydrophobic interaction and a force of demixing, the latter obeying Hildebrand's solubility theory of simple liquids [8]. The result allows us to attribute the interactions responsible for folding to quantifiable properties of individual amino acids. These properties suggest further experimental tests, and can be used for analyzing sequence-structure relation.

Eigenvalue decomposition is a general approach to analyzing matrices. A given $N \times N$ real symmetric matrix M can be reconstructed by the following formula:

$$M_{ij} = \sum_{\alpha=1}^{N} \lambda_\alpha V_{\alpha,i} V_{\alpha,j}\,, \tag{1}$$

where M_{ij} is the element of the matrix in row i and column j, λ_α is the αth eigenvalue, and $V_{\alpha,i}$ is the ith component of the corresponding eigenvector. We have analyzed the MJ matrix using eigenvalue decomposition. First, we subtract the mean $\langle M_{ij} \rangle$ from each element and then analyze the eigenvalue spectrum of the remaining matrix [9]. We find that the eigenvalue spectrum has two dominant eigenvalues which are much larger in magnitude than the rest. Specifically, we find $\lambda_1 = -22.49$, $\lambda_2 = 18.62$, while the rest of the eigenvalues have absolute values between 2.17 and 0.013. This suggests (as we shall demonstrate below) that the matrix can be accurately reconstructed using only the first two eigenvectors, $\tilde{M}_{ij} = \langle M_{ij} \rangle + \lambda_1 V_{1,i} V_{1,j} + \lambda_2 V_{2,i} V_{2,j}$. Further analysis shows that the second eigenvector is related to the first one by a shift and rescaling, i.e., $V_{2,i} = \beta + \gamma V_{1,i}$, with $\beta = -0.30$, $\gamma = -0.90$, and a correlation coefficient 0.986. Using this relation, the expression for $\tilde{M}_{ij}$ can be written simply as

$$\tilde{M}_{ij} = C_0 + C_1(q_i + q_j) + C_2 q_i q_j\,, \tag{2}$$

where $q_i \equiv V_{1,i}$, and the C's are constants, $C_0 = -1.492$, $C_1 = 5.030$, and $C_2 = -7.400$. Thus we can reconstruct the MJ matrix (which in principle could have 210 independent elements) by using only twenty parameters q_i, associated with the twenty amino acids, and three interaction coefficients. Such a simple interaction form is often the starting point for a theoretical modeling of proteins [10].

The spectrum of the MJ matrix (two large eigenvalues with corresponding eigenvectors related to each other) reflects the specific physical interaction between the amino acids. The connection between the interaction and the spectrum can be understood in the following general way: Consider a pairwise interaction matrix M_{ij} which is determined by certain properties of two species i and j, denoted by q_i and q_j. Assume, on physical grounds, that M_{ij} can be expressed as an analytical function $f(q_i, q_j)$ with a well defined converging power series, $f(q_i, q_j) = C_0 + C_1(q_i + q_j) + C_2 q_i q_j + C_3(q_i^2 + q_j^2) + C_4(q_i q_j^2 + q_j q_i^2) + \cdots$, where the C's are constants. Take first the example where the expansion ends at the C_2 term, i.e., $M_{ij} = C_0 + C_1(q_i + q_j) + C_2 q_i q_j$. Since any row of the matrix M is given by a vector $\mathbf{U}_i = (C_0 + C_1 q_i)\mathbf{I} + (C_1 + C_2 q_i)\mathbf{Q}$, which is a linear combination of $\mathbf{I}$ and $\mathbf{Q}$, where $\mathbf{I} \equiv \{1, 1, \ldots, 1\}$, and $\mathbf{Q} \equiv \{q_1, q_2, \ldots, q_n\}$, one can decompose the vector space $\mathcal{G}$ into the subspace $\mathcal{G}_\parallel$ spanned by $\mathbf{I}$ and $\mathbf{Q}$, and its perpendicular compliment $\mathcal{G}_\perp$. It is obvious that $\mathcal{G}_\perp$ gives rise to $n - 2$ zero eigenvalues, as $M\mathbf{V}_\perp = 0$ for any vector $\mathbf{V}_\perp$ in the subspace $\mathcal{G}_\perp$. Furthermore, the two eigenvectors with nonzero eigenvalues must be expressible as a linear combination of $\mathbf{I}$ and $\mathbf{Q}$, therefore they are related to each other by a shift and rescaling. Similarly, if the expansion ends at the C_4 term, there will be three nonzero eigenvalues, and the corresponding eigenvectors will lie in a subspace spanned by $\mathbf{I}$, $\mathbf{Q}$, and $\mathbf{Q}^2$, where $\mathbf{Q}^2 \equiv \{q_1^2, q_2^2, \ldots, q_n^2\}$. The same argument applies to all higher order expansions. This analysis applies to the ideal case where there is no noise in the matrix. Introducing noise leads to a slight mixing of $\mathcal{G}_\perp$ and $\mathcal{G}_\parallel$ and therefore to small nonzero values for the rest of the eigenvalue spectrum.

The reconstructed matrix in Eq. (2) reproduces the original MJ matrix to a high accuracy. Figure 1 shows the correlation between the original MJ matrix and the reconstructed one. The regression line is $y = 0.999x + 0.008$, and the correlation coefficient is 0.989. On average Eq. (2) gives matrix elements with only 5% error compared to the original matrix.

Notice that one can redefine the q's in Eq. (2) by a shift and rescaling while leaving the interaction form unchanged. Therefore any transformation $q \rightarrow Aq + B$ with a corresponding change in the C's yields an identical matrix. To better understand the physical meaning of Eq. (2), we rewrite it in the following form:

$$\tilde{M}_{ij} = h_i + h_j - C_2(q_i - q_j)^2/2, \tag{3}$$

where

$$h_i = C_0/2 + C_1 q_i + (C_2/2)q_i^2. \tag{4}$$

Now each term in Eq. (3) above is invariant under the transformation discussed above.

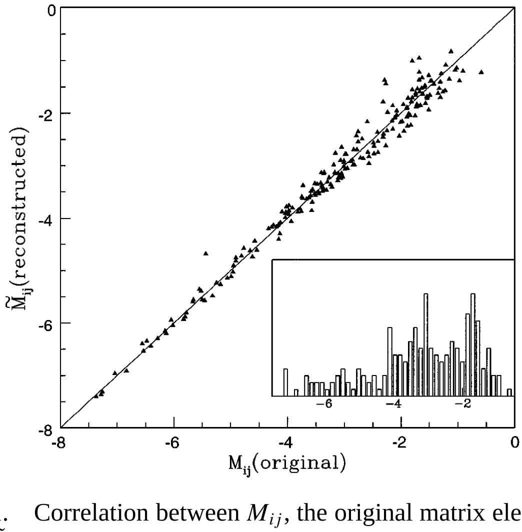

FIG. 1. Correlation between M_{ij}, the original matrix elements, and $\tilde{M}_{ij}$, the matrix elements reconstructed from Eq. (2). The regression line is $y = 0.999x - 0.008$. The correlation coefficient is 0.989. Inset: The distribution of the MJ matrix elements. The unit of energy is $k_B T$.

What is the physical basis for the simple interaction form in Eq. (3)? Consider the quantity $\chi_{ij} \equiv 2\tilde{M}_{ij} - \tilde{M}_{ii} - \tilde{M}_{jj}$. Since $\tilde{M}_{ij}$ is the energy of forming a contact between type i and type j amino acids in water, χ_{ij} gives the energy of breaking one i-i contact and one j-j contact and forming two pairs of i-j contacts; thus χ_{ij} is the energy change due to the mixing of the two types of amino acids. According to Eq. (3), $\chi_{ij} = -C_2(q_i - q_j)^2$. This form has a striking similarity to the mixing energy of two simple liquids as given by Hildebrand's solubility theory (HST) [8]. In his 1933 classic paper, Hildebrand derived the energy of mixing of two simple liquids by summing over the pairwise interactions throughout the mixture. Assuming that the mixing is random and that the potentials between molecules are of the Lennard-Jones type due to the London dispersion force [11], Hildebrand arrived at a formula which expresses the energy of mixing of liquids A and B as $E_{\text{mixing}} \propto (\delta_A - \delta_B)^2$, where $\delta_{A,B}$ are pure component properties related to the square root of the vaporization energies of liquids A and B, traditionally called the "solubility parameter."

Now we can imagine the formation of 2 i-j contacts in water by two steps, formation of an i-i contact and a j-j contact followed by a mixing [12]. The energy change for the first step is $2h_i + 2h_j$, and that for the second step χ_{ij}. As the formation of an i-i contact in water is related to the segregation of amino acids of type i in water, we expect that h_i is related to the hydrophobicity of amino acid i. Indeed, we find that h_i correlates very well with the hydrophobicity scales published in the literature [13] (see Fig. 2). Thus despite the complicated interactions in proteins, we find that the pairwise inter-residue interactions responsible for folding can be attributed to the hydrophobic force and a force of demixing, the latter obeying HST. (Although HST was derived for simple nonpolar molecules, it was found previously that the theory describes well the behavior of polymer blends [14]. The application to proteins is another example of the more general scope of HST.)

The above analysis presents a simple picture of the nature of interactions between amino acids. It also provides experimentally testable predictions. Comparison with HST indicates that the q_i we derive should be linearly related to the solubility parameter of amino acid i, which can be measured. Furthermore, we predict

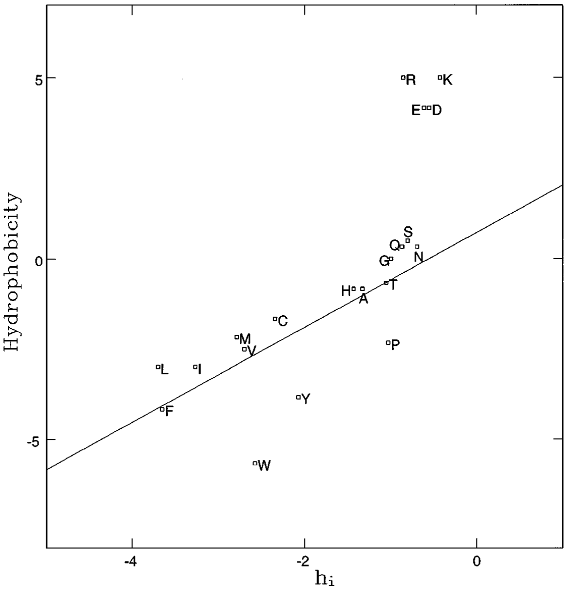

FIG. 2. Calculated h_i and measured hydrophobicities [13] of the 20 amino acids. The type of amino acid is indicated using the standard one letter code. The straight line is a linear fit (excluding the charged amino acids) with slope 1.314 and intercept 0.759. The correlation coefficient is 0.769.

from Eq. (4) that hydrophobicity can be expressed as a quadratic function of the solubility parameter. Since the solubility parameter and the hydrophobicity of an amino can be measured independently, this prediction can also be tested.

Comparison of the terms in Eq. (3) shows that the linear term $h_i + h_j$ is the dominant one in selecting the native structure. This is because the typical difference of the linear term δh (among different types of contacts) is much larger than the typical difference of the square term $\delta\chi/2$, specifically, $\delta h = 6.52(\delta\chi/2)$. Therefore the energy difference between different compact structures (due to different arrangements of the contacts) is mainly due to the linear term. Thus, through a quantitative analysis of the MJ matrix we arrive at the conclusion that the hydrophobic force is the dominant driving force for protein folding [15].

The term $-C_2(q_i - q_j)^2/2$ has an important consequence, however. This term favors demixing of amino acids (C_2 is negative). The microscopic basis for such a demixing force is the dissimilar polarizability of the two monomers [11]. Since the interior of a protein is composed of various types of amino acids which tend to segregate, an amino acid buried in the interior of a protein will experience an environment which is quite different from a uniform nonpolar environment. It has been controversial whether one can model the interior of a protein as a uniform nonpolar environment [16]. This study suggests that in general it is not adequate to do so.

It is worth noting that although Eq. (2) in general gives a very good fit for contact energies (as shown in Fig. 1), there are a few exceptions where the error is large. For example, Eq. (2) underestimates the attraction between positively and negatively charged amino acids (GLU-ARG, GLU-LYS, ASP-ARG, ASP-LYS). In other words, the χ_{ij} term (which measures the energy of mixing) does a poor job for pairs with opposite charges, which favor mixing due to Coulomb interaction. Another example is CYS-CYS contact, where the attraction is stronger than that given by Eq. (2) due to the formation of a disulfide bridge. Some exceptions are, of course, expected as we aimed at revealing the dominant features of the MJ matrix using only the two dominant eigenvalues and the corresponding eigenvectors. It is certain that some attributes of amino acids (such as Coulomb charge, sulfhydryl groups, etc.) will not be captured by a simple q value.

Notice that in Fig. 2 the charged amino acids fall into a separate group. Since Eq. (2) gives accurate values for all the pairs involving charged amino acids (except the four pairs formed between positively and negatively charged amino acids), we believe that h_i for a charged amino acid does measure the free energy change of hiding the side chain from water. The different behavior of charge amino acids in Fig. 2 is likely due to the fact that the theoretically constructed h_i and the experimentally measured hydrophobicity represent different quantities for charged amino acids. Experimentally hydrophobicity was obtained by measuring the relative solubilities of the amino acids in water and organic solvent, which involve different ionization states of the amino acids. On the other hand, h_i gives the free energy cost of bringing together two already ionized amino acids. Thus, it is not a surprise to find that these two quantities are not very well correlated.

The q values we obtain can be used to characterize amino acids. The distribution of the q values is bimodal (see Fig. 3), which supports the notion that amino acids naturally fall into two distinct groups: "polar" (P) and "hydrophobic" (H). This division also accounts for the three different regions in the distribution of the MJ matrix elements (see the inset of Fig. 1), which reflect

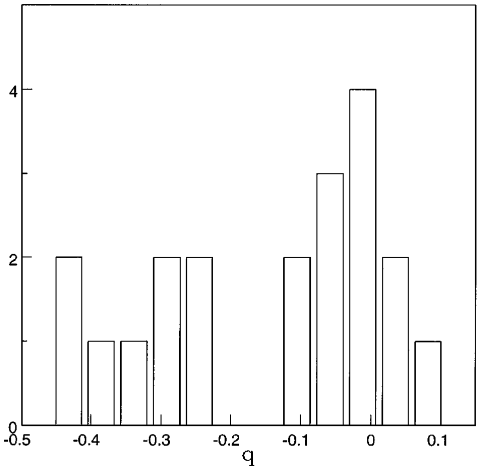

FIG. 3. Distribution of q values of the 20 amino acids. The amino acids fall into two groups: "polar," large q, and "hydrophobic," small q.

the three possible combinations of the two groups: polar-polar, polar-hydrophobic, and hydrophobic-hydrophobic. The sharp division between the two groups as indicated in Fig. 3 suggests that amino acids in the same group may play similar roles in structure determination [17]. There is experimental evidence to this effect insofar as certain proteins can be designed by specifying only the HP pattern of the sequence [18]. For the purpose of protein design, the q values can serve as a useful scale for selecting amino acids.

The q values can also be used to analyze the relation between sequence and structure. In previous studies, hydrophobicity scales have been used to analyze sequences and locate helical segments [19]. However, there exist many different hydrophobicity scales. Our q scale has the advantage of being more closely related to the interactions which determine structure. We find that for a given sequence, segments with alternating large and small q values usually correspond to α helices (consistent with the previous findings using hydrophobic scales), segments with long stretches of large q values usually correspond to loops, and segments with long stretches of small q values usually correspond to β strands.

To summarize, we were able to extract the regularity of the Miyazawa-Jernigan matrix of inter-residue contact energies between amino acids using the method of eigenvalue decomposition. The analysis reveals that the dominant driving force for protein folding is the hydrophobic force and a force of demixing between amino acids. We were able to construct a solubility scale for amino acids which can be tested experimentally. This scale can be used for selecting amino acids for the purpose of protein design, and for analyzing sequence-structure relation. We would like to point out, however, that due to the statistical nature of the MJ matrix, certain features of inter-residue interactions (such as orientational dependence of the interactions, side-chain packing, etc.) are averaged out. The specific features may be necessary for building a realistic potential for protein folding.

[1] C. Anfinsen, Science **181**, 223 (1973).

[2] S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985); J. Mol. Biol. **256**, 623 (1996).

[3] D. G. Covell and R. L. Jernigan, Biochemistry **29**, 3286 (1990).

[4] D. T. Jones, W. R. Taylor, and J. M. Thornton, Nature (London) **358**, 86 (1992); J. U. Bowie, R. Lüthy, and D. Eisenberg, Science **253**, 164 (1991).

[5] To derive the MJ matrix, the position of each amino acid residue in a protein structure was taken to be the center of its side chain atom positions. Two residues were considered to be in contact if the distance between corresponding center points was smaller than $r_C \approx 6.5$ Å. The contact energy e_{ij} between type i and type j amino acids was obtained by counting the number of contact between these two types of amino acids in the protein structure data bank. There are two energy matrices in the MJ paper. One matrix gives the contact energies $e_{ij} \equiv E_{ij} + E_{00} - E_{0i} - E_{0j}$, where E_{ij} measures the absolute contact energy, and index 0 refers to the solvent molecule. Thus e_{ij} measures the energy cost of forming type i-j contact inside the solvent. This is the matrix we analyze. The reference state is a state in which amino acids are randomly mixed with water molecules, and the effective number of water molecules is estimated from the volume a protein occupies when performing self-avoiding random walk.

[6] For a review, see R. L. Jernigan and I. Bahar, Curr. Opin. Struct. Biol. **6**, 195 (1996).

[7] E. I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994); V. S. Pande, A. Yu. Grosberg, and T. Tanaka, Phys. Rev. E **51**, 3381 (1995).

[8] J. H. Hildebrand and R. L. Scott, *The Solubility of Nonelectrolytes* (Reinhold Publishing Corporation, New York, 1950).

[9] Such a subtraction procedure is necessary to remove a trivial source of a large eigenvalue. Any matrix with a nonzero mean m_0 can have one dominant eigen-value proportional to Nm_0 if the dimension N of the matrix is large. Removing this trivial regularity enables us to clearly identify other intrinsic regularities which could be obscured in the spectrum of the unsubtracted matrix.

[10] A Hamiltonian containing the C_1 term has been derived for amphiphilic copolymers, see T. Garel, L. Leibler, and H. Orland, J. Phys. II (France) **4**, 2139 (1994). The more general form with C_2 has been used in S. P. Obukhov, J. Phys. A **19**, 3655 (1986). See also, T. Garel and H. Orland, Europhys. Lett. **6**, 597 (1988).

[11] R. Eisenschitz and F. London, Z. Phys. **60**, 491 (1930).

[12] Such a hypothetical physical process has been employed previously by Thomas and Dill. P. D. Thomas and K. A. Dill, J. Mol. Biol. **257**, 457 (1996).

[13] Y. Nozaki and C. Tanford, J. Biol. Chem. **246**, 2211 (1971); M. Levitt, J. Mol. Biol. **104**, 59 (1976); M. A. Roseman, J. Mol. Biol. **200**, 513 (1988).

[14] W. W. Graessley *et al.*, Macromolecules **28**, 1260 (1995).

[15] For a review, see K. A. Dill, Biochemistry **24**, 1501 (1985); **29**, 7133 (1990).

[16] A. M. Lesk and C. Chothia, Biophys. J. **32**, 35 (1980).

[17] We believe this is true for determining the overall tertiary structure. Determining the detailed local structure requires more specificity; thus the distinctions between amino acids in the same group could be important. Recently Jernigan and Bahar argued that classification of amino acids into two groups (H and P) is inadequate based on contact energies at shorter distance (with $r_C = 4$ Å) [6]. Our results indicate that H-P is a valid classification for contacts more loosely defined (with $r_C = 6.5$ Å), corresponding to structures coarse grained over a large length scale.

[18] S. Kamtekar *et al.*, Science **262**, 1680 (1993).

[19] M. Schiffer and A. B. Edmundson, Biophys. J. **7**, 121 (1967); V. I. Lim, J. Mol. Biol. **88**, 873 (1974); D. Eisenberg, R. M. Weiss, and T. C. Terwilliger, Nature (London) **299**, 371 (1982).