Codability criterion for picking proteinlike structures from random three-dimensional configurations

Hai-Bo Cao,[1] Cai-Zhuang Wang,[2] Drena Dobbs,[3] Yungok Ihm,[1] and Kai-Ming Ho[1]
[1]*Department of Physics and Astronomy, Iowa State University, Ames, Iowa 50011, USA*
[2]*Ames Laboratory-US DOE, Iowa State University, Ames, Iowa 50011, USA*
[3]*Department of Genetics, Development and Cell Biology, Iowa State University, Ames, Iowa 50011, USA*
(Received 6 October 2004; revised manuscript received 24 July 2006; published 28 September 2006)

We show that the dominant eigenvectors of real protein structural contact matrices are highly correlated with their amino acid sequences. These results suggests that an *ab initio* sequence-independent profile exists for every protein structure and that this profile is highly effective in differentiating the ordering of amino acids in natural protein sequences from random sequences. This profile provides a structural code and is a key for understanding the unique behavior of protein structures. Using a lattice model, we show that there are special codable structures highly separated from random structures in the dominant eigenvector space of their structural contact matrices. As an example, we show our results provide a good explanation to the "designable principle" of protein structures.

DOI: 10.1103/PhysRevE.74.031921 PACS number(s): 87.15.He, 87.15.By, 87.15.Cc

Deciphering the "structural code" that dictates how an amino acid sequence folds into a unique three-dimensional (3D) protein structure is a key problem in structural biology [1–3]. Nature is extremely selective in choosing the polypeptide sequences and native structures of proteins [4]. Among $\sim 20^{300}$ theoretically possible polypeptide sequences, only about 10^{12} occur in nature [5]. These natural proteins fall into about 2000 distinct structural families [6,7]. Many proteins with similar structures have very little sequence similarity [8]. This suggests that only part of a protein's sequence information is important in determining its native structure. From the viewpoint of information coding, a random 3D polypeptide structure contains many degrees of freedom which in general, cannot be encoded with just N amino acids with 20 restricted discrete choices. However, the fact that protein sequences encode unique structures indicates that protein structures must satisfy the criterion of being "codable." To investigate which components of a protein sequence are responsible for determining the protein's native configuration, it is advantageous to examine information which is carried inherently in the native structure without explicit consideration of the amino acid sequence information. This sequence-independent information can be viewed as a global consensus of the "structural coding" of those sequences that can adopt a given structure despite their sequence diversity. In this paper we show that such a structural code can be obtained by studying the correlation between real protein sequences and structures from a contact energy perspective.

In a contact energy scheme [9], the three-dimensional structure of a protein is represented by a $n \times n$ contact matrix C where n is the number of residues. The element $C_{i,j}$ of C is assigned a value of 1 if the ith and jth residue are in contact, otherwise, $C_{i,j}$ is 0. Two residues are considered to be in contact when they are not neighboring in sequence and the geometrical distance between them is within a certain cutoff distance. If two residues are in contact, a contact energy is assigned according to the residue types. The total interaction energy for a given protein structure is the summation of all pairwise contact energies of the conformation. There are various ways to weight contact energies for different residue pairs of the 20 naturally occurring amino acids. The simplest is the HP model [10–12] in which the amino acids are classified as hydrophobic (H) and polar (P), and pairwise contact energy is assigned according to the three different types of contact: H-H, H-P, and P-P. A more advanced scheme, the statistical potential obtained by Miyazawa and Jernigan [13–16] is a 20×20 matrix (MJ matrix) obtained from the contact statistics of different types of residues in the protein structure database. Li, Tang and Wingreen (LTW) [17] showed that the MJ matrix can be approximated as $E_{i,j} = c_2(q_i+a)(q_j+a)+E_0$ where q_i is a measure of the hydrophobicity of residue type i [17]. c_2, E_0, and a are constants. By replacing the value q_i by $q_i' = q_i + a$, it can be rewritten as $E_{i,j} = c_2 q_i' q_j' + E_0$. In this paper, we will refer to the modified q' as the q value of the ith residue in the rest of this paper. In this paper, we measure the sequence-structure fitness by the difference between the energy of native sequence on a given structure and the average energy of randomly shuffled sequences on the same protein structure. Because the native sequence and shuffled sequences are calculated using the same 3D structure, the constant shift (E_0) in energy will be canceled exactly and will be neglected in the rest of this paper. The constants c_2 can be viewed as a unit of energy and will be set to be 1.

Under the HP and LTW parametrized MJ interaction schemes, the sequence of a protein can be represented by a sequence vector $\mathbf{S}$ whose elements are either 1 or 0 for the HP model, or the corresponding q values of residues using the LTW representation. For a protein with contact matrix C, the conformational energy can be written as

$$E = \langle \mathbf{S} | C | \mathbf{S} \rangle. \tag{1}$$

Note that the energy form of Eq. (1) is similar to a standard quantum system [18] with C as its Hamiltonian. The

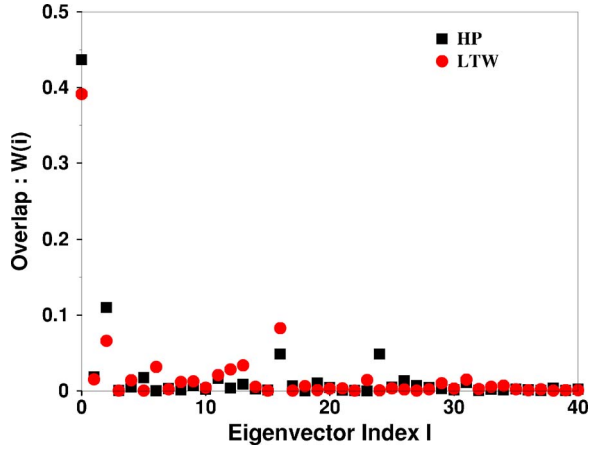

FIG. 1. (Color online) An example of the overlap ($W_i = |\langle \mathbf{S}|\mathbf{V}\rangle|^2$) between a protein sequence and eigenvectors of the protein's native structure contact matrix (protein PDB id 1a0b). Square and circle are results using HP and LTW scheme, respectively. Eigenvectors are ranked in decreasing order according to their eigenvalues. ($|\mathbf{V_0}\rangle$ is the vector with the largest eigenvalue.)

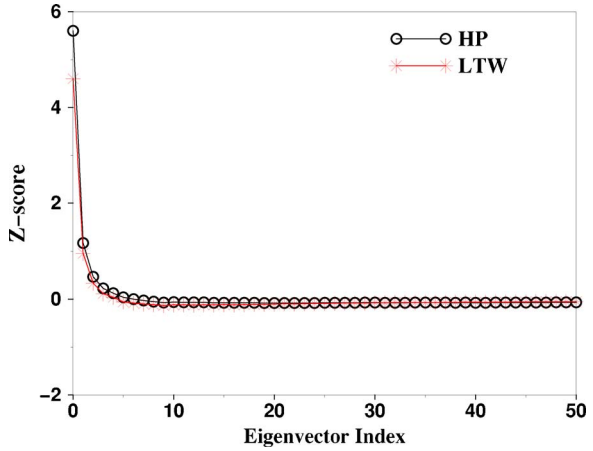

FIG. 2. (Color online) Average Z scores of different eigenvectors for 13 340 representative proteins [21] from ASTRAL database. 1000 shuffles are used to calculate Z-score Z_i^k for each protein k with eigenvector i. The average Z scores over the 13 340 proteins ($Z_i^{\text{ave}} = \Sigma_{k=1,13\,340} Z_i^k / N_{\text{total}}$, where N_{total} is the total number of proteins involved) were calculated. Results for the first 50 eigenvectors are plotted here. Results using HP model and LTW scheme are represented by circles and stars, respectively.

difference is in the vector space. For a quantum system, elements in vector $\mathbf{S}$ can be any complex number, while for a protein system, the elements in vector $\mathbf{S}$ are limited to the 20 q values (LTW) or (1,0) (HP model). C can be decomposed into its eigenstates $|\mathbf{V_i}\rangle$ corresponding to eigenvalues λ_i, i.e., $C = \Sigma_i \lambda_i |\mathbf{V_i}\rangle\langle\mathbf{V_i}|$ where λ_i is the solutions of the equation $C|\mathbf{V_i}\rangle = \lambda_i|\mathbf{V_i}\rangle$. Thus the total conformational contact energy can be expressed as the summation of the individual contribution of the eigenvectors of $C: E = \Sigma_i \lambda_i W_i$, where $W_i = |\langle \mathbf{S}|\mathbf{V_i}\rangle|^2$ is the overlap between sequence vector $|\mathbf{S}\rangle$ and eigenvector $|\mathbf{V_i}\rangle$.

For a quantum system, the ground state is $|\mathbf{V_0}\rangle$, with the W_i spectrum: $W_0 = 1$, $W_i = 0$ if $i \neq 0$. For a protein, however, because the vector space $|\mathbf{S}\rangle$ is restricted by the 20 naturally occurring amino acids, $|\mathbf{V_0}\rangle$ is generally unreachable, and W_i might be different from that of the quantum system even though the contact energy is indeed optimized in the protein folding process.

An example of the overlap W_i between a protein sequence and the eigenvectors of its native structure contact matrix is shown in Fig. 1 which shows the W_i spectrum of the protein structure 1a0b [Phosphotransfer Domain of E coli, Protein Data Bank [19] (PDB) id 1a0bi] using HP sequence and LTW q values. It is interesting to note that the dominant contribution to the contact energy comes from the first eigenvector ($\frac{\lambda_0 W_0}{E} = 0.688$ using LTW q values). The strong correlation between a protein's sequence vector and the dominant eigenvector of its native contact matrix implies that a linear structural code exists for a given protein's native structure. We hypothesize that this structural code differentiates a correctly ordered protein sequence from a "random" sequence. To test this idea, we compared the W_i spectrum of the native sequence with the spectrum that was obtained from those randomly shuffled of the same sequence. We calculated the "Z score" [20] of $W(i)$,

$$Z_i = \frac{W_i - \overline{W_i}^{\text{shuffle}}}{\delta}, \tag{2}$$

where $\overline{W_i}^{\text{shuffle}}$ is the average W_i for shuffled sequences, and δ is the root-mean-square deviation of the shuffled sequences. To get good statistics, we selected 13 340 representative single-domain structures [21] from the ASTRAL database [8] and calculated Z_i for each protein.

The average Z scores over all the structures are plotted in Fig. 2 for both the HP model and the LTW-MJ scheme to evaluate the sequence-eigenvector correlation for each eigenvector. According to Fig. 2, only the dominant eigenvector is statistically significant (Z score greater than 2.0) in distinguishing the correct ordering of amino acids that will fold to a given structure. This means that the majority of the eigenvectors are "blind" to the sequence order. Thus, in the process of assessing the sequence-structure fitness for a given amino acid sequence and a given protein structure one can consider the dominant eigenvector only instead of the entire contact matrix.

Because each eigenvector can be viewed as a "mode" and together these eigenvectors form a complete set, the correlation between the protein sequence and the dominant eigenvector of the contact matrix discussed above implies a way that a protein sequence encode its structural information. In natural proteins, the structural information is encoded by incorporating the dominant mode of the protein's native structure into the ordering of its hydrophobicity profile. This information will be recognized in a folding process that optimize hydrophobic interaction energy. The remaining modes (eigenvectors) with small eigenvalues are much less important than the dominant "coding mode" in determining the contact matrix. These other degrees of freedom may be used to incorporate other important aspects of protein function (e.g., enzyme active sites).

The observed correlation between a protein's sequence and the dominant eigenvector of its contact matrix provides a framework for mapping a 3D structure to a one-dimensional (1D) sequence. This framework should be useful in several applications. In previous work [21], we have shown that an effective structural threading method based on this correlation can be constructed. This one-dimensional representation of protein structure may also be useful in studies of the evolution of protein structures [22]. In particular, this framework can be used to understand the "designability principle" of protein folding proposed by Li, Tang, and Wingreen [23]. The designability of a given protein structure is the number of sequences which adopt that structure as their lowest energy state (ground state). The "designability principle," proposed by Li *et al.* based on a HP lattice model, states that natural protein structures correspond to highly designable structures because they have higher thermal stability than other configurations. Recently, Li *et al.* [24,25] postulated that, for a given 2D lattice structure, there is a corresponding vector in sequence space, such that all neighboring sequences adopt that structure as the ground state. Therefore, the density of structures in the neighborhood of this vector is a measure of the designability of the protein. In their study, this vector is the solvent exposure profile of the protein.

Here we propose that the dominant eigenvector of the contact matrix provides a better 1D structural representation of a protein. Eigenvector studies of protein contact matrix support this assumption [26,27]. Because the sequence vector and the dominant eigenvector of the contact matrix have the same dimension, a protein's structure and sequence can be represented by points in the same n- (where n is the length of sequence) dimensional vector space. The distance between the vectors of a sequence and a structure in this n-dimensional space is related to their vector dot product. Minimization of the free energy due to the hydrophobic interactions causes a polypeptide sequence to fold to a structure in its neighborhood in the n-dimensional space. However, if other structures are also close to the sequence vector, the presence of competing "decoy" structures having similar energies to the ground state will lead to a rugged energy landscape, and prevent the protein from adopting a unique fold. To achieve a "funnel"-like energy landscape [28,29], the native structure of a protein must be far away from other compact structures in the n-dimensional space. So the "designability" of a structure should be inversely related to the density of structures around it.

The above conjecture can be tested on the case of lattice proteins where all possible contact matrices can be enumerated. We restrict ourselves to a $3\times3\times3$ cubic lattice used by Li *et al.* and many others [23]. In agreement with past studies we found a total of 103346 possible different contact matrices. We enumerate all possible HP sequences on these lattice structures using the same interaction scheme as Li *et al.* (H-H, −2.3; H-P, −1; P-P, 0) [23] to obtain the designability of each lattice configuration. For each structure, the dominant eigenvector of its contact matrix is also generated as its structural coding vector to map the structure onto the 27-dimension vector space. To measure the density of structures, we calculate all pairwise overlaps for all lattice structural coding vectors. The overlap becomes large when two vectors

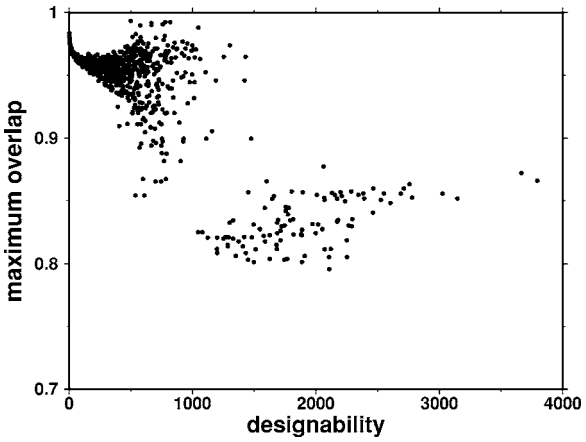

FIG. 3. The correlation between designability and the largest overlap of its eigenvector with all other eigenvectors ζ_i [see Eq. (3)] obtained from $3\times3\times3$ cubic lattice model.

are close to one another with a maximum value of 1 when the two vectors coincide. For a given lattice configuration i, we define its distance from the whole ensemble of other structures, ζ_i, as the largest overlap of its eigenvector with all other eigenvectors,

$$\zeta_i = \max(|\langle \mathbf{V}_i^0 | \mathbf{V}_j^0 \rangle|), \tag{3}$$

where j goes from 1 to 103 347, and $j \neq i$. Thus, ζ_i is inversely related to the density of structures around a given structure i.

The relationship of ζ_i and the designability for all 103 346 structures is plotted in Fig 3. All structures with designability greater than 1500 correspond without exception to vectors with very few neighbors: with ζ values are smaller than 0.88. The clear separation between low designability compact structures and highly designable "proteinlike" structures shown in Fig. 3 reveals a unique property of the "protein like" structures: the dominant eigenvectors of their contact matrices are well separated from those of other random compact structures. This property is a key requirement for the dominant eigenvector to provide a "structural code," because the dominant eigenvector of a randomly chosen contact matrix does not fully determine the contact matrix. Contact matrices from different structures may have indistinguishable dominant eigenvectors. Only those dominant eigenvectors well separated from other dominant eigenvectors can provide a unique and robust mapping to their corresponding contact matrices. This result is significant because it shows how the 3D structural information of a protein containing many degrees of freedom (at least $2N$ real numbers) can be encoded with just N amino acids of 20 discrete choices for special "proteinlike" structures. This codability requirement may place severe constraints on the number of possible protein folds that can exist in nature.

In summary, our study shows that a strong correlation exists between a protein sequence and the dominant eigenvector of its structure contact matrix. The dominant eigenvector provides an *ab initio* sequence-independent structural profile for sequence-structure mapping of proteins. This 1D profile gives a better explanation of the "designability prin-

ciple" found in the HP lattice models, and can be further extended to real off-lattice protein structures.

The authors would like to thank Robert Jernigan and Amy Andreotti for helpful discussions. The authors would like to thank the Institute for Physical Research and Technology, the Plant Science Institute, the Biotechnology Council, and the Lawrence H. Baker Center for Bioinformatics and Biological Statistics at Iowa State University for financial support and the Scalable Computer Laboratory for computational support in this project.

[1] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).
[2] R. Bonneau and D. Baker, Annu. Rev. Biophys. Biomol. Struct. **30**, 173 (2001).
[3] A. L. Watters and D. Baker, Eur. J. Biochem. **271**, 1615 (2004).
[4] C. B. Anfinsen, E. Haber, M. Sela, and F. H. White, Proc. Natl. Acad. Sci. U.S.A. **181**, 223 (1961).
[5] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. J. Wheeler, Nucleic Acids Res. **32**, D23-6 (2004).
[6] J. M. Chandonia, N. S. Walker, L. Lo Conte, P. Koehl, M. Levitt, and S. E. Brenner, Nucleic Acids Res. **30**, 260 (2002).
[7] S. E. Brenner, P. Koehl, and M. Levitt, Nucleic Acids Res. **28**, 254 (2000).
[8] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, J. Mol. Biol. **247**, 536 (1995).
[9] M. Levitt and C. Chothia, Nature (London) **261**, 552 (1976).
[10] H. S. Chan and K. A. Dill, J. Chem. Phys. **92**, 3118 (1990).
[11] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[12] K. A. Dill, S. Bromberg, K. Z. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan, Protein Sci. **4**, 561 (1995).
[13] S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985).
[14] S. Miyazawa and R. L. Jernigan, J. Mol. Biol. **256**, 623 (1996).
[15] S. Miyazawa and R. L. Jernigan, Proteins **36**, 347 (1999).
[16] S. Miyazawa and R. L. Jernigan, Protein Eng. **13**, 459 (2000).
[17] H. Li, C. Tang, and N. S. Wingreen, Phys. Rev. Lett. **79**, 765 (1997).
[18] E. Merzbacher, *Quantum Mechnics* (Wiley, New York, 1970).
[19] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucleic Acids Res. **28**, 235 (2000).
[20] S. H. Bryant and S. F. Altschul, Curr. Opin. Struct. Biol. **5**, 236 (1995).
[21] H. B. Cao, Y. Ihm, C. Z. Wang, J. R. Morris, M. Su, D. Dobbs, and K. M. Ho, Polymer **45**, 687 (2004).
[22] G. Tiana, B. E. Shakhnovich, N. V. Dokholyan, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **101**, 2846 (2004).
[23] H. Li, C. Tang, and N. S. Wingreen, Science **273**, 666 (1996).
[24] H. Li, C. Tang, and N. S. Wingreen, Proc. Natl. Acad. Sci. U.S.A. **95**, 4987 (1998).
[25] N. S. Wingreen, H. Li, and C. Tang, Polymer **45**, 699 (2004).
[26] J. L. England and E. I. Shakhnovich, Phys. Rev. Lett. **90**, 218101 (2003).
[27] M. Porto, U. Bastolla, H. E. Roman, and M. Vendruscolo, Phys. Rev. Lett. **92**, 218101 (2004).
[28] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **87**, 7524 (1987).
[29] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins **21**, 167 (1995).