Fast method for estimating the energy distribution of globular states of proteins

Hai-Bo Cao,[2] Cai-Zhuang Wang,[2] and Kai-Ming Ho[1,2]
[1]*Department of Physics and Astronomy, Iowa State University, Ames, Iowa 50011, USA*
[2]*Ames Laboratory-U.S. DOE. , Iowa State University, Ames, Iowa 50011, USA*

(Received 31 August 2004; published 22 August 2005)

By an enumeration study, we show that the energy distributions of a lattice protein sequence on all possible compact lattice configurations can be approximated by the energy distribution of shuffled sequences on a given lattice structure. We also show that the random energy model (REM) gives a good analytical approximation for the energy distribution of shuffled sequences on lattice structures. For real proteins, when a gapped threading method is used, REM calculations systematically underestimate the mean value of the energy distributions. We found that this discrepancy can be roughly compensated by a linear correction obtained from empirical fits. This result can be used to greatly reduce the computational effort in protein threading calculations.

DOI: 10.1103/PhysRevE.72.021907 PACS number(s): 87.15.−v, 87.14.Ee, 82.30.Vy

I. INTRODUCTION

The energy distribution of protein molten globular states [1] is important for understanding protein folding dynamics [2–4]. Wolynes and co-workers proposed a "funnel"-like energy landscape to explain the fast folding process of natural proteins [2]. According to this assumption, the ability of a protein finding its native configuration [5] correlates with the energy difference between the energy of native state and the average energy of its molten globular states. The energy distribution of a protein's molten globular states is the density of states of the given polypeptide in all physical compact configurations. However, for a real protein, it is an outstanding challenge to map out this distribution, because the conformation space of a real protein is so large that it is not feasible to enumerate all the possible compact structures. Any current method in sampling protein energy distributions will only be able to select a very small fraction of the entire landscape. Therefore, the resulting energy distribution might not be representative. On the other hand, amino acid sequence shuffling is commonly used to study the significance of protein sequence-structure compatibility (Z score) [6,7]. When a protein sequence is put onto its native structure, we expect a significantly lower energy than putting the protein sequence onto a random structure. However, if we have a scrambled sequence, it ought not matter whether the structure adopted is the native structure of the unshuffled sequence or a random structure. Thus, by using randomly shuffled sequences, we should be able to estimate the energy distribution of a protein's molten globular states. The results should be insensitive to which structure we choose and one convenient way is to impose the protein's "shuffled" sequences onto its native structure [8].

In this paper, we want to investigate how well the energy distribution of the shuffled sequences represent the energy distribution of the various molten-globule states for a given protein sequence. Using a simple lattice HP model [9–13], we compare the energy distributions from exact enumeration and the "shuffled sequence" scheme. Our study shows that the energy distribution for a given sequence on all possible three dimensional lattice structures can be effectively reproduced by shuffling the amino acid sequence while keeping its conformation as the native structure of the protein. This observation is useful because it is much easier to model structural evolution in sequence space than following the structural evolution of a connected chain in real space. Moreover, we found that the random energy model [14] (REM) gives good estimates for the energy distribution for shuffled sequences. Earlier work [15] has established this for the case of gapless threading onto a given structure. Here, we showed that good estimates can still be obtained even when a gapped threading method is involved. This makes it possible to establish a fast prescreening method in structural threading approaches for protein structure prediction.

II. THE LATTICE HP MODEL

In the lattice HP model, a polypeptide chain is modeled as a self-avoiding random walk on a square or cubic lattice. Amino acids are divided into two groups: hydrophobic (H) residues or polar (P) residues. Each of the residues occupies a site on a lattice. For a given lattice polypeptide chain configuration, those residue pairs that are geometrical neighbors but not adjacent in sequence are considered to be "in contact," and a contact energy is assigned. We use the energy scheme proposed by Chan and Dill [11,12] where contact energy for H-H residues are 1, and all other kinds of contacts are assigned energy 0.

In this paper, we use a $3\times3\times3$ cubic lattice which had been popular in previous studies [16–18]. For this lattice HP model, it is possible to enumerate all the compact structures that a HP sequence can have. Using the HP interaction scheme, a lattice protein sequence can be represented by a sequence vector q_i ($q_i=1$ if the ith residue is hydrophobic, otherwise, $q_i=0$). The lattice configuration can be represented by a "contact matrix" $C_{i,j}$ ($C_{i,j}=1$ if ith residue and jth reside are in contact, otherwise $C_{i,j}=0$). The energy of a given polypeptide configuration can be written as

$$E=\sum_{ij} c_{i,j}q_iq_j. \qquad (1)$$

According to this contact energy scheme, two lattice configurations are considered to be the same if they have a com-

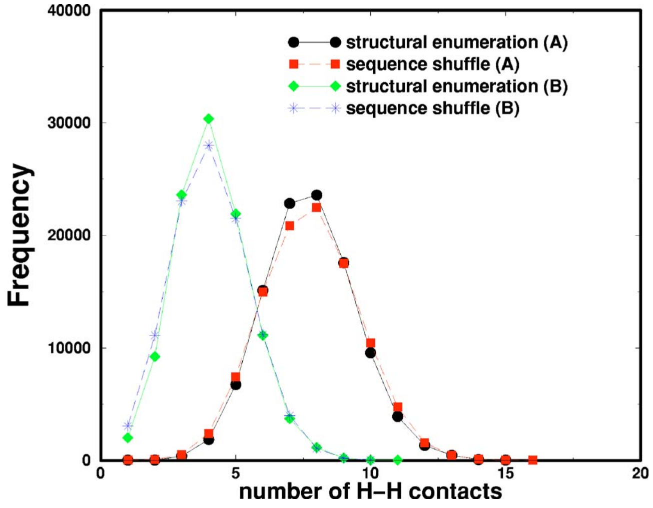

FIG. 1. (Color online) Comparison of energy distributions from structural enumeration and sequence shuffling. Two lattice proteins sequences on 3×3×3 lattice are selected for this study. (Lattice protein A:011010001111010110110100001, lattice protein B:011011000010010100000011001.) For each case, 103 346 (the same number as all distinct lattice structures) shuffled sequences are generated in each for this comparison.

mon contact matrix. (Here we consider two configurations with reverse labeling symmetry as two different configurations if their contact matrices are different.) There are 103 346 distinct contact matrices on a 3×3×3 cubic lattice. A lattice *HP* sequence is consider to be "proteinlike" if it can be mapped onto a unique lattice configuration which gives the highest number of *H-H* contacts among all the 103 346 different structures. On a 3×3×3 cubic lattice, there are 2^{27} different *HP* sequences. Through the enumeration of the energies of all these sequences on their possible configurations, we found that only 8140 444 of them are proteinlike.

III. SEQUENCE SHUFFLING SCHEME

For a given proteinlike sequence, its energy distribution can be obtained by mapping the sequence onto all the 103 346 cubic lattice structures. For comparison, we also calculated the energy distribution by randomly shuffling the sequences but keeping the structure as the native structure of the given sequence. As an example, Fig. 1 shows the comparison of the energy distributions from exact enumeration and the shuffled scheme for two protein-like sequences. As one can see from Fig. 1, the energy distributions of the lattice proteins are closely reproduced by random sequence shuffling. In order to get a good statistics, we randomly chose 1330 different sequences which fold on different lattice configurations from the 8140 444 proteinlike *HP* sequences for our study. We calculate the mean values and standard deviations of each chosen sequence using both structural enumeration and sequence shuffling. The results are shown in Figs. 2(a) and 2(b). From Fig. 2, we can see a linear correlation between the distributions from the two methods. This indicates that the energy distribution of lattice proteins can be effectively reproduced by the sequence shuffling method.

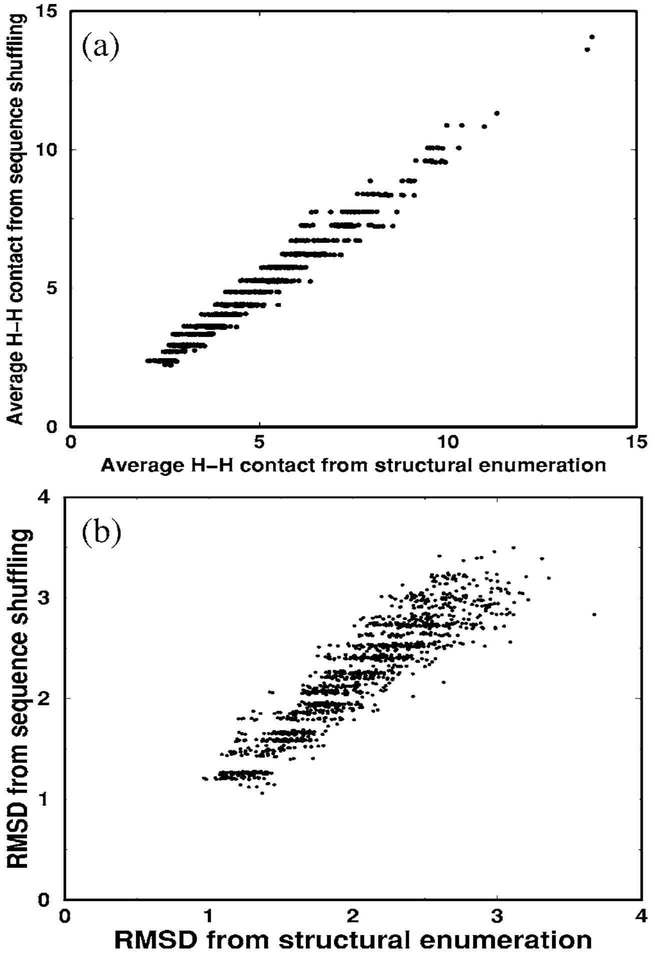

FIG. 2. The comparison of (a) average energy and (b) RMSD from exact structural enumeration and sequence shuffling. 1330 different lattice protein from a 3×3×3 cubic lattice were selected for the study. Each of these sequences were shuffled to generate 103 346 randomized sequences. The energies of these shuffled sequences on the corresponding native configuration were collected to calculate energy mean and RMSD. For each of the selected sequence, we also enumerate all possible 3×3×3 lattice configurations to get energy mean and RMSD. The unit for all axes in this figure is the number of *H-H* contacts.

In this paper, we enumerate only the maximally compact conformations on a 3×3×3 lattice. However, such maximally compact conformations may not always be the true ground-state structure when conformational search is not restricted and extended to all accessible lattice conformations [9]. In this regard, the complete density of states (instead of the one restricted to the 3×3×3 conformations here) estimated, for example, by Monte Carlo sampling would be more relevant and will be the subject of further studies. One should also note that in the 3×3×3 lattice model there is only one buried residue while the other 26 residues are all exposed to solvent. In real proteins, much higher ratio of buried residues is observed. It will be interesting to see if our conclusion for the 3×3×3 lattice model can be generalized to more realistic protein models.

IV. RANDOM ENERGY MODEL

The REM, introduced by Derrida to study spin glass [14], has been widely used to study the thermodynamical properties of polymers and proteins [2,18–22]. It has been shown that the energy distribution for the shuffled sequences can be analytically estimated using the REM approximation [15,18]. In particular, Mirny *et al.* [15] have shown that the energy distribution of the shuffled sequence on real protein structures in gapless threading can be approximated by REM. In this paper, we show that REM is also a good model to describe the shuffled sequence for all protein like sequences on the $3\times3\times3$ lattice. However, we found that the REM systematically underestimates the average threading scores for the shuffled sequences on real protein structures if a gapped threading method is used.

In REM model, the energy distribution of random shuffled sequences is approximated by a Gaussian function

$$P(E^s) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-(E^s - E^{\text{ave}})^2/2\sigma^2\right)}, \qquad (2)$$

where E_{ave} is the average of the distribution and σ is the root mean square deviation. Because we are considering only pairwise contact interactions, E_{ave} and σ can be calculated as [15]

$$E^{\text{ave}} = \frac{C}{C_{\text{total}}} \sum_{i,j} U_{i,j} \qquad (3)$$

and

$$\sigma = \sqrt{C}\left[\frac{1}{C_{\text{total}}} \sum_{i,j} U_{i,j}^2 - \frac{1}{C_{\text{total}}^2}\left(\sum_{i,j} U_{i,j}\right)^2\right]^{1/2}, \qquad (4)$$

where C is the total number of contacts in the configuration, and C_{total} is the total number of possible pairwise contacts between residues. $U_{i,j}$ is the contact energy of residue i and residue j if they are in contact.

Figure 3 shows that the energy distribution of all the proteins in the $3\times3\times3$ lattice calculated using REM in comparison with the results from shuffled sequences. The average *H-H* contact from shuffled sequence is in good agreement with the REM estimation as one can see from Fig. 3(a). The root mean square deviation (RMSD) from the two methods are also similar. These results indicate that the REM can provide good estimates of the energy distribution for shuffled sequences in the lattice model.

When gaps are permitted in the threading process, the resulting distribution does not obey the REM estimation. This is because, in the process of optimizing the sequence-structure fitness in the alignment process, bias was introduced toward the native structures. Fortunately, because the shuffling process completely destroy any correlation between the given sequence and structure, the chance for a threading method to find a long segment of good alignment are very small. Thus, we expect that the overall distribution (in terms of mean value and root mean square deviation) using a gapped threading method should still be correlated with the REM estimate. We chose 100 protein domains from the ASTRAL [23] database for investigation. For each of the 100 proteins, we calculated the mean value of the shuffled sequence's energy distribution using a gapped threading [8] approach as well as the REM approach. In our threading method, pair-wise contact energy between residues were calculated using a simplified Miyazawa–Jernigen matrix [24,25]. The total threading energy is the summation of contact energies from each residue-residue pair. The correlation between the two distributions (calculated using REM and actual sequence shuffling) is shown in Fig. 4. The calculations show that the mean value of the scores from the threading approach is systematically higher compared with the REM estimated value, because the threading method can optimize the alignment of sequence and structure to find the optimum conformation energy. However, it is interesting to note that the difference between the REM and threading score can be fitted by a linear line as shown in Fig. 4. Therefore, this difference can be empirically corrected using this fitting result.

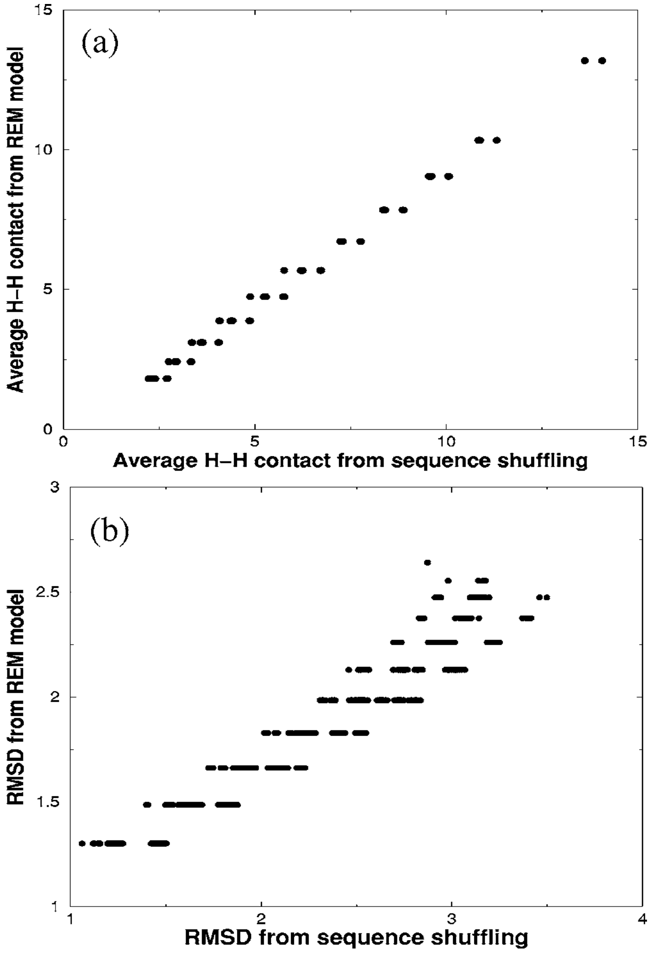

FIG. 3. The comparison of (a) average energy and (b) RMSD from REM estimation and sequence shuffling on a $3\times3\times3$ cubic lattice. The lattice proteins used are the same as those in Fig. 2. The unit for all axes in this figure is the number of *H-H* contacts.

V. APPLICATION IN THREADING CALCULATIONS

The result shown in Fig. 4 would be useful for structural threading studies which involves the calculation of average

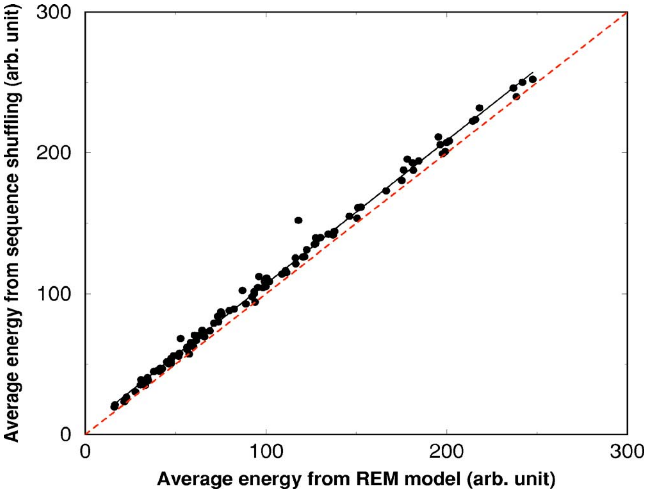

FIG. 4. (Color online) Comparison of average threading energies from sequence shuffling and REM approximation for real proteins. 100 nonredundant proteins were selected from ASTRAL database for this study. For each protein, we shuffled its amino acids sequence to generate 1000 randomized sequences. These shuffled sequences were aligned to the corresponding native structure by a gapped threading method (Ref. [8]). The resulting sequence-structure alignment is used to calculate the threading energies. For comparison, the REM estimation using same contact interaction model were also calculated for each protein. Note that the underestimation by REM can be empirically compensated use the linear fitting shown by the solid line in the figure.

score E_{ave} for shuffled sequences [8]. As has been shown in the above lattice studies, the energy distributions for shuffled sequences is similar to the energy distributions of lattice proteins in compact configurations. Therefore, it is reasonable to believe that the E_{ave} calculated using shuffled sequence for a real protein represent the average energy for molten globular states in protein folding process. Thus for a threading method using a pairwise interaction scheme, the relative energy $E_{\mathrm{relative}}=E_{\mathrm{threading}}-E_{\mathrm{ave}}$, where $E_{\mathrm{threading}}$ is the threading score using the original sequence, measures the energy difference between the native state and the molten globular state. E_{relative} has been shown to be a better scoring function in determining the fitness of sequence-structure pairs [8]. Since the calculation of average score E_{ave} is the most time-consuming process, the speed of gapped threading methods can be increased significantly if E_{ave} is estimated analytically using REM. Even though the REM estimates may not be exact, the effects of such inaccuracy on the result of threading is negligible if the REM calculation is used as an initial screening tool. Thus, although the REM calculation is not accurate enough to distinguish high score sequence-structure pairs, it is good enough to distinguish low score sequence-structure pairs from high score sequence-structure pairs. In genome-wide threading studies, an overwhelming majority of sequence-structure pairs have average scores E_{ave} far below the threshold value beyond the errors of the REM. The REM calculation can be used to screen out such low score alignments, so that more accurate E_{ave} calculations are used to investigate only the most promising sequence-structure pairs.

As an example, we applied this prescreening technique in our threading method [8] for genomic scale search. In our threading approach, we use the relative energy E_{relative} instead of the Z score as scoring function for measuring sequence-structure fitness. A given query protein sequence is threaded on a representative structure set which consist of more than 13 000 PDB structures. Before this prescreening technique was developed, for each sequence-structure pair, 20 shuffled sequences were generated to calculate E_{ave} due to limited computational resources. Thus, without prescreening process, 21 pair-wise threading calculation has to be performed to generate E_{relative}.

With the prescreening process, an estimated relative score E_{estimate} is calculated using the REM model. Thus, there is only one pair-wise threading calculation needed (to calculate $E_{\mathrm{threading}}$) for each sequence-structure pair. All protein structures in the representative database are ranked according to their corresponding estimated relative score (E_{estimate}). In practice, we found that all hits above threshold can be found in the top 1%–5% of promising structures according to E_{estimate}. This prescreening process resulted in 10–16 times speed up of the threading process for sequence-structure studies involving large databases.

VI. CONCLUSION

Using a $3\times3\times3$ cubic lattice model, we compared the distribution of energies from two different approaches: enumerating all the possible compact configurations of a given sequence vs shuffling the sequence on a fixed native configuration. We found that the energy distribution obtained from these two approaches are similar and the latter one can be well described by the REM approximation. For real proteins, we found that REM calculations systematically underestimate the mean value of the energy distribution if the energy calculation involves gapped threading. However, the discrepancy can be corrected empirically. These results can be used to reduce substantially the computational efforts involve in protein structure predictions using gapped-threading methods.

[1] O. B. Ptitsyn, in *Protein Folding*, edited by T. E. Creighton (W. H. Freeman, New York, 1992), p. 243.

[2] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins **21**, 167 (1995).

[3] R. Miller, C. A. Danko, M. J. Fasolka, A. C. Balazs, H. S. Chan, and K. A. Dill, J. Chem. Phys. **96**, 768 (1992).

[4] K. A. Dill and H. S. Chan, Nat. Struct. Biol. **4**, 10 (1997).

[5] C. B. Anfinsen, E. Haber, M. Sela, and F. H. White, Proc. Natl. Acad. Sci. U.S.A. **181**, 223 (1961).

[6] S. H. Bryant and S. F. Altschul, Curr. Opin. Struct. Biol. **5**,

236 (1995).

[7] J. Meller and R. Elber, Proteins **45**, 241 (2001).

[8] H. B. Cao, Y. Ihm, C. Z. Wang, J. R. Morris, M. Su, D. Dobbs, and K. M. Ho, Polymer **45**, 687 (2004).

[9] K. Yue, K. M. Fiebig, P. D. Thomas, H. S. Chan, E. I. Shakhnovich, and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **92**, 325 (1995).

[10] K. A. Dill, S. Bromberg, K. Yue, K. M. Feibig, D. P. Yee, P. D. Thomas, and H. S. Chan, Protein Sci. **4**, 561 (1995).

[11] H. S. Chan and K. A. Dill, J. Chem. Phys. **95**, 3775 (1991).

[12] H. S. Chan and K. A. Dill, J. Chem. Phys. **92**, 3118 (1990).

[13] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).

[14] B. Derrida, Phys. Rev. B **24**, 2613 (1981).

[15] L. A. Mirny, A. V. Finkelstein, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **97**, 9978 (2000).

[16] H. Kaya and H. S. Chan, Proteins **52**, 524 (2003).

[17] H. Li, C. Tang, and N. Wingreen, Science **273**, 666 (1996).

[18] E. Shakhnovich and A. Gutin, J. Chem. Phys. **93**, 5967 (1990).

[19] E. Shakhnovich and A. Gutin, Biophys. Chem. **34**, 187 (1989).

[20] J. D. Bryngelson and J. D. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987).

[21] E. I. Shakhnovichn and E. M. Gutin, J. Theor. Biol. **149**, 537 (1991).

[22] A. V. Finkelstein, A. Gutin, and A. Badretdinov, Proteins **23**, 151 (1995).

[23] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, J. Mol. Biol. **247**, 536 (1995).

[24] S. Miyazawa and R. L. Jernigan, J. Mol. Biol. **256**, 623 (1996).

[25] H. Li, C. Tang, and N. S. Wingreen, Phys. Rev. Lett. **79**, 765 (1997).