Simplified calculation of folding energies and residue coordination numbers in random heteropolymers

A. E. Carlsson
Department of Physics, CB 1105, Washington University, St. Louis, Missouri 63130

(Received 19 October 1998)

I develop a formalism for calculating effective pair and higher-order interactions between residues in random heteropolymers that approximately predict the folding enthalpy and the coordination numbers of individual residues. In a simple model heteropolymer with additive couplings between residues, the folding enthalpy is written in terms of two-, three-, and four-body interactions between residues. The coordination numbers are expressed in terms of interactions between up to three residues. Application to a 6×6 square model shows that the folding enthalpy is obtained to an accuracy of better than 1%. The coordination numbers are obtained with a rms error of 1.2 neighbors. [S1063-651X(99)15905-1]

PACS number(s): 87.15.Cc, 87.14.Ee

I. INTRODUCTION

Prediction of the three-dimensional structure of proteins in terms of the amino acid sequence continues to be a daunting theoretical problem. Although more efficient algorithms are regularly being devised, there is not a single example of successful folding of a real protein on a computer. For this reason, it is of interest to investigate other properties that are related to the three-dimensional structure, but are easier to access on the basis of computer calculations. The folding enthalpy, or the free-energy difference between the fully folded state and the random coil state, is a key thermodynamic parameter in the theory of protein folding [1]. It provides the driving force for protein folding and is related to the folding temperature. Computer simulations [2] have suggested that the folding temperature is closely related to the foldability of proteins. Thus a calculation of the folding energy that avoids the full complexity of folding the protein can be very advantageous. The present calculations are aimed at the zero-temperature enthalpy $E_{\min}$ of the folded state, which is a key ingredient of the finite-temperature free energy. (Other simulations [3] have pointed to the importance of the energy gap, or the difference in energy between the lowest- and second-lowest folds; our method is not able to predict such delicate quantities with useful accuracy). The coordination number vector of a protein has components that are the coordination numbers of the residues, ordered according to their position along the chain. It provides a simplified one-dimensional description of the structure of the protein. It has been shown by Galaktionov and Marshall [4] that an accurate estimate of the coordination number vector of a protein can be used as a basis for predicting first the matrix of contacts between residues, and subsequently the three-dimensional structure, with promising accuracy. Thus the development of a simple method of predicting the coordination number vector of a protein would be very useful.

This paper presents a formalism for predicting the folding enthalpy and the coordination number vector in a model random heteropolymer in a simple fashion, using simple interactions between the residues. In addition to computational economy, such a description of the folding enthalpy may help predict the foldability of given sequences of residues. Unger and Moult [5] have argued that the reliable folding of simple model proteins is determined by the number of favorable local residue-residue interactions along the chain. The simplified description of the folding enthalpy can determine which sequences of residues have generally favorable local residue-residue interactions, so that one can predict some aspects of foldability from the primary sequence itself, without explicitly folding the protein. The method developed here includes up to four-body interactions, which renders it computationally tractable even for sequences having up to 1000 residues.

The formalism, described in Sec. II, is based on a lattice heteropolymer model in which the residues are represented by hydrophobicity energies ϵ_i, describing the extent of their preference for high coordination numbers. I examine the properties of the energy density of states (DOS) associated with a particular sequence of residues. This DOS represents the collection of energies of this sequence in the possible folds of the polymer, so that the lower end of the DOS corresponds to the minimum-energy fold. The moments of the DOS are obtained rigorously in terms of pair or multibody interactions between the residues. Using the moments, I develop an approximate folding enthalpy function motivated by the known behavior of the model in certain cases. This approach is analogous to previous applications of moment analysis in analyzing the energetics of metallic bonding [6]. The coordination number vector is then given as the gradient of the folding enthalpy function with respect to the ϵ_i. This function is straightforwardly differentiated to obtain estimates for the coordination-number vector based on interactions of up to three residues.

In Sec. III, I describe tests of this methodology in a 6×6 square lattice model with compact conformations. Diagonal as well as vertical and horizontal neighbors are included in the coordination number. At the level of two-body interactions, the folding enthalpy is obtained with an accuracy of 1.0%; the inclusion of four-body interactions improves the accuracy to 0.7%. The coordination number is obtained with a rms error of 1.31 neighbors at the level of an additive function of the hydrophobicity energies; a description including interactions of up to three residues reduces the error to 1.19 neighbors.

II. MODEL AND FORMALISM

The heteropolymer model is based on residues on a lattice. The formalism applies to general lattices; I will later consider a square lattice model. The interaction energy in a given fold has the form

$$E=1/2\sum_{i\neq j} h_{ij}\chi_{ij}\,. \tag{1}$$

Here the h_{ij} are interaction energies between residues i and j, which are determined by the amino-acid types of residues i and j. The quantity χ_{ij} describes the proximity of residues i and j in the given fold. It is 1 if i and j are neighbors, and zero otherwise. (χ_{ij} can be generalized to have several different values, allowing for neighbors at different distances, but I have not yet done so.)

The analysis of this type of model is simplified if one makes the assumption that the h_{ij} are additive, in the sense that $h_{ij}=\epsilon_i+\epsilon_j$. This approximation is well justified, as has been shown by Li *et al.* [7]. They analyzed a matrix of statistical potentials [8], between residue types derived from observed contact frequencies, and found that the interactions could be represented to good accuracy by an additive form; the corrections to this form were generally smaller than the additive terms by more than an order of magnitude. Given the additivity assumption, one has from Eq. (1) that

$$E=\sum_i \epsilon_i Z_i\,, \tag{2}$$

where Z_i is the coordination number of residue i in the given fold. In this description, the ϵ_i are energies that correspond the hydrophobicities of the residues. A large negative value of ϵ corresponds to a residue that prefers a large number of neighbors and in this sense acts hydrophobically. A similar random-hydrophobicity model, but off-lattice (continuous spatial variables), has been used by Garel and collaborators [9] and Obukhov [10] to treat some aspects of the phase diagram of proteins and other heteropolymers.

I now turn to the task of describing the zero-temperature energy of the most favorable fold in terms of the hydrophobicity energies ϵ_i. To accomplish this task, I take a fixed sequence $\{\epsilon_i\}$, and consider the density of states $\rho(E)$ that is obtained by taking the energies of that sequence in all the possible folds, or in a subset of all the possible folds. This density of states has the form

$$\rho(E)=(1/N)\sum_\alpha \delta[E-E_\alpha(\vec{\epsilon})]. \tag{3}$$

Here, N is the number of folds that are considered, and $E_\alpha(\vec{\epsilon})$ is the energy of the fold α. The folding enthalpy $E_{\min}$, or the value of E_α for the minimum-energy fold, is then the lower limit of the support of $\rho(E)$. The description of $E_{\min}$ in terms of residue-residue interactions is based on the energy moments of $\rho(E)$. These are defined as

$$\mu_1=(1/N)\int E\rho(E)dE=(1/N)\sum_\alpha E_\alpha(\vec{\epsilon}) \tag{4}$$

and

$$\mu_n=(1/N)\int (E-\mu_1)^n\rho(E)dE$$
$$=(1/N)\sum_\alpha [E_\alpha(\vec{\epsilon})-\mu_1]^n \quad (i\geqslant 2). \tag{5}$$

Thus μ_1 represents the average energy of all the folds. This may be thought of as the energy of a molten-globule state, in which all compactly folded states are equally represented. μ_2 represents the mean-square width of the energy distribution, and the higher moments describe various aspects of the shape of the distribution. From Eqs. (2), (4), and (5), one readily shows that

$$\mu_1(\vec{\epsilon})=\sum_i \epsilon_i\langle Z_i\rangle, \tag{6}$$

$$\mu_2(\vec{\epsilon})=\sum_{i,j} \epsilon_i\epsilon_j\langle \Delta Z_i\Delta Z_j\rangle, \tag{7}$$

and analogous relations hold for the higher-order moments. Here, $\Delta Z_i=Z_i-\langle Z_i\rangle$, and the brackets denote averages over the set of folds that is included in $\rho(E)$, i.e., $\langle Z_i\rangle=(1/N)\Sigma_\alpha Z_i^\alpha$, where Z_i^α is the coordination number of site i in fold α.

Thus the moments can be written in terms of simple interactions between the residues, and these interactions are determined by statistical properties of the set of possible folds. I now develop an approximate method of writing $E_{\min}(\vec{\epsilon})$ in terms of low-order moments. I first note that when $E_{\min}$ is written in terms of the moments it has the form $E_{\min}=\mu_1+f(\mu_2,\mu_3,\ldots)$, where f is a function having units of energy. This is because changes in μ_1 that do not affect the higher moments (these are defined relative to μ_1) correspond to a shift in the energy zero that does not change the difference between $E_{\min}$ and μ_1. I also note two important properties of the function $E_{\min}(\vec{\epsilon})$:

(i) If $\mu_2(\vec{\epsilon})$ vanishes, then the DOS has zero width, and $E_{\min}(\vec{\epsilon})=\mu_1(\vec{\epsilon})$. Thus $f=0$ if $\mu_2=0$.

(ii) Under uniform scaling of the ϵ_i by a positive factor, the energy must scale linearly, i.e., $E_{\min}(\eta\vec{\epsilon})=\eta E_{\min}(\vec{\epsilon})$. This follows because each of the folding energies is multiplied by the same constant factor, so the lowest-energy fold remains the same.

We first restrict ourselves to energy functions containing only μ_1 and μ_2. In this case, the function f depends only on μ_2. Since μ_2 is quadratic in uniform scalings of the ϵ_i, property (ii) above implies that f is proportional to $\sqrt{\mu_2}$. Thus

$$E_{\min}(\vec{\epsilon})=\mu_1(\vec{\epsilon})-a\sqrt{\mu_2(\vec{\epsilon})}, \tag{8}$$

where a is a positive dimensionless constant. The sign of the μ_2 term is negative because the minimum energy will always be lower than the average energy μ_1 of all the folds.

The form of the energy is somewhat unconventional, because of the square root term, but in terms of computational complexity the differences are practically negligible.

I will consider folding-enthalpy functions based on up to four-body interactions. In this case, f has as its arguments μ_2, μ_3, and μ_4; we can just as well express f in terms of μ_2, $\gamma_3=\mu_3/\mu_2^{3/2}$, and $\gamma_4=\mu_4/\mu_2^2$. Note that γ_3 and γ_4 are invariant under uniform scalings of the ϵ_i, while μ_2 scales quadratically. Then property (ii) above implies that $f/\sqrt{\mu_2}$ is invariant under such uniform scalings, and thus is a function g of only γ_3 and γ_4. Thus

$$E_{\min}(\vec{\epsilon})=\mu_1(\vec{\epsilon})-\sqrt{\mu_2(\vec{\epsilon})}g(\gamma_3,\gamma_4). \tag{9}$$

For simplicity, the linear form $g(\gamma_3,\gamma_4)=a+b\gamma_3+c\gamma_4$ is used here.

I now turn to the calculation of the coordination-number vector in terms of the residue hydrophobicity energies. The coordination-number vector for a given sequence ϵ_i contains the values Z_i for the minimum-energy fold. Thus

$$E_{\min}(\vec{\epsilon})=\sum_i \epsilon_i Z_i^{\min}(\vec{\epsilon}). \tag{10}$$

Then

$$\partial E_{\min}/\partial\epsilon_i=Z_i^{\min}(\vec{\epsilon})+\sum_j \epsilon_j\partial Z_j^{\min}/\partial\epsilon_i. \tag{11}$$

The second term in Eq. (11) vanishes for the exact $Z_i^{\min}$. This is because $Z_i^{\min}$ is a piecewise constant function of $\vec{\epsilon}$; since $Z_i^{\min}$ only takes on integer values, the space of possible values of $\vec{\epsilon}$ is divided into regions, each one corresponding to a certain value of $Z_i^{\min}$. Therefore the derivative $\partial Z_j^{\min}/\partial\epsilon_i$ vanishes except on a set of measure zero.

This implies that

$$Z_i^{\min}=\partial E_{\min}/\partial\epsilon_i, \tag{12}$$

except on a set of measure zero, and I will take this form form $Z_i^{\min}$ in the approximate calculations as well.

For the energy functions given in Eqs. (8) and (9), the corresponding forms of the coordination-number vectors, as obtained from Eq. (12), are

$$Z_i^{\min}=\langle Z_i\rangle-(a/\sqrt{\mu_2})\sum_j \langle\Delta Z_i\Delta Z_j\rangle\epsilon_j \tag{13}$$

and

$$\begin{aligned}Z_{\min}=&\langle Z_i\rangle-(a'/\sqrt{\mu_2})\sum_j \langle\Delta Z_i\Delta Z_j\rangle\epsilon_j\\&-(3b/\mu_2)\sum_{j,k}\langle\Delta Z_i\Delta Z_j\Delta Z_k\rangle\epsilon_j\epsilon_k\\&-(4c/\mu_2^{3/2})\sum_{jkl}\langle\Delta Z_i\Delta Z_j\Delta Z_k\Delta Z_l\rangle\epsilon_j\epsilon_k\epsilon_l,\end{aligned} \tag{14}$$

where $a'=a-2b\gamma_3-3c\gamma_4$. Thus the evaluation of $\vec{Z}$ at the μ_2 level involves a two-index tensor, and the μ_4 level involves a four-index tensor.

III. APPLICATION TO THE SQUARE LATTICE MODEL

I have evaluated the accuracy of the above functional forms in a model of compact folded conformations on a square lattice, via a comparison of the results from the approximate analytic forms (8) and (9) with exact results. The heteropolymers have 36 residues, and are constrained to have conformations contained inside a 6×6 square. These conformations (57 337 in number) can be enumerated exactly, so that for each sequence $\{\epsilon_i\}$ the lowest-energy fold can be found. Residues are considered to be neighbors if they are touching in the $(\pm1,0)$, $(0,\pm1)$, or $(\pm1,\pm1)$ directions; adjacent residues along the chain are not counted as neighbors. Thus each site potentially has eight neighbors minus its number of adjacent residues on the chain, which is 1 for the two residues at the ends, and 2 for the others. [The reason for counting $(\pm1,\pm1)$ pairs as neighbors is that otherwise artificial degeneracies arise in the model.]

The sequences $\{\vec{\epsilon}\}$ are chosen from a random distribution. In each sequence, the ϵ_i are chosen randomly on the interval $[-h,0]$, where h is the energy unit of the model. (Note that shifting of the average energy around which the ϵ_i are chosen would not affect the folding properties in the space of compact conformations. In this space, all folds have a total of 150 neighbors, so addition of a constant shift to the ϵ_i would not affect the relative energies of folds, but rather shift the energies of all the folds by the same constant.) Thus ϵ_i has a continuum of possible values rather than a finite number corresponding to the 20 different amino acids in a protein. The two-dimensional model is not sufficiently accurate to justify a literal association of ϵ_i with particular amino acids at this point. Using a large number of sequences, we have used least-squares fitting methods to adjust the parameters a, b, and c [cf. Eqs. (8), (9), (13), and (14)] to provide an optimal agreement between the analytic results and the exact ones.

Figure 1 shows a comparison between the exact folding energies and the analytic ones. A set of 5000 sequences was used to obtain the parameters, and the results are shown for a distinct test set consisting of 1000 sequences. Frame (a) shows results at the μ_2 level. Already at this level, the analytic estimate [Eq. (8)] is quite close to the exact results. The standard deviation is $0.86h$, roughly a percent of the typical values of the folding energies. It is illuminating to assess the accuracy from the point of view of the ordering energy, or the energy difference between the minimum-energy fold and the average energy of all possible compact folds (which is simply μ_1). This was termed the ‘‘chain reconfiguration energy’’ by Dill [1]. In our calculations, the average value of the ordering energy is $-10.8h$. Thus the ordering energy is obtained to about 8% accuracy. Upon going to the μ_4 level the improvement in the results is substantial. The standard deviation is $0.63h$, or less than a percent of the typical folding energies and about 6% of the ordering energy.

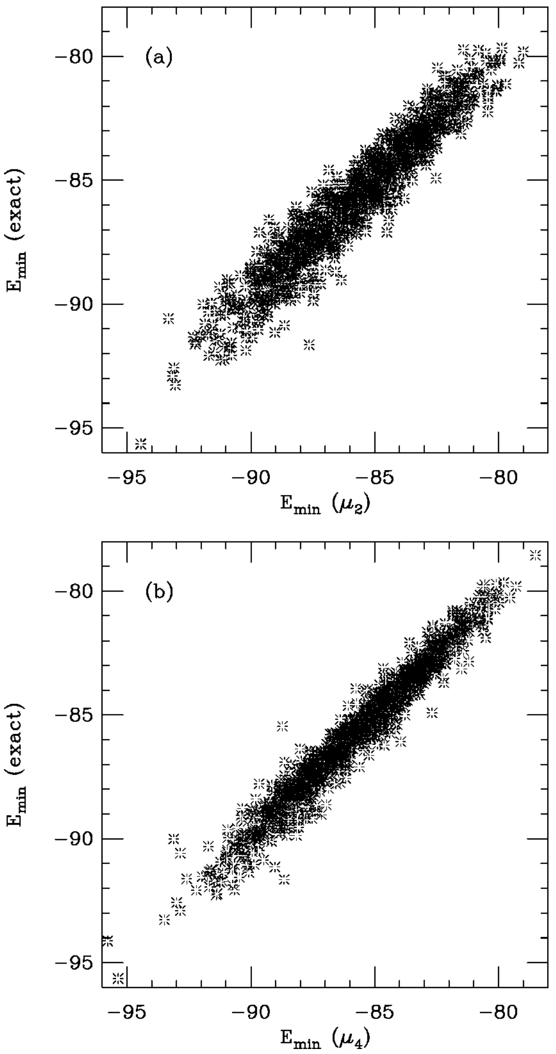

FIG. 1. Comparison between exact and model folding energies for the 6×6 lattice protein model. Energies are given in units of *h*, the energy scale of hydrophobicity energies. (a) μ_2-level results. (b) μ_4-level results.

For the coordination-number vector $\vec{Z}$, the rms error in the coordination number per residue is found to be 1.31 at the μ_2 level. At the μ_4 level, this drops to 1.19. For comparison, the possible coordination numbers range from 1 (for a corner residue in the middle of the chain) to 7 (for a residue inside the square, at the end of the chain). If one ignores the effects of sequence completely, and simply takes all of the sites to have the database-averaged coordination number, the rms error is 1.79. Figure 2 shows a comparison of the analytic estimate (14) of $\vec{Z}$ [frame (a)] at the μ_4 level with the exact result [frame (b)], for a sequence that is typical in that it has a rms error of 1.19. The analytic estimate obtains the gross features of the exact results, in particular the peaks centered on residues 6, 25, and 35, as well as the dips centered near residues 19 and 35. However, there are also important missing features in these results, in particular the low coordination numbers of residues 3 and 10. In general, the analytic results do not reach as high or as low as the exact results. Analysis of these results in view of the hydrophobicity sequence [frame (c)] illustrates the major effects included in the method. In the absence of frustration effects resulting from the connectedness of the residues, which make it im-

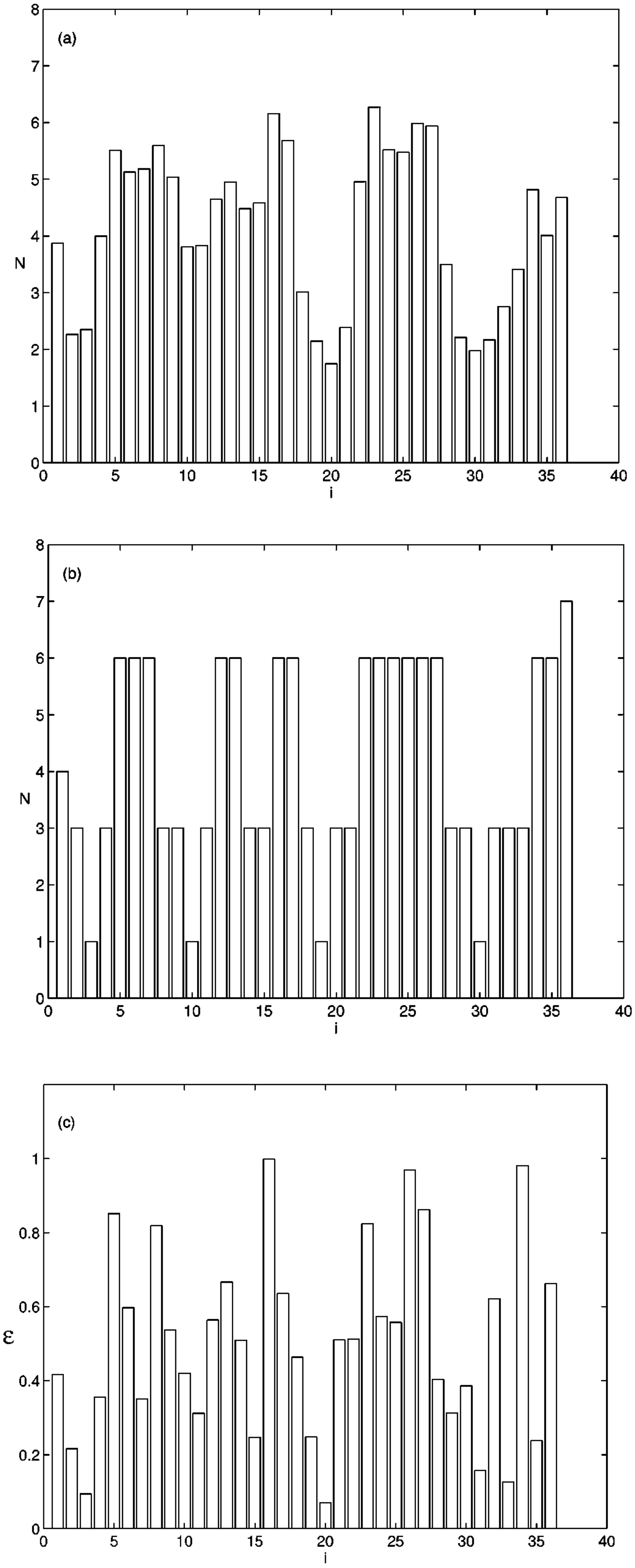

FIG. 2. Coordination-number vectors and hydrophobicity energies for 36-residue proteins in the 6×6 lattice model. (a) Coordination-number vector obtained at the μ_4 level, for a sequence with a rms error of 1.19, which is average for sequences tested. (b) Exact coordination-number vector for the same sequence. (c) Hydrophobicity energies ($-\epsilon_i$) for the above sequence.

possible for all of the residues to simultaneously find their optimal place in the folded configuration, one would expect the coordination number of a residue to be determined by its hydrophobicity: the larger the hydrophobicity, the larger the coordination number. Application of this simple theory to the sequence shown in Fig. 2(c) would yield a coordination-number vector with very frequent jumps. Both the analytic and exact results have a smoother behavior, with fewer jumps. Apparently the frustration effects reduce the frequency of jumps between different coordination numbers, and this effect is taken into account in the analytic estimates.

IV. CONCLUSION

The above results are encouraging, and it is in fact surprising that one can obtain the limit of the distribution of fold energies with such accuracy on the basis of only a few moments of this distribution. Ordinarily, one expects moment analysis to work well for averaged properties of a distribution, but not something as specific as its limits. However, more accuracy in the calculation of the energy, and in particular the coordination number, is certainly desirable. Methods based on higher moments are probably not feasible, as they scale unattractively with the length of the chain. However, it is possible that one could get useful information out of calculating certain parts of all the higher moments. This procedure would be analogous to the well-established use of diagrammatic resummation techniques [11] in many-body perturbation theory.

Unfortunately, we do not have a systematic collection of measured or calculated folding energies available for proteins or other types of heteropolymers. In addition, the present model leaves out many physical effects present in real proteins. Nevertheless, it does treat the competition between the preference of individual residues for surface versus interior positions on the one hand, and the constraints of the chain topology on the other hand. This competition has been a major stumbling block in predicting coordination-number vectors of proteins. Therefore, I feel that the functional form developed here may be of use in the prediction of coordination-number sequences in real proteins. One needs a hydrophobicity energy scale, as well as the statistical functions $\langle Z_i \rangle$, $\langle \Delta Z_i \rangle$, etc., entering the formalism. Considerable work has gone into establishing hydrophobicity scales for amino acids. They can be derived [7] from statistically obtained residue-residue potentials [8]. Since this approach provides energies directly, it would probably be the best for models of the present type. One can also obtain hydrophobicity scales via experiments transferring amino acids from aqueous to hydrophobic solvents [12]; in this case one needs a way to convert these hydrophobicities to energies. The statistical functions can in principle be evaluated from the large collection of known protein structures. It is not immediately clear if this collection is large enough to evaluate the correlation functions with the accuracy needed. In addition, the existing database contains proteins with many different lengths, so one would likely have to find a way to merge these data into a useful form. These efforts should await a more accurate method of predicting the coordination-number vector.

It is of interest to compare the present approach based on the coordination-number vector to recent work based on the ''contact map,'' a residue-residue matrix having elements 1 or zero according to whether two residues are in contact or not. Since the contact map contains more information than the coordination-number vector, one would expect folding on the basis of a known contact map to be, if anything, simpler than folding on the basis of the coordination-number vector; several works have demonstrated practical algorithms for contact-map based folding [13–15]. On the other hand, the prediction of the coordination-number sequence from an amino-acid sequence may be simpler than the prediction of the contact map, since there are fewer numbers to predict. This becomes important when dealing with fits to a fixed database, such as the Protein Data Bank. It is not clear at present which is the preferred method. One should also note that recent work [16] has shown that real proteins cannot be folded accurately with a pair energy function based on the contact map. The same must then hold for energy functions based on the coordination-number vector. Clearly, additional physical effects must be added to reliably fold real proteins. For example, rather than defining residue-residue contact in terms of a sharp threshold distance, it may be preferable to allow the extent of contact to go smoothly to zero with increasing distance; in this case the contact maps and coordination-number vectors would take on real rather than integer values. It is also likely that specific terms for electrostatics and hydrogen bonding must be included to obtain quantitatively accurate structures.

ACKNOWLEDGMENTS

This work received support from the U.S. Department of Energy under Grant No. DE-FG02-84ER45130, which is gratefully acknowledged. I appreciate stimulating conversations with Garland Marshall and Stan Galaktionov.

[1] K. A. Dill, Biochemistry **24**, 1501 (1985).

[2] D. K. Klimov and D. Thirumalai, Phys. Rev. Lett. **76**, 4070 (1996).

[3] A. Sali, E. Shakhnovich, and M. Karplus, Nature (London) **369**, 248 (1994).

[4] S. G. Galaktionov and G. R. Marshall, Biotechnology Computing **5**, 326 (1994).

[5] R. Unger and J. Moult, J. Mol. Biol. **259**, 988 (1996).

[6] A. E. Carlsson, in *Solid State Physics: Advances in Research and Applications*, edited by H. Ehrenreich and D. Turnbull (Academic Press, New York, 1990), Vol. 43, p. 1.

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

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

[9] T. Garel, L. Leibler, and H. Orland, J. Phys. II **4**, 2139 (1994).

[10] S. P. Obukhov, J. Phys. A **19**, 3655 (1986).

[11] A. L. Fetter and J. D. Walecka, *Quantum Theory of Many-Particle Systems* (McGraw-Hill, New York, 1971).

[12] A. L. Lehninger, D. L. Nelson, and M. M. Cox, *Principles of Biochemistry* (Worth Publishers, New York, 1993).

[13] M. Vendruscolo, E. Kussel, and E. Domany, Folding Des. **2**, 295 (1997).

[14] A. T. Brunger, P. D. Adams, and L. M. Rice, Curr. Opin. Struct. Biol. **5**, 325 (1997).

[15] L. Mirny and E. Domany, Proteins **26**, 391 (1996).

[16] M. Vendruscolo, R. Najmanovich, and E. Domany, Phys. Rev. Lett. **82**, 656 (1999).