A Deterministic Approach to the Protein Design Problem

Maksim Skorobogatiy, Hong Guo,* and Martin J. Zuckermann

Centre for the Physics of Materials and Department of Physics, McGill University, Rutherford Building, 3600, rue Université, Montréal, Québec, Canada H3A 2T8

Received October 22, 1996; Revised Manuscript Received March 10, 1997[⊗]

ABSTRACT: We have considered the problem of protein design based on a model where the contact energy between amino acid residues is fitted phenomenologically using the Miyazawa−Jernigan matrix. Due to the simple form of the contact energy function, an analytical prescription is found which allows us to design energetically stable sequences for fixed amino acid residue compositions and target structures. The theoretically obtained sequences are compared with real proteins, and good correspondence is obtained. Finally, we discuss the effect of discrepancies in the procedure used to fit the contact energy on our theoretical predictions.

I. Introduction

It is well-known that natural proteins fold into their native structures remarkably easily in spite of the enormous number of possible physical configurations.[1] For small proteins the native structure can be determined by the global minimum of the free energy.[2] It has been conjectured[3−5] that protein sequences are "optimized" such that not only is there a stable unique structure for the ground state but also the free energy landscape is funnel-like, which leads to efficient folding kinetics. A principle of minimal frustration was proposed[6] to enforce a selection of the interactions between monomers such that as few energetic conflicts occur as possible. Among other things, considerable theoretical effort has concentrated on finding proper models for protein folding and investigating various sequencings which lead to fast folding kinetics. In this regard, statistical analysis has played a very useful role in identifying the most relevant factors which determine the process of protein folding.

A statistical mechanical treatment of the protein folding problem requires a tractable form for the interactions between the various amino acid residues. One approach is to determine the contact interactions between each pair of amino acid residues using experimental data. Since there are 20 different amino acid residues, a total of 210 such interaction parameters is required for a complete description. This is the approach of Miyazawa and Jernigan.[7] Another simpler approach used by many researchers is to replace the detailed interaction between amino acid residues by a minimal two-parameter model where one parameter represents the attractive interactions between nonpolar groups and the second represents the interactions between the polar and nonpolar groups. Clearly, the simple models based on the second approach are easier to analyze and they have been successfully used for qualitative studies of protein folding, but they are still far from reality. However, even though more involved models give reasonably good agreement with experiment, they are unfortunately difficult to analyze theoretically. From this point of view, it is useful to introduce a compromise between minimal and more complete models, such that the resulting model is detailed enough to capture most of the essential characteristics of protein folding while at the same time being sufficiently simple for a tractable analytic approach. Indeed, it was shown by Grossberg *et al.* that, when the properties of real proteins are studied using an energy interaction matrix, sufficiently stable ground states can still be obtained even if there are some errors in the numerical values used for this interaction matrix.[8]

In a recent article,[9] Li, Tang, and Wingreen (LTW) suggested a particularly interesting parameterization of a statistical potential which was originally derived from known protein structures. By analyzing the Miyazawa−Jernigan (MJ) interaction matrix,[7] they found that the entire 20 × 20 MJ matrix can be fitted very well by a simple form,

$$E^{\theta\sigma} = q^{\theta} + q^{\sigma} + \beta q^{\theta} q^{\sigma} \tag{1}$$

where $E^{\theta\sigma}$ is the contact energy between amino acid residues of type $\theta \in (1, .., 20)$ and type σ and q^{θ} is a negative real number which is assigned to amino acid residues of type θ. In their fit to the MJ matrix, Li *et al.*[9] found numerical values lying in the range [−3.0, 0.0] for the quantities $\{q^{\theta}\}$. The form for the MJ matrix given by eq 1 thus has the following physical interpretation. The $(q^{\theta} + q^{\sigma})$ term corresponds to solvent exclusion, which is responsible for the formation of the hydrophilic surface and the hydrophobic core of the folded protein, whereas the $\beta q^{\theta} q^{\sigma}$ term represents segregation, which is responsible for the differentiation of secondary structures inside of the hydrophobic core. This fitting form, while not necessarily unique, reveals the intrinsic regularity of the interactions between the various amino acid residues and reduces the total of 210 interaction parameters to essentially 20. Hence this is clearly a useful formal step in the theoretical analysis of protein folding. It is also interesting to notice that the particular form of the contact energy $E^{\theta\sigma}$, being a combination of linear and quadratic terms, has also been discussed in previous protein folding literature.[10−12]

In this work we use this form of contact interaction with the fitting parameters given by the work of Li, Tang, and Wingreen to examine the following questions *analytically*. For a given protein compact target structure and a given amino acid composition, how can we find a sequence of the parameters $\{q^{\theta}\}$ that minimizes the total energy? Once a sequence that gives the minimum energy for the target structure is predicted, how does this "optimal" sequence compare with the protein sequence in the native state? How sensitive are our predictions to the fitted form given by (1)? Obvi-

[⊗] Abstract published in *Advance ACS Abstracts,* May 1, 1997.

S0024-9297(96)01564-1 CCC: $14.00

ously, these are important questions related to the protein folding problem. The motivation to investigate these questions analytically comes from the inspiring work of Shakhnovich and Gutin[13] who devised a numerical approach for the design of stable proteins by randomly permuting the amino acid residues for a given target structure using a Monte Carlo algorithm. We use the considerable intuition obtained from various important numerical calculations[13−15] to serve as a guide for our *analytical* examination of the above questions.

Because we are mainly concerned with the energetics of protein design, the sequence will be specified only by the parameters $\{q^\theta\}$. Hence a given composition is equivalent to a particular set of values for these parameters. We next fix a target structure for the protein. This is basically a three-dimensional space curve along which nodes labeled by $i \in (1, .., L)$ are located at equal distances from each other. Here L is the total number of monomers (amino acid residues) in the protein chain. A "model protein" is obtained by placing amino acid residues on these nodes. Since each amino acid residue corresponds to a specific value of the parameter q^θ, a sequence of these parameters must be obtained such that the total energy is a minimum for the given target structure. This sequence can be achieved by permuting the amino acid residues on the nodes until the energy reaches a minimum. However, such an exhaustive search quickly becomes intractable in the limit of long proteins. Numerically, one can improve the search by using important sampling techniques such as the Monte Carlo methods.[13] Once the minimum energy configuration is found, we will have obtained, at least theoretically, a "model protein" which is stable energetically. Clearly, if *nature* produces proteins only according to energy minimization, our "model protein" obtained in this manner should be very similar to the native state of the actual protein. We will thus compare our predicted amino acid sequence for the given target structure with the native state of the corresponding protein as given in the Protein Data Banks (PDB). Clearly, some differences should be expected since proteins also possess functional properties and they can not be considered as purely energetic units.

In our analytical work, we use the expression of eq 1 as a model (referred in the following as the LTW model) for the interaction matrix between monomers. It is then reasonably straightforward to make some general statements concerning the above questions while maintaining a good correspondence with the behavior of real proteins. In particular, we derive an expression for sequencing the parameters $\{q^\theta\}$ for a given target structure such that the total energy of the protein is minimized. Our expression is exact if the segregation term is neglected. We also show that the segregation can be included in an extremely good approximation, which we confirm by comparing results of our analytical predictions to those from exact numerical exhaustive search. Finally, we confirm the results of our calculation by numerically calculating the overlap between predicted sequences and the native protein sequences using 84 randomly chosen proteins from the Brookhaven Protein Data Bank with lengths ranging from $L = 21$ to $L = 680$.

The article is organized as follows. In section II we first derive the relevant formula for sequencing without the segregation term and we next treat the segregation as a perturbation. In section III we present our numerical tests on the perturbation treatment of the segregation term. The comparison of our predictions to those from PDB will then be presented. Section IV includes an estimate of the range of validity of our predictions in relation to the possibility of discrepancies in the fitted form of the MJ matrix, as given by eq 1. A summary of the main results is included in the last section.

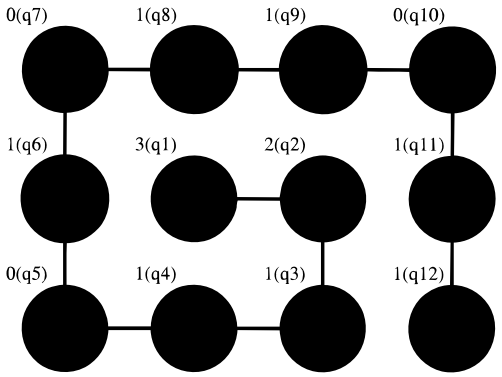

Figure 1. Sketch of a typical target structure on a 2D lattice with 12 nodes. The notation $n(q_m)$ on each node states that the node associated with parameter q_m has n nearest neighbors.

II. Protein Design Using the LTW Model

In the following we use the LTW model of eq 1 to "design" a stable sequence for a "model protein" with respect to a given amino acid composition and a given target structure. It is worth noting that the calculations presented below do not require the presence of a lattice, although they can also be applied to lattice models. For our problem, while we should denote q_i^θ as the parameter q^θ on node i of the target structure, without causing confusion from now on we shall simplify notation by dropping the Greek superscripts. Thus the contact interaction between monomer i and monomer j is written as

$$E_{ij} = q_i + q_j + \beta q_i q_j \tag{2}$$

with the understanding that q_i is given by one of the 20 possible values. As mentioned in the original work of ref 9, it follows from the values of the fitted q parameters that the solvent exclusion term $(q_i + q_j)$ gives the main contribution to the MJ energies E_{ij}. Hence it is reasonable to consider first the interaction $E_{ij} = q_i + q_j$ only and then investigate the influence of the segregation term $\beta q_i q_j$. This will be our approach.

A. Solvent Exclusion Term. As stated above, we first examine the LTW model for the case when the interaction matrix between monomers is given by the solvent exclusion term $E_{ij} = q_i + q_j$ only. Here we will assume that two monomers are in contact if the distance between them is smaller than a length scale of the order of a few angstroms. Following Li *et al.*,[9] we take this scale to be 6.5 Å. Then, if n_i denotes the number of closest neighbors to the ith node on our target structure, the total energy of the protein structure is given by

$$E = \sum_{i,j} E_{ij} = \sum_{i,j} q_i + q_j = \sum_i n_i q_i \tag{3}$$

and $\sum_i n_i = 2N$ and N is the total number of contacts.

As an example, we apply eq 3 to the target structure with 12 nodes on a 2D lattice shown in Figure 1. By placing 12 amino acid residues (monomers) with pa-

rameters q_1, q_2, ..., q_{12} on these nodes, we obtain a "model protein". For this structure there are 6 pairs of contacts: the monomer on the first node with "amino acid residue" q_1 has three contacts, the monomer on the second node with q_2 has two contacts, while all others have either one or no contacts. Using eq 3 the energy is given by $E = 3q_1 + 2q_2 + q_3 + q_4 + q_6 + q_8 + q_9 + q_{11} + q_{12}$. This example shows that it is natural to specify the target structure by a vector with elements representing the number of closest contacts to each node. Hence for a particular structure with L nodes on a 2D lattice, the geometric conformation is represented by the "contact vector" $\vec{n} \equiv \{n_i\}$ where $n_i \subset \{0, 1, 2\}$ if $i \subset \{2, ..., L - 1\}$; and $n_i \subset \{0, 1, 2, 3\}$ if $i \subset \{1, L\}$. In this notation the ith component of $\vec{n}$ gives the number of closest neighbors to the ith node. A similar prescription can easily be written down for 3D systems. It is clear that the *length* of a vector $\vec{n}$ will in general increase as the number of contacts in a given structure increases.

Next we introduce a second vector with L components, $\vec{q} \equiv \{q_i\}$, which specifies a particular sequence of the values of $\{q_i\}$ imposed on the geometrical structure defined by $\vec{n}$. The placing of amino acid residues on the nodes of the target structure is equivalent to assigning the corresponding values of q_i to each node. Then using eq 3 the energy of the "model protein" can be rewritten as

$$E = \sum_i n_i q_i = \vec{n} \cdot \vec{q} \tag{4}$$

Equation 4 shows that the energy is separable in the geometrical factors and the details of a particular sequence in this model. Using our notation, if one draws all possible vectors of type $\vec{n}$ corresponding to all different geometrical structures of a protein, the configurational space of the protein will be represented by a vector bundle generated by the set of all $\vec{n}$ vectors, while a particular sequence will be represented by a single vector $\vec{q}$. Then, as seen from eq 4, the energy spectrum for a particular set of amino acid residues on a given target structure will be determined by the projection of the vector bundle onto the vector $\vec{q}$. As mentioned above, more compact conformations will in general have longer $\vec{n}$ vectors since compact structures tend to have more contacts between monomers and the corresponding length will be proportional to $L^{1/2}$. Given that the most compact conformation is represented by the longest vector, $\vec{n}_{max}$, all less compact conformations will lie inside a sphere of radius equal to the magnitude of $\vec{n}_{max}$. Nearly compact conformations will then lie in the neighborhood of this sphere.[16] This is shown in Figure 2.

We now consider a specific sequence defined by a vector $\vec{q}$ with magnitude Q. We next denote a given compact geometrical target structure for a given protein as $\vec{n}_\alpha$. If we are *not* limited by a particular composition represented by a fixed set of values of the parameters $\{q_i\}$, the energy minimization and design is straightforward. Equation 4 shows directly that the system energy is minimized if we choose $\vec{q}$ to be antiparallel to the vector $\vec{n}_\alpha$, as shown in Figure 2. For this trivial case we thus obtain the "ideal" sequence, $\vec{q}_{ideal}$, which gives the lowest possible energy for the target structure represented by $\vec{n}_\alpha$

$$\vec{q}_{ideal} = \vec{n}_\alpha \left(\frac{Q^2}{\vec{n}_\alpha \cdot \vec{n}_\alpha} \right)^{1/2} \tag{5}$$

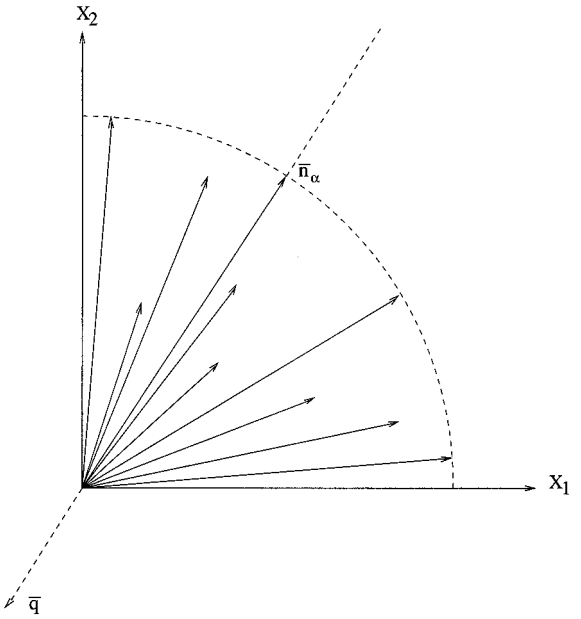

Figure 2. Configurational space of the protein as represented by a vector bundle generated from all vectors of type $\vec{n}$. A particular sequence is represented by a vector $\vec{q}$.

However, such an "ideal model protein" is not realistic, as it does not respect the particular values of the parameters $\{q_i\}$ corresponding to the actual set of amino acid residues defining the primary structure of the protein. Hence the set of values of the parameters $\{q_i\}$ must be fixed in order to specify a particular protein. The problem then becomes more complicated as the total energy must now be minimized subject to this constraint. Furthermore, we can only minimize the energy by shuffling the given parameter set of the parameters $\{q_i\}$ among the different nodes of the target structure. This corresponds to performing discrete rotations of the vector $\vec{q}$ in its specific vector space rather than the continuous rotations used to find the "ideal" sequence. However, even under this constraint we can still solve the problem close to analytically.

We begin by stating the following well-known inequality. If

$$n_1 \geq n_2 \geq n_3 \ldots \geq n_L$$

$$q_1 \geq q_2 \geq q_3 \ldots \geq q_L$$

where n_i and q_i are arbitrary real numbers, then

$$\sum_{i=1}^{L} n_i q_i \geq \sum_{i=1}^{L} n_i q_{k_i} \tag{6}$$

where q_{k_i} represents any permutation of the set of parameters q_i. We now fix the amino acid composition by fixing the values of the components, q_i, of a given vector $\vec{q}$, where the amino acid residue corresponding to q_i is placed on the ith node of the target structure. We change the amino acid sequence on the target structure by permuting the values q_i among the nodes of the target structure represented by the vector $\vec{n}_\alpha$. This gives us a new vector representing a different amino acid sequence on the target structure. Now the minimization of the energy with respect to the target structure given by $E = \vec{n}_\alpha \cdot \vec{q}$ clearly requires us to find a vector, $\vec{q}_{min}$, which is as antiparallel to $\vec{n}_\alpha$ as possible. Since the fitted numerical values for the parameters q_i

are all negative, we devise the following procedure on the basis of the inequality of eq 6. In this procedure we minimize the energy by first sorting all the components q_i according to their absolute values and then sorting all nodes on the geometrical target structure according their number of nearest neighbors. We then place the amino acid residues corresponding to larger values of $|q_i|$ on the nodes of the target structure which have larger numbers of contacts, and the amino acid residues corresponding to smaller values of $|q_i|$ on the nodes with a smaller number of contacts. This is a systematic way of finding the sequence which gives a stable (minimum energy) protein target structure. Since no exhaustive search is involved, essentially no computer is needed. However, from the energy point of view, this procedure, because it is based on eq 3, will produce sequences of $\{q_i\}$ in which as many hydrophilic amino acid residues as possible are on the surface of the target structure and as many hydrophobic amino acid residues as possible are in its interior. This will be corrected in the next section when the segregation term is included.

The method discussed here may lead to degeneracies of the final sequence $\{q_i\}$; *i.e.* there may be more than one sequence which gives the same minimum energy for the target structure $\vec{n}_\alpha$. This is easy to understand by noticing that a compact structure predominantly consists of "surface" monomers with a small number of nearest neighbors and "interior" monomers with a large number of such neighbors. Thus, the geometrical permutation of "interior" monomers among themselves or "surface" monomers among themselves made by permuting the relevant parameters, q_i, will not alone change the energy. The sequence obtained from the above procedure hence specifies the structure up to a differentiation of "surface" and "interior" monomers only.

B. Segregation Term. In order to treat the segregation term $\beta q_i q_j$ in our analysis, we use the result of ref 9 that it is small in comparison with the solvent exclusion term studied in the last subsection and that it should not alter the relationship between "surface" and "interior" monomers. Thus the inclusion of this term should not cause a substantial change in the geometric vector, $\vec{n}$, of our model protein. However, this term will at least partially break the degeneracy in the secondary structure. Because the segregation term is quadratic in $\vec{q}$, it energetically favors segregation between different amino acid residues and leads to an increased specificity of the overall structure. Including this term, we now investigate which sequence should be chosen so that the given target structure will have minimal energy.

As in the previous section, we first find the ideal sequence $\vec{q}_{\text{ideal}}$ by only fixing its length to be Q and then minimizing the energy of the target structure denoted by vector $\vec{n}_\alpha$. Mathematically, the problem is to minimize the function $\sum_{ij}E_{ij} = \sum_{ij}q_i + q_j + \beta q_i q_j$, $\beta < 0$, $q_i < 0$, while keeping the length of $\vec{q}$ fixed at the value Q. For this purpose, we introduce the contact matrix, C^α, which has elements $C^\alpha_{ij} = 0$ if ith monomer does not have jth monomer as a nearest neighbor and $C^\alpha_{ij} = 1$ otherwise. We can then rewrite the energy of the target structure with a sequence $\vec{q}$ as follows

$$\begin{cases} E^{(\alpha)} = \vec{n}_\alpha\vec{q} + \dfrac{\beta}{2}\vec{q}C^\alpha\vec{q} \\ \vec{q}\cdot\vec{q} = Q^2 \end{cases} \tag{7}$$

Using the method of Lagrange multipliers one finds the following sequence which gives the minimum energy,

$$\vec{q}_{\text{ideal}} = [\lambda I - \beta C^\alpha]^{-1}\cdot\vec{n}_\alpha \tag{8}$$

where I is the identity matrix and λ is the solution of equation

$$Q^2 = \vec{n}_\alpha\cdot[\lambda I - \beta C^\alpha]^{-2}\cdot\vec{n}_\alpha \tag{9}$$

These equations can easily be solved, and they can be replaced by their expansions in β for proteins with small numbers of monomers.

The sequence $\vec{q}_{\text{ideal}}$ as solved above gives the lowest possible energy for a given target structure. However, it does not respect the actual amino acid composition of the real protein. We should instead fix the composition and only shuffle the elements, q_i, of the vector $\vec{q}$ instead of changing their values. In the previous subsection, we gave a prescription for finding a sequence *exactly* for a given composition. However, with the inclusion of the segregation term we can no longer access an exact solution, but we can make an extremely good approximation for the desired sequence. The approximation we use here has the character of a mean field analysis in that we replace one of the vectors, $\vec{q}$, in the quadratic term of eq 7 by $\vec{q}_{\text{mf}}$, *i.e.*

$$E^{(\alpha)} \approx \left(\vec{n}_\alpha + \frac{\beta}{2}\vec{q}_{\text{mf}}C^\alpha\right)\cdot\vec{q} \equiv \vec{n}^{\text{mf}}_\alpha\cdot\vec{q} \tag{10}$$

Here $\vec{n}^{\text{mf}}_\alpha$ is the sum inside the bracket and results in a slight change of the elements of the nearest neighbor vector. The choice of $\vec{q}_{\text{mf}}$ may be $\vec{q}_{\text{ideal}}$ using eq 8, which we obtained in this section, or the *optimized* sequence, which we obtained without the segregation term. With this new form of energy, we can use the same sorting prescription discussed at the end of the last subsection to find the minimal sequence $\vec{q}$ under the constraint of fixed composition. In the next section we shall confirm this mean field like analysis by comparing results to those from the exhaustive numerical search.

III. Numerical Results

To confirm the predictions by our analytical design method discussed in the last section, in the following we shall examine the quality of our mean field like analysis of the segregation term and compare our results to those from real protein structures using 84 randomly chosen proteins from PDB.

A. Lattice Enumeration. While our method of finding an optimal sequence without segregation term was exact, the validity of the mean field approach to include the segregation, eq 10, needs to be investigated. To this purpose, we have used the compact structure of Figure 1 as the target structure $\vec{n}_\alpha$, and fixing a composition of the parameters $q_i \in(-4.0, 0.0)$, we designed the optimal sequence $\vec{q}_{\text{opt}}$ using our analytical method including the segregation term. With a choice of $\vec{q}_{\text{mf}}$ (see below), this procedure minimized the energy $E^{(\alpha)} \approx \vec{n}^{\text{mf}}_\alpha\cdot\vec{q}_{\text{opt}}$ within the mean field approximation to give $\vec{q}_{\text{opt}}$ by eq 10. We then exhaustively generated all other possible structures of this 12 monomer self-avoiding chain on a 2D lattice, and we denote these structures by $\vec{n}_\eta$ where $\eta \neq \alpha$. The energies, $E^{(\eta)}$, of these other structures are calculated using eq 7. We then compare $E^{(\alpha)}$ with the smallest $E^{(\eta)}$ to see which is lower. Finally, we have checked 60 different amino acid

compositions with q_i's randomly generated from the above range. We have fixed the parameter $\beta = -0.476$ in this numerical check, and several observations are in order.

First, for 57 out of the 60 random compositions tested, $E^{(\alpha)} < E^{(\eta)}$. Hence our analytical design with the mean field approximation indeed generated the sequence which guarantees the compact target structure to be the ground state. For the other 3 compositions of the q_i's, each case has only one chain structure which gave a slightly lower energy than $E^{(\alpha)}$. However, the difference is very small, being 0.3%, 0.6%, and 2.8%. We may thus conclude that the accuracy of our mean field treatment of the segregation term is acceptable. Second, we found no difference to this conclusion using two different $\vec{q}_{\mathrm{mf}}$ in eq 10. The two most evident choices of $\vec{q}_{\mathrm{mf}}$ are the optimized sequence which we obtained by neglecting the segregation term (see section II.A), or the $\vec{q}_{\mathrm{ideal}}$ of eq 8 where we included the segregation term. They work equally well. Finally, our numerical test also gave a measure of the relative importance of the segregation term. If we directly use the optimized sequence determined without the segregation term as $\vec{q}_{\mathrm{opt}}$ to compute the energy using eq 7, rather than determine $\vec{q}_{\mathrm{opt}}$ as we have done so far, the comparison with the lattice enumeration is worse. In this case 27 out of the 60 random compositions gave $E^{(\alpha)} < E^{(\eta)}$. On the other hand, the other 33 compositions produced a chain structure with lower energy than the target structure $\vec{n}_\alpha$, with differences which are less than 10%. This means that choosing a sequence which is only optimal without the segregation term, we have a lower probability of making a target structure to be the ground state, although the result is still not far from it.

Our mean field approach can also be applied to the model where the contact energy is determined solely by the segregation term: $E_{i,j} = \beta q_i q_j$. In terms of the total energy, we have $E^{(\alpha)} = (\beta/2)\vec{q}C^{\alpha}\vec{q}$. With the approach discussed above, we again tested 60 randomly generated sequence compositions by comparing our analytically designed results to the exact enumerations using the 12 monomer chain of Figure 1 as the target. Of these, in 49 cases the designed sequences made the energy of the target structure to be the ground state. In the remaining 11 cases, the designed sequences produced higher energies than some structures other than the target, but the differences were less than 10% with only one case of about 28%. Hence, for this purely quadratic model, our design method also works quite well.

With the above comparisons, we may conclude that the mean field approach to the segregation term is an acceptable approximation for finding an optimal sequence for general models described by contact energies in the form $E_{i,j} = C_0 + C_1(q_i + q_j) + C_2(q_i q_j)$.

B. Comparison to PDB. In the previous sections we have derived analytical expressions and devised exact or approximate methods to find sequences with fixed composition which minimize the total energy of a given target structure. Using this method, we showed how "model proteins" can be generated. The next step is to see how closely these "model proteins" resemble the native states of the real proteins with the same composition. This is important since natural proteins have other functional properties and do not just minimize their energy during the folding process. In addition, our analysis so far is based on the model described by eq 1, which is a result of fitting to experimental data.[9] Hence, some numerical discrepancies in the fitted parameters, $\{q_i\}$, can be expected. These will give rise to differences between our model proteins and their real counterparts.

In order to check the accuracy of our predicted sequences with fixed composition and fixed target structure, we randomly chose 84 proteins with lengths between $L = 21$ and $L = 680$ from the Brookhaven Protein Data Bank as our target structures. We examine the quality of the predictions by using two parameters that give a measure of the overlap of our "model proteins" with real protein structures. The first parameter uses a scale that only distiguishes between hydrophobic and hydrophilic amino acid residues

$$\mathrm{PH}_0 \equiv 1 - \frac{1}{L}\sum_i |\alpha_i - \alpha_i^{\mathrm{real}}| \tag{11}$$

Here α_i and α_i^{real} equal 1 if the ith amino acid residue is hydrophobic and 0 otherwise. In the following we consider an amino acid residue to be hydrophobic[9] if its strength $q_i \le -1.5$. Clearly, PH_0 approaches unity if the predicted values of α_i are close to those of the real protein, α_i^{real}. The second parameter we use is defined as

$$S_0 \equiv 1 - \left(\frac{\sum_i |q_i - q_i^{\mathrm{real}}|^2}{\sum_i q_i^{\mathrm{real}\ 2}}\right)^{1/2} \tag{12}$$

where $\{q_i\}$ is the predicted sequence while $\{q_i^{\mathrm{real}}\}$ is the real sequence. This quantity is a more refined measure than PH_0 since it uses the complete 20 letter code instead of the two letter code used to define PH_0. Throughout the calculations we have used the values of $\{q_i\}$ as fitted in ref 9. Again, we consider two monomers in contact with each other if the distance between them is less than 6.5 Å.

Before proceeding any further, we present an expression for PH_0 for a *random* sequence of the same composition as a real protein. This expression can easily be obtained from eq 11. If n_0 is the total number of hydrophilic amino acid residues and n_1 is the number of hydrophobic ones, we obtain

$$\mathrm{PH}_0 = \frac{n_1^{\ 2} + n_0^{\ 2}}{(n_1 + n_0)^2} \tag{13}$$

From this expression we conclude that PH_0 is usually greater than 0.5 for a random sequence and it equals to 0.5 when $n_1 = n_0$.

The first set of results from our calculations gives the correspondence between the exact solution of the model with the total energy given by eq 3, where the segregation term is neglected, and the real protein sequences. Using our method described in the last section, we minimized the energy of geometrical target structures taken from PDB while keeping the composition fixed and identical to that of real proteins. The results for 12 typical proteins are tabulated in Table 1. The first column gives codes for the 12 proteins randomly selected from PDB. The second column gives values for the two letter code measure, PH_0, as obtained by our minimization procedure. The data show that the "model protein" sequences thus obtained have a 61%−71% correspondence with real proteins for the two letter code measure. On the other hand the best random sequence (third column) gives only a 59% correspondence. Since the

Table 1. Values of the Overlap between Real Protein Structures and Those of the Predictions[a]

protein code	PH_0 minimized	PH_0 random sequence	PH_0 minimized degeneracy limits	S_0
621p	65%	54%	71% ≥ 56%	28%
1291	66%	53%	73% ≥ 59%	27%
4mbn	62%	57%	62%	23%
144l	64%	55%	72% ≥ 58%	26%
451c	61%	59%	61% ≥ 59%	20%
181l	64%	56%	70% ≥ 62%	26%
7api	64%	55%	67% ≥ 59%	25%
6dfr	70%	55%	80% ≥ 61%	33%
2pal	67%	58%	70% ≥ 64%	21%
2gch	63%	56%	67% ≥ 61%	28%
1edn	71%	53%	81% ≥ 71%	47%
1epg	66%	51%	70% ≥ 58%	37%

[a] Only the solvent exclusion term is included in the analysis. Columns 1, protein codes from PDB; 2, the measured PH_0 obtained from the analytical predictions; 3, PH_0 from random sequences; 4, the best and worst predicted PH_0 values among all the degenerate sequences; 5, the 20 letter code overlap S_0.

model sequences are degenerate, as discussed before, we computed all possible degenerate sequences and the corresponding PH_0 values.[17] The best and the worst correspondences with real proteins are listed in column four. From these numerical values we conclude that, when using the two letter code measure, PH_0, even the simple model with only the solvent exclusion term can lead to good correspondence between our predicted "model protein" and the real protein. On the other hand, if we use the more stringent measure S_0, which is based on the 20 letter code, the "model protein" and the real protein have considerably less overlap, as shown in the fifth column of Table 1. Here the best case is less than 50%.

The overlap is greatly improved if the segregation term is included. This term gives an overlap parameter S_0 of 0.87 between $\vec{q}_{\text{ideal}}(\beta = 0)$ and $\vec{q}_{\text{ideal}}(\beta = -0.467)$. This implies that the inclusion of the segregation term will give similar energy behaviors, and we therefore do not expect any significant changes when using the solvent exclusion term only. On the other hand, inclusion of the segregation term should improve the S_0 overlap parameter with real protein sequences because this term lifts the degeneracies in the position of amino acid residues. Using our analytical design procedure, we obtained the results listed in Table 2. In particular, in comparison with Table 1, the parameters PH_0 do not change very much even though the degeneracies discussed in ref 17 have largely been lifted. When the segregation term is included, the difference between the best and worst limits of Table 1 decreased to around 1% only due to the lifting of the degeneracies. On the other hand, column four of Table 2 shows clearly that the 20 letter code measure S_0 has been significantly improved by about a factor of 2 to values ranging from 0.36 to 0.53 when the segregation term is present.

The last three columns of Table 2 give the total energies of the 12 proteins. Column five gives the energies for the real protein structures as computed using the MJ matrix. Columns six and seven correspond to the optimized "model protein" sequence when the MJ matrix or equivalently the LTW model eq 1 is used to calculate the energies. This column shows that theoretical predictions using the MJ interaction matrix give energies lower than those of the real proteins by 20%−42% for almost all the proteins tested. The exception is the protein coded 1ed, for which the predicted energy is higher than that of the true native

Table 2. Values of the Overlap between Real Protein Structures and Those of the Predictions[a]

protein code	PH_0 minimized	PH_0 random	S_0	energy: real protein	energy: using MJ	energy: using LTW
621p	65%	54%	42%	−1123	−1414	−1326
129l	67%	53%	41%	−1112	−1378	−1622
4mbn	62%	57%	34%	−1131	−1430	−1322
144l	66%	55%	40%	−1101	−1353	−1648
451c	56%	59%	35%	−395	−563	−465
181l	64%	56%	39%	−1110	−1393	−1781
7api	63%	55%	36%	−2572	−3376	−3037
6dfr	66%	55%	42%	−993	−1184	−1174
2pal	67%	58%	35%	−563	−801	−529
2gch	63%	56%	41%	−1718	−2122	−2178
1edn	71%	53%	53%	−113	−107	−97
1epg	62%	51%	42%	−313	−382	−230

[a] Both the solvent exclusion and segregation terms are included in the analysis. Columns: 1, protein codes from PDB; 2, the measured PH_0 obtained from the analytical predictions; 3, PH_0 from random sequences; 4, the 20-letter code overlap S_0; 5, energy of the real protein sequence as computed using the MJ matrix; 6, energy of the predicted sequence computed from the MJ matrix; 7, energy of the predicted sequence computed from the LTW fitting form.

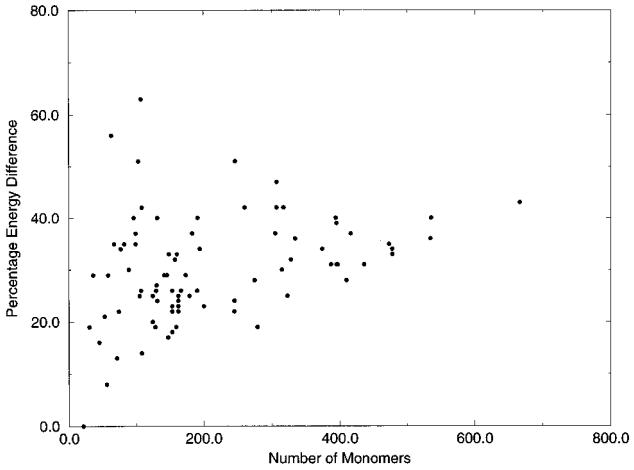

Figure 3. Percentage energy difference between real proteins and "model" proteins versus the number of monomers in a protein. A value of zero on the scale of the ordinate is equivalent to 100% correspondence.

state by 5%. Similar behavior was found using the LTW model, but in this case there are three proteins for which the theory prediction gives higher energies than for the native states. In this regard the difference between the two theory predictions reflects the quality of the fit of the MJ matrix as given in eq 1.

The complete data for the energy comparisons of 84 proteins with their "model" counterparts are presented in Figure 3. The horizontal axis corresponds to the number of monomers in a chosen protein. The vertical axis gives the percentage difference in energies between real proteins and the corresponding "model" or "designed" proteins. Several comments concerning this figure need to be made. First of all, the proteins examined for this figure were chosen randomly from the Brookhaven Protein Data Bank and we concentrated on those proteins consisting of one chain and representing autonomous units. An interesting feature of Figure 3 is the broad scatter of points for short proteins on an energy difference scale along the vertical axis. From general arguments it is clear that the energy optimization of a structure should be much easier in nature for shorter proteins than for longer proteins. This is why we expect a greater similarity between real and "model"

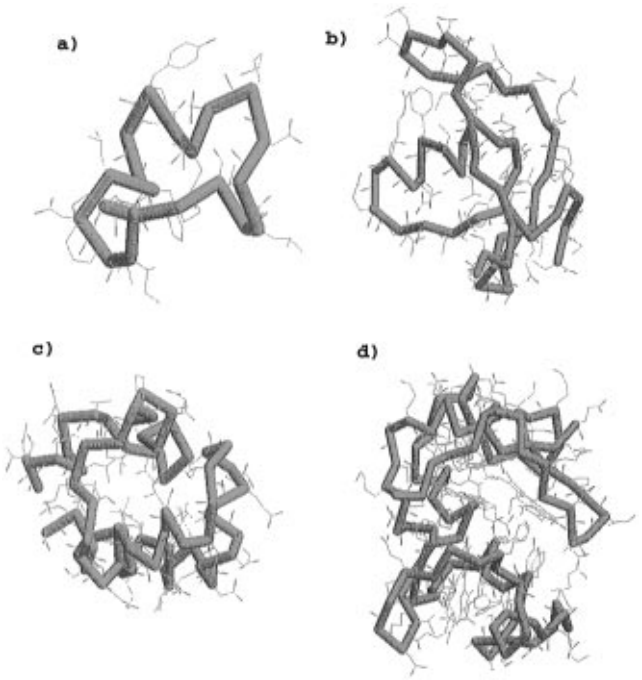

Figure 4. Native protein conformations taken from the PDB: (a) 1edn; (b) 1hpt; (c) 1r69; (d) 2cym.

proteins for low monomer number. This tendency is examined in Figure 3, which shows that not all of the short proteins chosen have good correspondence with the related "model" proteins. A possible explanation for the broad scatter is that the energy of a short protein is much more sensitive to structural details such as side chains, rigid bonds, etc. than the energy of a long protein. Such structural details are not considered in simple theories, and they clearly impose additional constraints on the geometrical and energy landscapes of a protein. For the number of interactions present in short proteins these structural details can be expected to give substantial deviations from the "model proteins" that were obtained by energy minimization alone. In contrast the large number of interactions present in long proteins can be expected to suppress the influence of structural details on their energies. This implies that there should be less scatter in the energy difference between real proteins and "model proteins" obtained from energy minimization using simple theories in the case of high monomer number.

In order to clarify these considerations we investigated the structure of two best case scenarios (1edn, 1hpt) and two worst case scenarios (1r69, 2cym) for short proteins. These structures are shown in Figure 4. We notice that while 1edn and 1hpt are very compact proteins, 1r69 and 2cym have some interior cavities related to their function (1r69 for example is an amino terminal domain). The presence of structural cavities is probably due to packing arrangements among the amino acid residues. This packing constraint plays the role of a "structural perturbation" from the "unperturbed" state defined by the minimal energy requirements of abstract point residues interacting via the MJ matrix as used in our simple model. In contrast, for the 1edn protein, the abundance of hydrophilic groups leads to a very compact, energetically minimized structure where there are no "structural perturbations".

Even though there can be other contributions to the energy differences between "model" and real proteins, the energy comparison with the true native structure of real proteins clearly suggests that while energy minimization plays an important role in protein folding, it is definitely not the only rule that nature follows.

IV. Discussion

We have shown that the LTW model and the related analytic procedure for the design of stable "model proteins" leads to reasonable results which compare well with real proteins for the two measures used.

The one point of concern in this discussion relates to the actual fit of the MJ matrix by Li *et al.*[9] which was based on eq 1 and which resulted in the specific values of the parameters $\{q_i\}$ used here in the analysis for protein design. However, we recognize that there are always some small uncertainties to any numerical fit; hence it is useful to determine to what extent these small uncertainties affect the predictions and conclusions of sections II and III. To this purpose, we notice that using quantities defined as $\delta q_i^j \equiv (\mathbf{MJ}(i,j) - E_{i,j})/(1 + \beta q_j)$ and then substitution of $q_i + \delta q_i^j$ into eq 1 give the correct values of the elements of the MJ energy matrix, $\{\mathbf{MJ}(i,j)\}$. This suggests that we may use $\delta q_i = (\sum_{j=1}^{20}|\delta q_i^j|)/20$ as a measure of the fitting quality of the parameter q_i. Obviously, the better the fit, the smaller the value of $\delta q_i/|qi|$. For the LTW fit, this quantity is not greater than 20%. Now let us assume that there are L monomers in a compact target structure and that the errors for each parameter q_i are independent. For a compact conformation there are $\sim dL^{(d-1)/d}$ "surface" monomers and there are $\sim L$ "interior" monomers. When the monomers are shuffled in order to find the most stable sequence, as discussed in section II, the possible difference in energy given by the LTW model will be of the order of $\Delta E \sim dL^{(d-1)/d}\langle q\rangle n_s$ where $\langle q\rangle$ is the average of the parameters $\{q_i\}$ and n_s is close to the average number of nearest neighbors for the "surface" monomers. In the fit of ref 9, $\langle q\rangle$ is of the order of unity. On the other hand, the errors in the energy due to the small discrepancies δq_i are of the order of $\delta E \sim \delta(\vec{n}\cdot\vec{q}) \sim L(\langle\delta q\rangle n_b/20^{1/2})$ since there are 20 independent parameters q_i. Here $\langle\delta q\rangle$ is the average of the quantities δq_i. Clearly, if $\delta E \sim \Delta E$, we cannot make any reasonable predictions. From this discussion we conclude that the use of the LTW model for our calculation is justified if

$$L \ll \left(20^{1/2}\frac{n_s}{n_b}d\frac{\langle q\rangle}{\langle\delta q\rangle}\right)^d \tag{14}$$

Hence, as the fitting uncertainty is at most 20%, *i.e.* $\langle q\rangle/\langle\delta q\rangle \sim 5$, our predictions should be valid for $L \ll 500$ in 2D and $L \ll 5000$ in 3D. Thus even for a 20% error in the parameters $\{q_i\}$, our procedure based on the LTW model can still describe and make predictions for relatively long proteins.

V. Summary

In this work we applied the model of Li, Tang, and Wingreen[9] to design "model proteins" that have minimum energy for a fixed amino acid composition and a given target structure. The model is well suited to this procedure because it is reasonably accurate and yet sufficiently simple for analytical or deterministic calculations. Using the vector notation of section II for the target structure and the sequence, we were able to find a simple method which determines model protein sequences based on the LTW model. We estimated that our method can be applied to protein chains with a few

hundred monomers even if there are substantial errors in the parameters $\{q_i\}$. Using 84 randomly chosen real proteins from protein data banks, we confirmed that our predicted sequences are reasonably realistic. Furthermore, our "model protein" sequences have total energies, as computed from the LTW model or from the original MJ matrix, that are for most cases lower than those of the real proteins. While several factors could be responsible for this difference, it suggests that energy minimization is indeed important but it is not the only factor that determines the native structure of proteins. We have also found that the segregation term in the LTW model plays the important role of lifting the degeneracies of the sequence that occur when only the solvent exclusion term is included in our calculations. In addition, this term improves the 20 letter code comparison between our theory and real proteins by a major factor. On the other hand for the two letter code overlap measure used in section III to compare with real proteins, our predictions for the proteins examined here are usually 10% (the best 19%) better than those obtained by using a random sequence. As pointed out in ref 13, one does not expect a 100% homology between the real and the predicted sequence for the design problem, as degeneracies and "structural perturbations" are present. The merit of our method lies in the fact that it is easy to use, allows analytical or partially analytical solutions of the problem, gives simple geometrical interpretation of the results, and uses essentially no computer time while giving reasonable comparisons with real proteins.

Acknowledgment. We gratefully acknowledge support by the Natural Sciences and Engineering Research Council of Canada and le Fonds pour la Formation de Chercheurs et l'Aide à la Recherche de la Province du Québec.

References and Notes

(1) See, for example, articles in: *Protein Folding*, Creighton, T. E., Ed.; W. H. Freeman and Co., New York, 1992.

(2) Anfinsen, C. *Science* **1973**, *181*, 223.

(3) Go, N. Abe, H. *Biopolymers* **1981**, *20*, 991.

(4) Shakhnovich, E. I.; Farztdinov, G.; Gutin, A. M.; Karplus, M. *Phys. Rev. Lett.* **1991**, *67*, 1665. Sali, A.; Shakhnovich, E. I.; Karplus, M. *J. Mol. Biol.* **1994**, *235*, 1614. Sali, A.; Shakhnovich, E. I.; Karplus, M. *Nature* **1994**, *359*, 248.

(5) Wolynes, P. G.; Onuchic, J. N.; Thirumalai, D. *Science* **1995**, *267*, 1619.

(6) Bryngelson, J. D.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 7524.

(7) Miyazawa, S.; Jernigan, R. L. *Macromolecules* **1985**, *18*, 534.

(8) Pande, V. S.; ; Grosberg, A. Yu.; Tanaka, T. *J. Chem. Phys.* **1995**, *103*, 1.

(9) Hao Li, Chao Tang and Ned Wingreen, cond-mat/9512111.

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

(11) Goldstein, R. A.; Luthey-Schulten, Z. A.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 4918, 9029.

(12) Rammanathan, S.; Shakhnovich, E. *Phys. Rev. E* **1994**, *50*, 1303.

(13) Shakhnovich, E. I.; Gutin, A. M. *Protein Eng.* **1993**, *6*, 793.

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

(15) Honeycutt, J. D.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 3526; *Biopolymers* **1992**, *32*, 695. Guo, Z.; Thirumalai, Honeycutt, J. D. *J. Chem. Phys.* **1992**, *97*, 525.

(16) The length of the most compact conformations can slightly vary due to the differences in the distribution between monomers in the "interior" and those on the "surface" of the compact structure, but we can neglect this variation when the protein length L is large enough.

(17) The origin of degeneracy in the absence of the segregation term can easily be explained. In this case the energy of a structure $\vec{n}$ with sequence $\vec{q}$ is $E = \vec{n}\cdot\vec{q}$. The ith component of $\vec{n}$ corresponds to the number of nearest neighbors of the ith monomer. Usually, n_i varies in the range of 0−12 taking only integer values. A permutation of amino acid residues corresponding to monomers with the same number of nearest neighbors will clearly not lead to a change in energy since the solvent exclusion term is additive, which leads to the degeneracy with respect to the shuffling of amino acid residues. Although the energy remains unchanged, shuffling these amino acid residues will influence the overlap parameter PH_0.

MA961564J