Protein ground state candidates in a simple model: An enumeration study

V. Shahrezaei,[1,2] N. Hamedani,[1,2] and M. R. Ejtehadi[1]
[1]*Institute for studies in Theoretical Physics and Mathematics, P.O. Box 19395-5531, Tehran, Iran*
[2]*Department of Physics, Sharif University of Technology, P.O. Box 11365-9161, Tehran, Iran*
(Received 11 May 1999)

The concept of the reduced set of contact maps is introduced. Using this concept we find the ground state candidates for a hydrophobic-polar lattice model on a two-dimensional square lattice. Using these results we exactly enumerate the native states of all proteins for a wide range of energy parameters. In this way, we show that there are some sequences which have an absolute native state. Moreover, we study the scale dependence of the number of members of the reduced set, the number of ground state candidates, and the number of perfectly stable sequences by comparing the results for sequences with lengths of 6 up to 20.
[S1063-651X(99)07710-7]

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

I. INTRODUCTION

The proteins are biomacromolecules, which are made from thousands of atoms. These atoms are in interaction with each other and water molecules, which surround them. Basically, to determine the states of a protein one needs to solve the problem with standard quantum mechanical calculations, however, the complexity of these macromolecules renders this impossible. A feasible approach to this problem is based on a coarse-grained view. In this viewpoint the proteins are made from 20 types of monomers (amino acids). The most important point in this approach is the determination of the effective interactions between the amino acids [1]. It seems that the information about effective intermonomer interaction energy and the coding of the amino acids in the sequence is sufficient to determine the protein characteristics.

The structural information for protein structures can be coded in a contact map [2]. A contact map is a binary $L \times L$ matrix C. The element c_{ij} of this matrix is nonzero if ith and jth monomers are in contact. The contact may be defined in several ways. It is obvious that the information coded in contact maps is not sufficient for a complete characterization of the spatial configuration. The short-range nature of intermonomer interactions suggests that one can determine the configuration energy in terms of contacts. Thus if one knows the effective intermonomer interactions in this coarse-grained approximation, the contact maps have sufficient information to calculate the configuration energy. There are many papers which study the thermodynamical and structural properties of proteins by using contact maps [3].

It is well known that the biological functionality of proteins depends on the shape of their native states. The native structure is the unique minimum free energy structure for the protein sequence [4]. As any protein in nature must have a well-defined function, the uniqueness of native states is a biological necessity for these molecules of life. Thus searching the configuration space to find native states by using the Monte Carlo methods [5] or exact enumerations [6–8] has been the subject of many papers. In most previous works, the problem was studied for given values of intermonomer energy parameters. As our knowledge about the effective interactions is not certain, and the native structures of proteins may be sensitive to these parameters [9], looking at the native states for different energy parameters is relevant [8,10]. By using a simple hydrophobic-polar (HP) lattice model we have shown in a recent work [11] that the number of ground state candidates for any sequence is unexpectedly small. This suggests that the problem can be studied for a wide range of interaction parameters by exact enumeration. We study this problem on a two-dimensional square lattice. In this approach a protein structure is modeled by a self-avoiding walk on the lattice, and any pair of monomers which are nearest neighbors and are not adjacent according to sequence (nonsequential neighbor) are in contact.

The number of possible configurations for an L-mer is equal to the number of self-avoiding walks (N_{SAW}) with $L-1$ steps. We have

$$N_{\mathrm{SAW}} \sim L^{\gamma-1} z_{\mathrm{eff}}^{L}, \tag{1}$$

in which γ is a dimension-dependent constant, and z_{eff} is the effective coordination number. For a two-dimensional square lattice, $\gamma = \frac{43}{32}$ and $z_{\mathrm{eff}} = 2.64$ [12]. Since many of these walks give the same contact matrix, the number of possible contact matrices (physical maps) N_c is much smaller, although it is still very large. In a recent work [13] the number of physical maps was fit to a formula similar to Eq. (1) and a value of $z_c = 2.29$ was obtained.

If one is interested only in the native structure of proteins, the set of the contact maps can be reduced further by removing all maps which have no chance to be a native state. We call the remaining maps the *reduced set of contact maps*. Indeed, this reduction is due to the physical fact that all effective interactions between amino acids are negative [1]. This reduced set of contact maps can be used in enumeration studies to find the possible ground states and the native states of proteins. In this paper we use a simple HP lattice model to address the problem for proteins with various lengths in more detail. We obtain some ground state candidates that possess some known properties common to real proteins. Also a stability against the variation of interaction parameters is shown. Some evidence for this stability has been reported in some other works [14].

II. REDUCED CONTACT MAPS

The effective potential energies between the 20 types of amino acids can be described by a 20×20 interaction matrix [1]. The energy of a given sequence $\boldsymbol{\sigma}$ in any structure can be determined from

$$E=\sum_{i,j} c_{ij}m_{\sigma_i\sigma_j}. \tag{2}$$

The c_{ij} and m_{ij} are, respectively, the elements of the contact matrix (C) and the interaction matrix (M). This shows that all configurations which have the same contact map have equal energies. If we look at the energy spectrum of one sequence, the states corresponding to such maps are degenerate. We call such degeneracies, type-1 degeneracies to distinguish them from other kinds of degeneracies, which we shall encounter later [11]. If the energy of a sequence is minimum in such states, this sequence does not have a unique native state. Such sequences are not proteinlike. The states corresponding to such *degenerate contact maps* can never be a native state, however, we cannot exclude them from our search, because they compete with other maps. On the other hand, there are some maps which cannot be the ground state and do not have a role in the competition for the ground state. To see that, consider two contact matrices C_1 and C_2 and their subtraction ($C'=C_1-C_2$). We call C_2 a *component* of C_1 if all elements of C' are non-negative ($c'_{ij}=0$ or 1). Note that C' has at least one nonzero element. Using Eq. (2), the energy of an arbitrary sequence $\boldsymbol{\sigma}$ in the configuration(s) corresponding to the map C_1 can be written as

$$\begin{aligned} E_1&=\sum_{i,j} c_{1,ij}m_{\sigma_i\sigma_j}\\ &=\sum_{i,j} c_{2,ij}m_{\sigma_i\sigma_j}+\sum_{i,j} c'_{ij}m_{\sigma_i\sigma_j}\\ &=E_2+\sum_{i,j} c'_{ij}m_{\sigma_i\sigma_j}. \end{aligned} \tag{3}$$

According to experimental data all elements of interaction matrix M are negative [1]. Thus the second term in the right-hand side gives a negative contribution to energy, and $E_1<E_2$ for any sequence. Then map C_2 can never be a ground state. One can find all component maps such as C_2, and remove them from the set of contact maps. Indeed such component maps are related to configurations which can fold to more compact shapes without losing any of their old contacts. By this procedure, a reduced collection of maps is found. We call this collection the *reduced set of contact maps*, and we represent the number of its elements by N_r. We have enumerated N_r for sequences with lengths up to 20. The results are shown in Fig. 1. In this figure the number of reduced maps (N_r) are compared with the number of self-avoiding walks (N_{SAW}) and the number of physical maps (N_c), on a two-dimensional square lattice. Although all of these quantities have similar behaviors, the growth rate of N_r is much slower than the others. If we fit the data to Eq. (1) we obtain $\gamma_r=1.37$ and $z_r=2.01$. These results are not enough to see whether the value of γ_r is lattice dependent. In the case of self-avoiding walks it is lattice independent [12]. In Fig. 1 there are other points which show the number of native states. We will discuss this matter in Sec. IV.

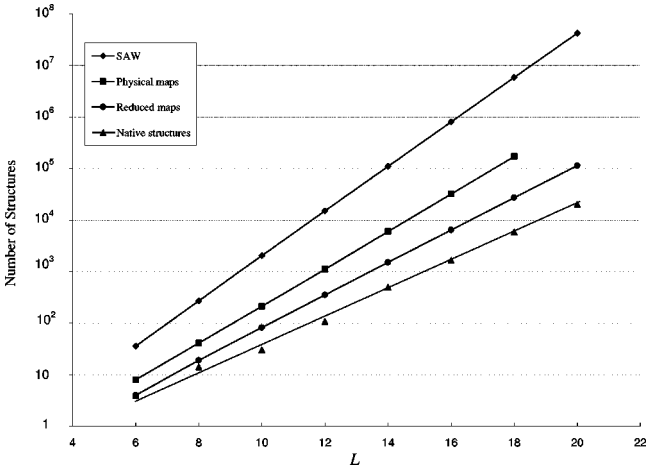

FIG. 1. The number of self-avoiding walk structures, physical contact maps, reduced set of contact maps, and native structures, vs the number of monomers in sequences.

Let us consider the number of contacts ($b=\frac{1}{2}\Sigma_{i,j}c_{i,j}$) as a measure for the compactness of configurations. Indeed, a better parameter is the relative compactness Γ,

$$\Gamma=\frac{b}{b_{\max}}, \tag{4}$$

where $b_{\max}$ is the maximum number of possible contacts for sequences of the same length. The maximum of contacts $b_{\max}$, for sequences of length 6, 8, 10, 12, 14, 16, 18, and 20 are 2, 3, 4, 6, 7, 9, 10, and 12, respectively. In Fig. 2, the number of members of the reduced set of contact maps vs the number of contacts is compared with the corresponding number of SAWs and physical maps for proteins of length 18. We see that the reduced set of contact maps contains only highly compact configurations. This shows why the results of studies on compact structure spaces are reasonable. In Fig. 3, the average compactness for SAWs, physical maps, and reduced maps is compared for sequences of various

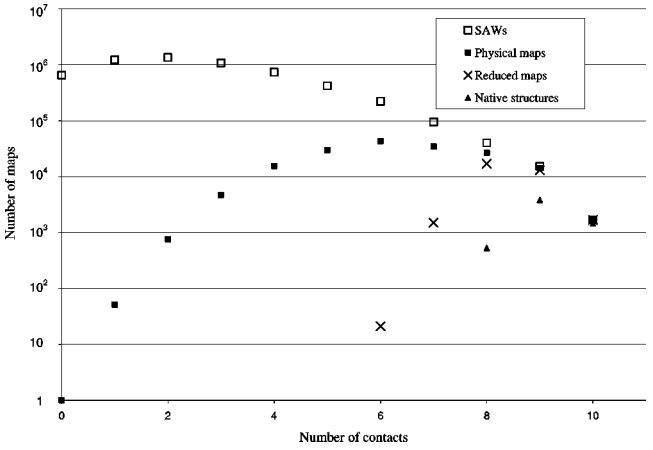

FIG. 2. Distribution of the number of structures vs the number of contacts for sequences of length 18.

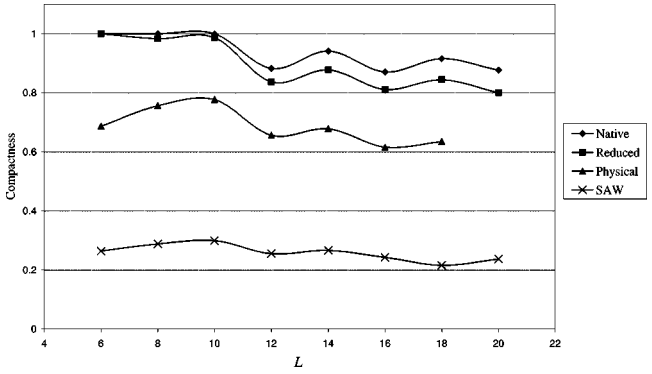

FIG. 3. The average compactness of structures for SAW, physical maps, reduced maps, and native structures, vs the number of monomers in sequences.

lengths. As one can see, $\langle\Gamma_{\mathrm{SAW}}\rangle<\langle\Gamma_c\rangle<\langle\Gamma_r\rangle$. There is an oscillatory behavior in the graphs. Note that the $b_{\max}$ is an integer. The highest ratio of $b_{\max}$ to length (L) is for sequences can be fitted to a square structure. Thus, sequences with such lengths have lower average compactness. This is due to the finite size effect and also the fact that the number of contacts has to be an integer; the same behavior can be observed in our other results in this paper too.

If one scales the number of reduced maps (N_r) by the number of total structures (N_{SAW}) at each compactness, a scale-independent behavior can be seen (Fig. 4). It also seems that there is a critical compactness, below which the compactness of the members of the reduced set never drops. We do not have an exact analytical proof, but it seems from these data that a transition occurs in the number of reduced maps near the compactness of 0.8 and it vanishes for a compactness below 0.5.

Contact maps can correspond to more than one structure. We call such maps *degenerate maps*. These maps cannot correspond to the native state of any sequence. Within the set of reduced maps there are fewer of such degenerate maps than within the set of physical maps. Figure 5 compares the percentage of nondegenerate maps for reduced and physical maps. It seems that both of them approach asymptotic values.

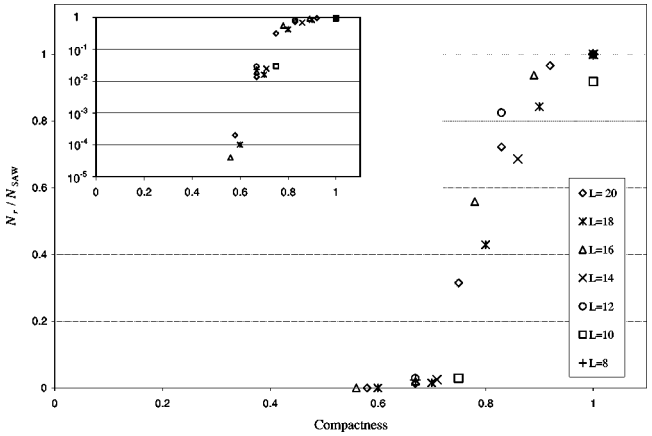

FIG. 4. The number of reduced maps that scaled by the number of all structures at each compactness, for sequences with length 8–20. There is a transition near 0.8 and a cutoff near 0.5. The latter can be seen better in logarithmic scale (inner graph).

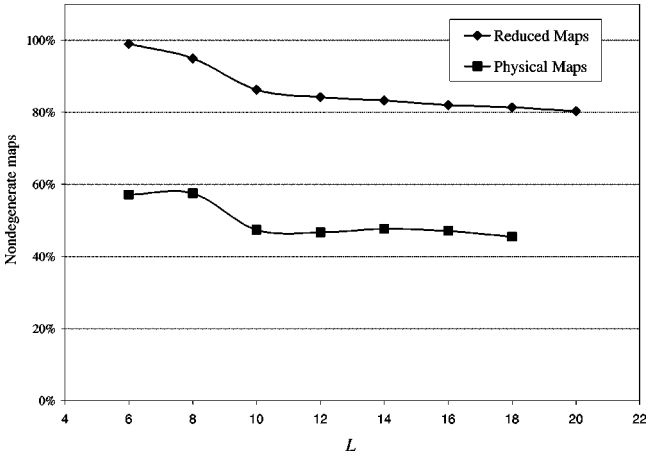

FIG. 5. The percentage of nondegenerate maps for reduced and physical maps.

III. GROUND STATE CANDIDATES FOR THE HP MODEL

The native states of proteins are to be found among the structures corresponding to the reduced set of contact maps. The sequence of the amino acids along the protein chain and their interactions have an essential role in the selection of a particular structure as the native state. In the coarse-grained viewpoint, the interaction between the amino acids is characterized by the effective energies. These effective interactions depend on the properties of the solutions. A relevant question is how sensitive the native structures are to changes in these interactions. We address this question by enumerating the possible ground states of protein sequences for a wide range of effective intermonomer interaction energies.

Without any loss of generality, we use a hydrophobic-polar two-dimensional lattice model [15] in this paper. The general form of the interactions between H and P monomers in a HP model can be written as follows [8,16]:

$$E_{HH}=-2-\gamma-E_c,$$

$$E_{HP}=-1-E_c, \tag{5}$$

$$E_{PP}=-E_c,$$

where $E_{\sigma\sigma'}$ is the contact energy between monomers of types σ and σ'. These potential energies are only between nonsequential nearest neighbors. Here γ and E_c are the mixing and compactness potentials, respectively, two parameters which are determined from experimental data. There are many publications based on this model, and in most of them the values of γ and E_c are fixed [16,15]. Here, we consider them as two free parameters and discuss our results in terms of them.

It has been argued that the following relations should hold between intermonomer energies:

$$E_{HH}<E_{HP}<E_{PP},$$

$$E_{HH}+E_{PP}<2E_{HP}. \tag{6}$$

These arguments are based on the compactness of the native states [17] and some calculations on 20×20 intermonomer interaction matrix M [18]. These restrict γ and E_c to positive values $(\gamma,E_c>0)$.

At first sight, it might seem possible to arrive at any native state for a given sequence by changing γ and E_c. But when we consider the geometrical properties of the ground state, we will find that these parameters are not powerful enough to select any configuration as the native state. In other words, the native states are stable against the change of interaction parameters.

If we consider $H=-1$ for hydrophobic monomers and $P=0$ for polar monomers, a given sequence can then be represented by a binary vector $(\boldsymbol{\sigma})$ [8]. The energy of this sequence in a configuration characterized by a contact matrix C can be written as

$$E=-m-a\gamma-bE_c, \tag{7}$$

where m, a, and b are three integers, related to $\boldsymbol{\sigma}$ and C as follows:

$$\begin{aligned} m&=-\boldsymbol{\sigma}^t\cdot C\cdot\mathbf{1},\\ a&=\frac{1}{2}\boldsymbol{\sigma}^t\cdot C\cdot\boldsymbol{\sigma},\\ b&=\frac{1}{2}\mathbf{1}^t\cdot C\cdot\mathbf{1}. \end{aligned} \tag{8}$$

It can be seen that m is equal to the number of all nonsequential neighbors of H monomers in the configuration, a is the number of $H-H$ contacts, and b is the number of all contacts. It can be shown that the following inequalities hold between these parameters [19].

$$m-b\leqslant a\leqslant\frac{m}{2}\leqslant b. \tag{9}$$

Equation (7) suggests that the energy levels of a given sequence can be described by three integer numbers (m,a,b). It is highly probable that these states are degenerate. There are three types of degeneracy.

Type 1: $C=C'$. In this case two or more configurations with different shapes have the same contact matrix. These configurations will remain degenerate for any sequence, and any choice of γ and E_c. These are the configurations corresponding to the degenerate maps already mentioned in Sec. II. This type of degeneracy is more probable for configurations with low compactness (see Fig. 2). Note that we are not talking about the configurations which are related to each other by spatial symmetries, i.e., rotation, reflection, etc., for our purpose such configurations are identical.

Type 2: $(m,a,b)=(m',a',b')$ but $C\neq C'$. In this case one particular sequence has the same m, a, and b values in two or more configurations. This degeneracy persists for any value of γ and E_c, but may disappear for another sequence. Although this degeneracy depends on sequence coding, the $b=b'$ condition is purely geometrical, and is a necessary condition for this degeneracy.

Type 3: $E=E'$, but $(m,a,b)\neq(m',a',b')$. One sequence has the same energy in two different states (m,a,b) and (m',a',b'), provided that γ and E_c obey the following relation:

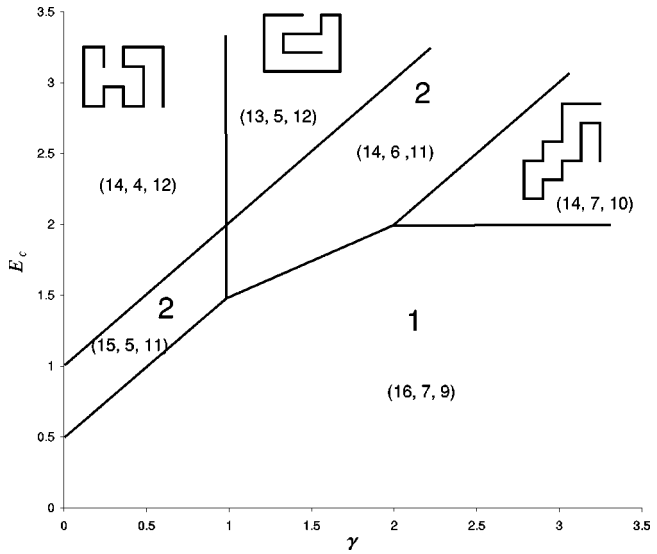

FIG. 6. The space of energy parameters (arbitrary units) for sequence $HPPPHPHPHPPHPHPHPHHP$ is divided into six cells. The integer numbers (m,a,b), inside any cell indicate the ground state corresponding to the cells. Three of these states are degenerate. The types of degeneracies for degenerate states and shape of structures for nondegenerate states are indicated in the cells.

$$(m-m')+(a-a')\gamma+(b-b')E_c=0. \tag{10}$$

This degeneracy is related to both sequence coding $\boldsymbol{\sigma}$ and intermonomer interactions.

The first type of these degeneracies is completely geometric. The second one depends on both geometry and the amino acids' coding sequence. These two types do not depend on the values of the interaction energies. Thus, in the energy spectrum of any sequence there are some states which are degenerate independently from the potential. If the ground state of a particular sequence is one of these degenerate states, that sequence does not have a unique native structure.

The third type is not actually a degeneracy at all. Equation (10) corresponds to a line in the parameter space of E_c and γ. This line is a level crossing line. Degeneracy actually occurs only on the line, and a highly accurate fine-tuning is needed to reach a point on this line. For the two sets of interaction energy parameters on the two sides of this line, the energy ordering of the states is different. For any pair of states such an ordering line exists. By drawing all ordering lines in the space of E_c and γ, this space is divided into many ordering zones. We are only interested in the ground state, which means that many of these ordering lines are not relevant. Some of them only govern the ordering of the excited states. By removing the irrelevant lines, one gets a diagram which shows the ground state cells (Fig. 6). As mentioned before, changing the intermonomer interaction parameters inside any of these cells does not change the ground state. In some recent works [20] this picture is introduced to show the stability of native states against change in the interaction parameters [21]. They only looked at one of these cells in the neighborhood of some selected interaction values. But by looking at the whole energy space, one can find all possible ground states and their corresponding cells. Any such cell in the space of energy parameters is associated

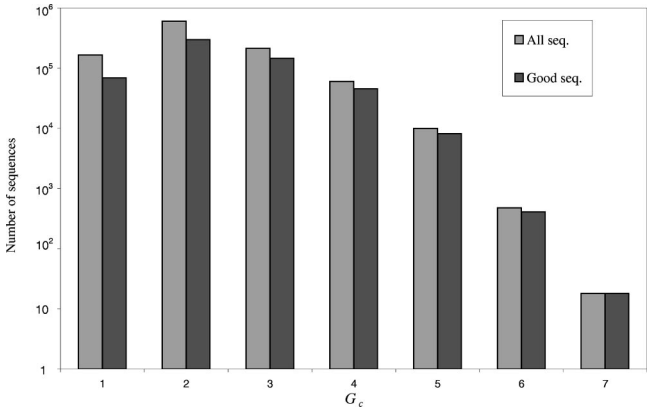

FIG. 7. The histogram of the number of ground state candidates for 20-mers. The light and dark gray areas show the results for all sequences and good sequences, respectively. There are some ''good sequences'' with only one ground state candidate.

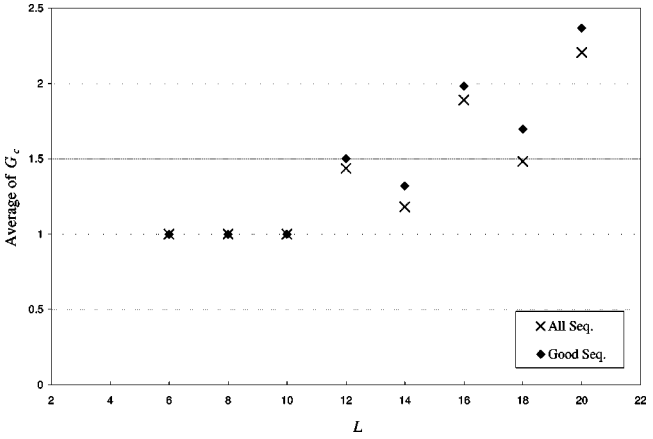

FIG. 8. The average of the number of ground state candidates for all sequences and good sequences vs length of sequences.

with one ground state candidate. The number of cells is equal to the number of ground state candidates $[G_c(\boldsymbol{\sigma})]$. By drawing such diagrams, one can easily find the ground state for any choice of E_c and γ. Figure 6 shows this diagram for a 20-mer. In this example there are only seven possible ground states. The cells marked with the numbers ''1'' and ''2'' correspond to type-1 and type-2 degenerate states, respectively, therefore there is no unique native structure for these cells. The sequence in this example has three nondegenerate states. These structures are shown in the figure. It is possible that all the ground state candidates of a given sequence are degenerate. These sequences constitute universally bad sequences, i.e., for any set of interaction parameter values they do not have a native structure. Any sequence which is not a bad sequence we call a *good sequence*. Nearly 54% of the sequences of length 20 are good sequences, i.e., for some specific set of energy parameters they have a native state.

The interesting point in Fig. 6 is that the number of ground state candidates is very small. The largest values of G_c, for sequences with length 6,8,10,12,14,16,18,20 are 1,1,1,3,4,5,6,7 respectively. Figure 7 shows the histogram of $G_c(\boldsymbol{\sigma})$ for all sequences with $L=20$. The light gray area in this figure shows the result for all 2^{20} sequences, and the dark area shows the results for good ones. From this diagram it can be seen that the mean value of $G_c(\boldsymbol{\sigma})$ is very small. The average of $G_c(\boldsymbol{\sigma})$ for various lengths is shown in Fig. 8. However, the data in hand are not enough to draw a reliable conclusion about the number of ground state candidates for sequences of large length, but the average number does not seem to grow very rapidly, and the growth rate appears to be linear. Extrapolating the growth rate to sequences of length 200, 30 ground state candidates are predicted on average. Comparison of the average value of $G_c(\boldsymbol{\sigma})$ for these sequences with the number of all configurations (i.e., for sequences with length 20 the number of sequences is on the order of 10^8) shows that the geometric constraints play an important role in selecting a state as the ground state. The reason that there are few ground state candidates for any sequence can be given by a geometrical argument [11]. This argument shows that the upper estimate for maximum G_c is L^2.

As Fig. 8 shows, there are some good sequences with $G_c=1$. This means that for any set of energy parameter values, they have the same unique ground state. Figure 9 shows some of these sequences and their unique native structures. Indeed the native states of these sequences have perfect stability with respect to a change of the energy parameters. Our enumeration shows that these *absolute native structures* are to be found among the most compact structures. As Fig. 10 shows, although the ratio of the number of *perfectly stable sequences* to the number of all possible proteins decreases with increasing L, their actual number increases. This suggests that for the proteins with typical lengths near that of natural proteins, perfectly stable sequences constitute a small but nonzero fraction of all possible sequences. A relevant question is whether the existence of these perfectly stable sequences is due to the simplifications in our model. Actually we cannot give an exact answer to this question, but such sequences may exist in models with more monomer types.

The existence of these sequences may answer some questions about protein folding. Their number is small compared with the huge number of the possible amino-acid sequences, their native states are highly compact and are stable against the changes in the intermonomer interactions (i.e., the properties of the solution).

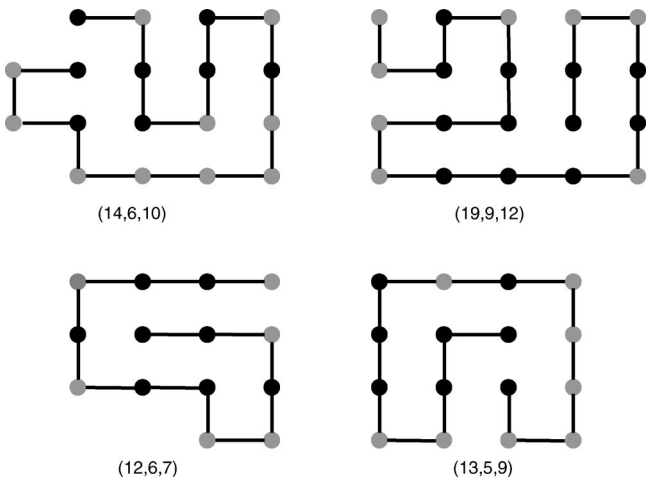

FIG. 9. Four examples of perfectly stable sequences and their absolute native structures. For any positive value of γ and E_c these sequences are folded uniquely in the structures shown.

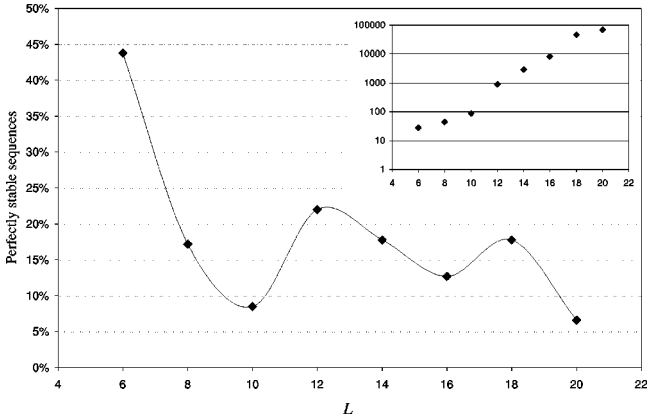

FIG. 10. The ratio of the numbers of perfectly stable sequences to all sequences decreases with length of sequences, but their absolute numbers increase (inner graph).

IV. NATIVE STRUCTURES

In Sec. II we introduced the reduced set of contact maps. As was shown, the number of maps belonging to this set, N_r, is much smaller than the number of structures, N_{SAW}. But the number of those structures which can be the native state is still much smaller. The number of possible native structures, N_{native}, is shown in Fig. 1. In this figure all those structures which have been the native state of some sequence for at least one set of energy parameter values have been counted. Fitting the data on an equation similar to Eq. (1), gives $\gamma_{\mathrm{native}}=1.87$ and $z_{\mathrm{native}}=1.68$. In Fig. 2, we have compared the number of native structures as a function of their compactness with the total number of physical maps and with the number of maps in the reduced set for $L=18$. It can be seen that there are no native structures with fewer than eight contacts. Also the average compactness of native states is compared in Fig. 3.

We can also look at the designability of structures. The designability shows how many times a structure is selected as the native state for a fixed set of interaction parameter [7]. We observed that the distribution of designability is energy dependent, but the most highly designable structures for the different values of energy parameters are almost the same. We can introduce a global designability parameter D for the native structures. It is a bit different from the above designability definition. In our case we correspond to any structure a set of sequences. Any of these sequences select this structure as the unique ground state, at least in a small region of energy parameter space. The global designability of any structure is the number of the members of its corresponding set, i.e., we count how many times a structure becomes the candidate for a nondegenerate ground state.

Figure 11 shows the histogram of designability for structures with length 20. As one can see, the results are very similar to those for a fixed set of energy parameters in the space of compact structures [8,7]. The average designability as a function of compactness for $L=20$ is shown in Fig. 12. As the diagram shows, the peak average designability occurs for the most compact structures and it falls sharply with decreasing compactness. Thus if one is only interested in highly designable structures, it is reasonable to search the space of compact structures.

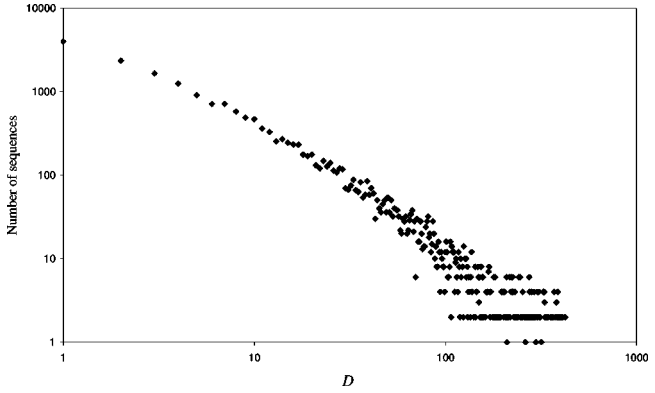

FIG. 11. The histogram of number of structures with a given designability.

V. SPACE OF ENERGY PARAMETERS, E_c AND γ

One of the important aspects of the work done in this paper is that we can find the exact results for any range of energy parameters. The time it takes for this program to find the ground state candidates for all sequences by exact enumeration is on the same order as that of the usual enumeration schemes for only one particular set of energy parameters. Because the average number of ground state candidates is very small, the determination of the native ground states for any range of interest only takes a little time. We found the native states of all sequences of length 20, for all pairs of energy parameters within a 12×12 square in arbitrary units, with a grid size of 0.1 (14 400 points). The number of proteinlike sequences (sequences which have unique ground states) is shown in Fig. 13. As one can see, there are jumps in the number of proteinlike sequences. These jumps specify the borders of regions of relative stability within the space of energy parameters. A closer examination shows that these borderlines contain sharp dips adjacent to the jumps. The large changes in the number of proteinlike sequences show that when we cross these borders the ground states of many sequences change, and the degenerate ground states are replaced by nondegenerate ones (or vice versa). However,

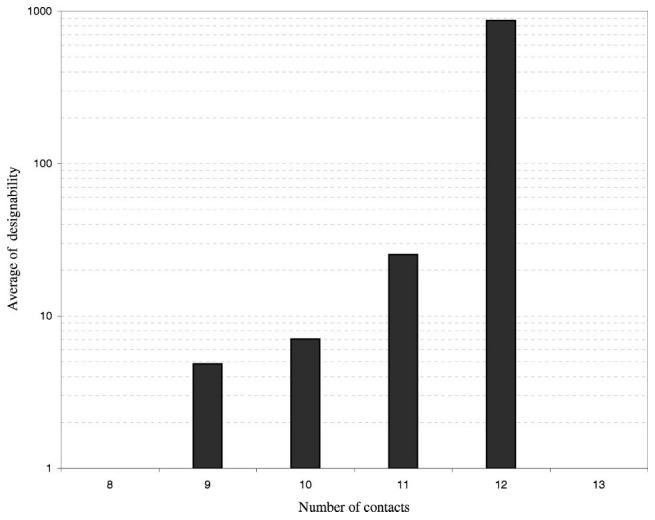

FIG. 12. The average designability for structures with a given number of contacts, for $L=20$.

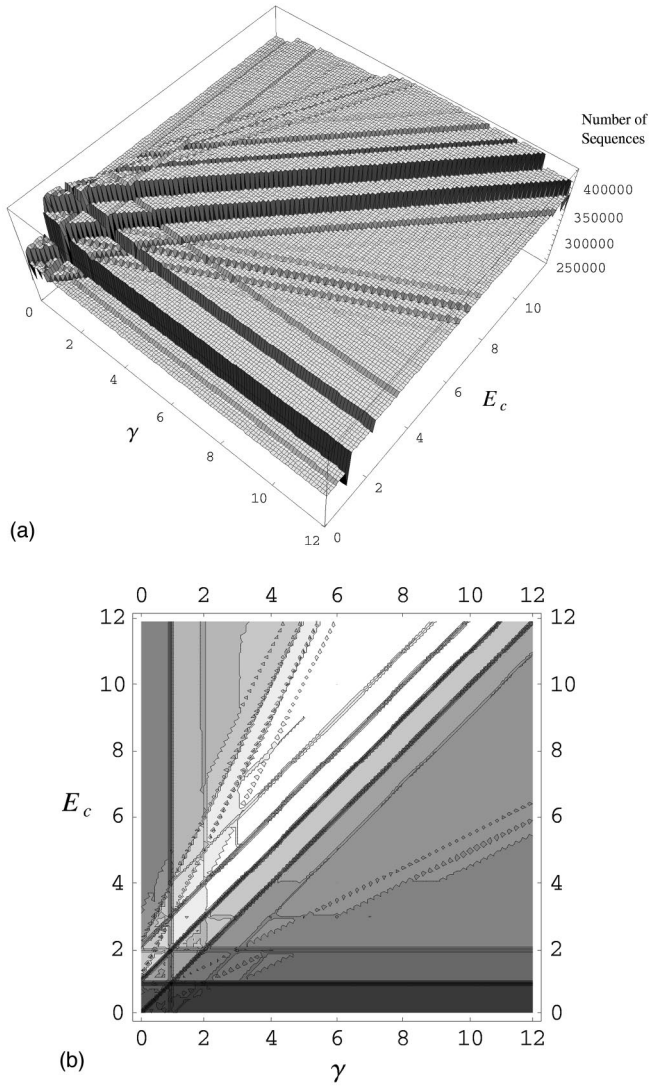

FIG. 13. The number of proteinlike sequences of length 20, for given values of energy parameters in a 12×12 square region (arbitrary units); (a) three-dimensional plot, (b) contour plot.

nothing can be said about the details of these changes. One can get some idea about what is happening on these borderlines by comparing the contour plot for Fig. 13(a) [Fig. 13(b)] with the ordering lines diagram for one particular sequence (Fig. 6). As mentioned in Sec. III, the ordering lines specify level crossings and type-3 degeneracies only occur on the ordering line itself. These ordering lines constitute the underlying cause of the sharp dips observed in the borders. This is more evident in Fig. 14. In this figure we have shown those points in the energy parameter space where at least one type-3 degeneracy occurs. This diagram is in fact a superposition of diagrams like Fig. 6, for all sequences, and any line in it corresponds to many ordering lines between ground state candidate cells.

We can find similar information for other types of degeneracies. For example, the number of sequences which have type-1 degenerate ground states is shown in Fig. 15. As one can see in this diagram, the number of such sequences vanishes for large E_c and small γ. For large E_c the number of contacts b plays an essential role in the selection of the

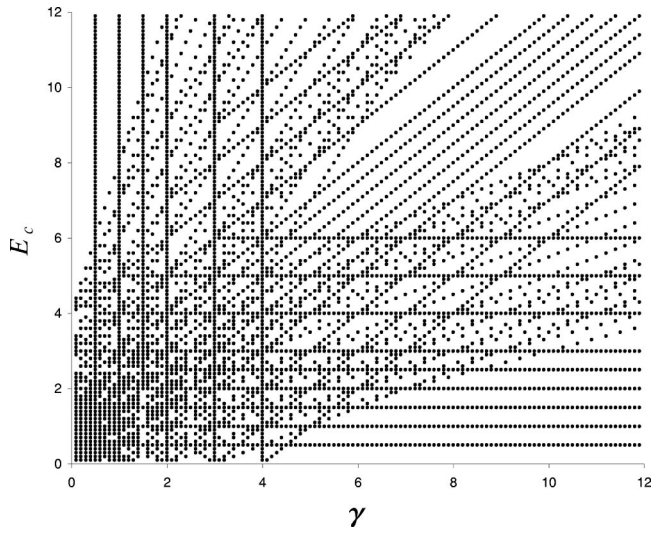

FIG. 14. The points in the energy parameter space (arbitrary units), where type-3 degeneracies occur, for sequences of length 20. The grid size is 0.1.

ground state [Eq. (7)]. Type-1 degeneracies do not occur for highly compact sequences (see Fig. 2). Thus this type of degeneracy is more relevant in the region $E_c<\gamma$. We have not shown the corresponding information for type-2 degeneracies as they contain no new information; similar border jumps can be observed in the number of sequences with this type of degeneracy too. The maximum percentage of sequences with nondegenerate, type-1 degenerate, and type-2 degenerate ground states in the chosen region are 40.0%, 5.06%, and 64.9%, respectively.

In addition to obtaining information about the sequences, this procedure also finds the ground states. Since the energy parameters determine which states are the ground states, the number of structures which can be the native state of some particular sequence also depends on the energy parameters. Figure 16 shows the number of native states as a function of the energy parameters. The importance of compactness for large values of E_c can also be seen in this diagram. Note that the smallest value for the number of native states is 503. This number corresponds to the number of most compact struc-

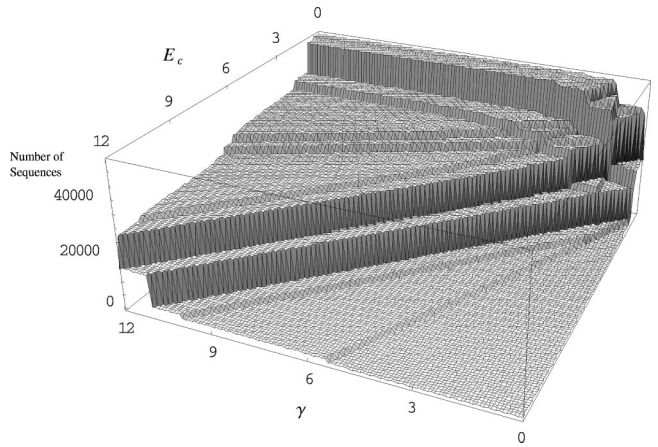

FIG. 15. The number of sequences of length 20, with type-1 degenerate ground states, for given values of energy parameters in a 12×12 square region (arbitrary units).

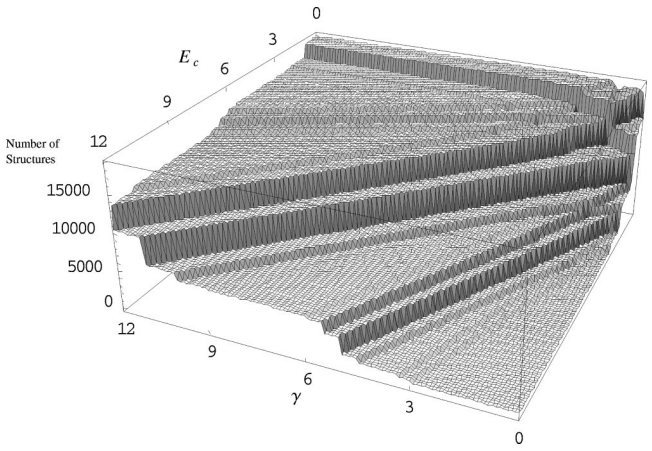

FIG. 16. The number of native states for sequences of length 20, for given values of energy parameters in a 12×12 square region (arbitrary units).

tures of length 20. Again, large jumps in the number of native states are observed. One can also find the average designability of the structures by dividing the data of Figs. 13 and 16 (the ratio of the number of sequences to a corresponding number of native structures).

VI. CONCLUSION

Due to the short-range nature of intermonomer interactions, the configuration energy of protein sequences can be determined by using configuration contact matrices. In this paper, it has been shown that for this class of problems, where one is interested in native states of proteins, the space of physical contact maps can be reduced to a much smaller set by removing all irrelevant maps. We have found the reduced set of contact maps for sequences of lengths up to 20 in this paper by exact enumeration. This reduced set of contact maps shows a scale-independent behavior as shown in Fig. 4.

Using the reduced set of contact maps, the ground state candidates for all sequences were found in the HP model. The number of these ground state candidates is quite small. However, we limited ourselves to the HP model in this work, but some arguments [11,19] show that the number of ground state candidates is restricted in models with more monomer types. The ground state candidates divide the space of energy parameters into several cells. By finding this cell structure for all sequences, we have found the native states for all sequences of different lengths, for a wide range of energy parameters. Jumps are observed in the number of proteinlike sequences. These jumps are related to boundaries of the aforementioned cells.

Another interesting result is that we find some sequences with absolute native states, i.e., their native states are not sensitive to the values of energy parameters. Our results show that the number of such perfectly stable sequences grows with length, however, their percentage decreases.

Because the key tool used in this paper has been the structural information contained in the contact maps, the qualitative results can be generalized to all contact models, regardless of the details of the lattice, the contact rules, and the number of monomer types.

ACKNOWLEDGMENTS

We would like to thank S.E. Faez, R. Gerami, R. Golestanian, A.Yu. Grosberg, N. Heydari, M. Khorami, S. Rouhani, and H. Seyed-Allaei for helpful comments.

[1] S. Miyazawa and A. Jernigan, Macromolecules **18**, 534 (1985).
[2] S. Lifson and C. Sander, Nature (London) **282**, 109 (1979).
[3] For a recent review, see M. Vendruscolo and E. Domany, e-print cond-mat/9901215.
[4] C.B. Anfinsen, Science **181**, 223 (1973).
[5] C. Camacho and D. Thirumalai, Proc. Natl. Acad. Sci. USA **90**, 6369 (1993); A.M. Gutin, V.I. Abkevich, and E.I. Shakhnovich, Phys. Rev. Lett. **77**, 5433 (1996).
[6] H.S. Chan and K.A. Dill, J. Chem. Phys. **95**, 3775 (1990).
[7] H. Li, R. Helling, C. Tang, and N. Wingreen, Science **273**, 666 (1996).
[8] M.R. Ejtehadi, N. Hamedani, H. Seyed-Allaei, V. Shahrezaei, and M. Yahyanejad, Phys. Rev. E **57**, 3298 (1998); J. Phys. A **31**, 6141 (1998).
[9] V.S. Pande, A.Yu. Grosberg, and T. Tanaka, J. Chem. Phys. **103**, 9482 (1995).
[10] E.L. Kussell and E.I. Shakhnovich, e-print cond-mat/9904377.
[11] M.R. Ejtehadi, N. Hamedani, and V. Shahrezaei, Phys. Rev. Lett. **82**, 4723 (1999).
[12] N. Madres and G. Slade, *The Self-avoiding Walk* (Birkhauser, Boston, 1993).
[13] M. Vendruscolo, B. Subramanian, I. Kanter, E. Domany, and J. Lebowitz, Phys. Rev. E **59**, 977 (1999).
[14] V.I. Abkevich, A.M. Gutin, and E.I. Shakhnovich, Biochemistry **33**, 10 026 (1994); E.I. Shakhnovich, V.I. Abkevich, and O.B. Ptitsyn, Nature (London) **379**, 96 (1996).
[15] H.S. Chan and K.A. Dill, J. Chem. Phys. **90**, 492 (1989); H.S. Chan, K.A. Dill, and D. Shottle, in *Princeton Lectures on Biophysics*, edited by W. Bialek (World Scientific, Singpore, 1992).
[16] R. Melin, H. Li, N. S. Wingreen, and C. Tang, J. Chem. Phys. **110**, 1252 (1999).
[17] K.A. Dill, Biochemistry **24**, 1510 (1985); K.A. Dill, S. Bromberg, K. Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas, and H.S. Chan, Protein Sci. **4**, 561 (1995).
[18] H. Li, C. Tang, and N. Wingreen, Phys. Rev. Lett. **79**, 765 (1997).
[19] N. Hamedani, V. Shahrezaei, and M.R. Ejtehadi (unpublished).
[20] J. Mourik, C. Clementi, A. Maritan, F. Seno, and J. R. Banavar, e-print cond-mat/9801137; C. Clementi, A. Maritan, and J. R. Banavar, Phys. Rev. Lett. **81**, 3287 (1998); F. Seno, A. Maritan, and J. R. Banavar, Proteins: Struct., Funct., Genet. **30**, 244 (1998).
[21] In their work they parametrize the energy space by E_{HP} and E_{PP} instead of γ and E_c.