Stability of preferable structures for a hydrophobic-polar model of protein folding

M. R. Ejtehadi,[1,2,*] N. Hamedani,[1] H. Seyed-Allaei,[1] V. Shahrezaei,[1] and M. Yahyanejad[1]
[1]*Department of Physics, Sharif University of Technology, Tehran 11365-9161, Iran*
[2]*Institute for Studies in Theoretical Physics and Mathematics, Tehran 19395-5531, Iran*
(Received 6 October 1997)

By exact computer enumeration we have calculated the designability of proteins in a simple lattice hydrophobic-polar model for the protein folding problem. We show that if the strength of the nonadditive part of the interaction potential becomes larger than a critical value, the degree of designability of structures will depend on the parameters of potential. [S1063-651X(98)01103-9]

PACS number(s): 87.15.−v, 82.20.Wt, 36.20.Ey, 82.20.Db

Biologically active proteins fold into a native compact structure despite the huge number of possible configurations [1]. In addition to the paradoxical problem of kinetics and time scales of the folding process, there is another mystery. If proteins are made randomly by amino acids, the number of all possible such proteins with typical length of 100 is far larger than the number of proteins which actually occur in nature. Some efforts have been made in order to study the stability of proteins against mutation by searching the two dimensional configuration space [2,3]. One simple model used in these studies is the hydrophobic-polar (H-P) model [4]. In this model there are two types of monomers, H which refer to hydrophobic monomers, and P for polar ones. Recently Li *et al.* [5] have looked at this problem in three dimensions. Calculating the energy of all possible 27-mers in all compact three dimensional configurations, they have found that there are a few structures into which a high number of sequences uniquely fold. These structures were named ''highly designable'' and the number of sequences which fold into each state was named its ''designability.'' In their H-P model, they choose the contact energy between H and P monomers by some physical arguments [5,6]. Other significant points of their work are (a) only a few percent of sequences have unique ground state; (b) there is a jump in energy gap for these highly designable structures. Thus the highly designable structures are more stable against mutation and thermal fluctuation.

Chan and Dill [7] have argued that many of the phenomena observed in proteins can be adequately understood in terms of the H-P model, but according to the work of Pande *et al.* [8] the designability of a conformation does depend on the nature of interactions between monomers. Maybe any interaction leads to some highly designable structures, but different interactions yield different patterns.

In this paper we consider a H-P lattice model, with different intermonomer interactions. We can write the general form of the interaction potential energy in an arbitrary energy scale as

$$E_{PP}=0, \quad E_{HP}=-1, \quad E_{HH}=-2-\gamma, \tag{1}$$

where γ gives the energy change due to the mixing of two types of amino acids [9]. The most usual choice of H-P model potential corresponds to the limit $\gamma \gg 1$ [2–4,7], however, physical arguments are consistent with a smaller value for γ, for instance, $\gamma=0.3$ was used by Li *et al.* [5]. They have calculated the energy of all of 2^{27} sequences in 103 346 compact configurations for a 27-site cube, by a huge enumeration. In particular, it has been suggested that for a random mixing of hydrophobic-polar chain it is reasonable to assume γ to be zero [6,9]. In the case $\gamma=0$, we have an additive potential. If we let $H=-1$, and $P=0$, we can rewrite the potential in the form

$$E_{\sigma_i\sigma_j}=\sigma_i+\sigma_j. \tag{2}$$

Following Li *et al.* [5], we consider only compact structures of sequences with length 27, occupying all sites of a $3\times3\times3$ cube [10]. There are 103 346 compact configurations which are not related to each other by rotation and reflection symmetries.

A protein of length N may be shown by an N-component vector

$$|\sigma\rangle=|\sigma_{i_1},\sigma_{i_2},\ldots,\sigma_{i_N}\rangle, \tag{3}$$

where $i_n=1,2$ refers to P and H residues. Thus the number of such N-component vectors for proteins with length 27 is 2^{27}.

Because of the additive form of the potential, we can write the energy of a given $|\sigma\rangle$ in any spatial configuration as

$$E=\sum_{i=1}^{27} g_i\sigma_i, \tag{4}$$

where g_i's are the number of nonsequential neighbors of the ith monomer, or by introducing the neighborhood vector $|G\rangle$,

$$E=\langle\sigma|G\rangle. \tag{5}$$

The vector $|G\rangle$ has 27 components and at ith component has the number of neighbors of the ith monomer. Due to the shape of $|G\rangle$ the type of neighbors is not relevant and all we

*Electronic address: reza@netware2.ipm.ac.ir

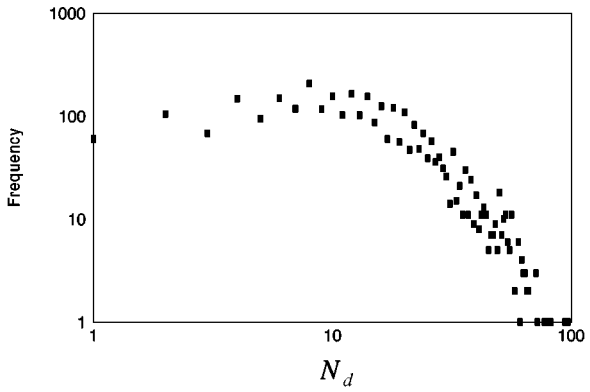

FIG. 1. Histogram of N_d for members of structure space. It is interesting that there are some G sets with $N_d=1$.

have to do is count the nonsequential neighbors. This gives us an additional symmetry for the energy that is different from spatial symmetries.

Due to this additional symmetry the space of all three dimensional structures, which has 103 347 members for all compact fully filled structures in a $3\times3\times3$ cube, is divided into 6291 subspaces, where all members of each subspace have the same $|G\rangle$. Let the number of members of a subspace be N_d. The range of N_d is from 1 to 96. Figure 1 shows that the frequency of large values of N_d is low. Interestingly there are a lot of $|G\rangle$'s which only point to one structure.

We have calculated the energy of all 2^{27} $|\sigma\rangle$ on all $|G\rangle$. We find the degeneracy of ground state in structure space. There are only a few sequences which have nondegenerate ground state; this corresponds to 8.47% of sequences. If energy of one sequence is minimized in a $|G\rangle$ with N_d greater than one it has degenerate ground state. According to the definition of designability, such sequences should not be considered. But if we consider all of the sequences which have nondegenerate ground state, we get a new picture for designability. This means that we calculate the designability of all $|G\rangle$'s, and not only those with $N_d=1$. This is in contrast to N_s, which had only $N_d=1$. To recognize this difference, we show designability of structures by N_s'. Figure 2 shows the distribution of N_s'. Many of the points in this figure are related to some $|G\rangle$'s with $N_d\neq1$. We shall use this picture to express the nature of the energy gap in the case $\gamma\neq0$.

In our enumeration we have calculated the energy of any sequence in all 6291 $|G\rangle$'s, but in Fig. 2, we show the results for 3153 $|G\rangle$'s which are not related to each other by reverse labeling. We cannot reduce the structure space according to this symmetry before enumeration. Reverse labeling for a nonsymmetric sequence gives two different configurations which may have different energies.

The energy gap for all of the sequences is equal to 2. We find it by enumeration in the first, but there is a simple proof. The number of nonsequential neighbors is related to type of site. A $3\times3\times3$ cube has 8 corner sites (C), 12 link sites (L), 6 face sites (F), and one center site (O). C sites have three neighbors, where two of them are connected by sequential links and there is only one nonsequential neighbor. Similarly L, F, and O sites have two, three, and four nonsequential neighbors, respectively. We must add 1 to these numbers

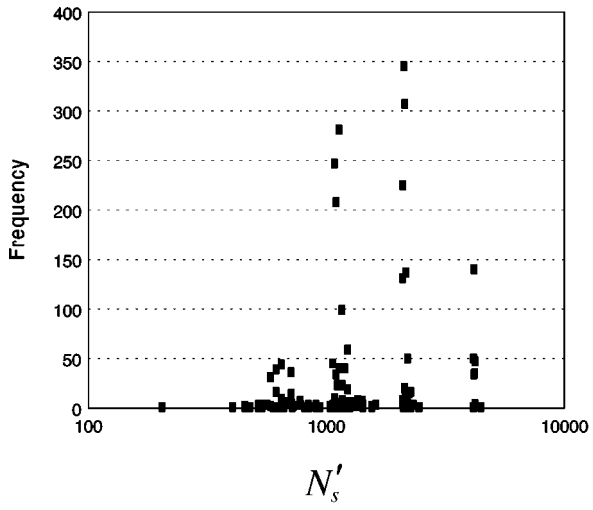

FIG. 2. Histogram of N_s' for additive potential. Note that many of the points in this diagram correspond to some $|G\rangle$'s which point to more than one spatial configuration.

for the two ends of the chain. These sites are divided in two classes, $\{C,F\}$ and $\{L,O\}$. In a self-avoiding walk in this cube, we must jump in any step from one set to the other. The first set has 14 members and the second has 13. Thus a walk passes through C and F sites in odd steps, and through L and O sites in even steps. In other words, the odd components of $|G\rangle$ are 1 or 3, and even components are 2 or 4 (except the first and 27th components which are like even components). Thus

$$|G\rangle=|g_1,\ldots,g_{27}\rangle, \tag{6}$$

where

$$g_i=\begin{cases}1,3, & \text{odd} \quad i\text{'s}\\ 2,4, & \text{even} \quad i\text{'s.}\end{cases} \tag{7}$$

Therefore, the energy for a sequence σ_α in a structure G_μ is

$$E_{\alpha\mu}=\langle\sigma_\alpha|G_\mu\rangle=\sum_{i\in\text{odd}}(g_{\mu i}-1)\sigma_{\alpha i}+\sum_{i\in\text{even}}(g_{\mu i}-2)\sigma_{\alpha i}$$

$$+\sum_{i\in\text{odd}}\sigma_{\alpha i}+2\sum_{i\in\text{even}}\sigma_{\alpha i}. \tag{8}$$

By introducing the new binary variable x the above can be rewritten as

$$E_{\alpha\mu}=\sum_{i=1}^{27}2x_{\mu i}\sigma_{\alpha i}+\sum_{i\in\text{odd}}\sigma_{\alpha i}+2\sum_{i\in\text{even}}\sigma_{\alpha i}, \tag{9}$$

where

$$x_i=\begin{cases}0, & g_i=1 \text{ or } 2\\ 1, & g_i=3 \text{ or } 4.\end{cases} \tag{10}$$

The two last terms in Eq. (9) are independent of $|X\rangle$ or $|G\rangle$, thus they result in a constant, which can be ignored when comparing energies of a sequence in different configurations. The first term in Eq. (9) is an integer times two, thus it

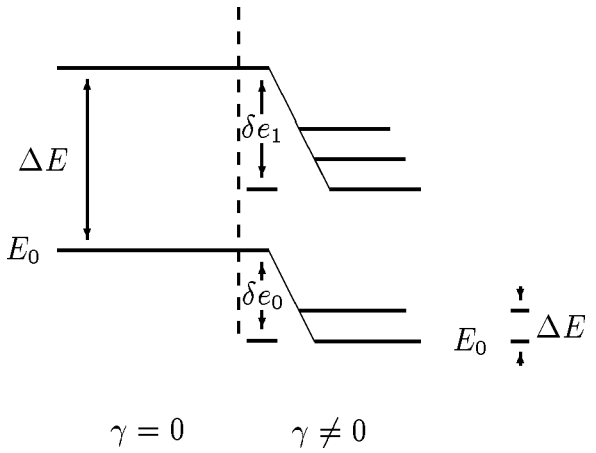

FIG. 3. Energy levels of additive potential split to sublevels for nonadditive potential.

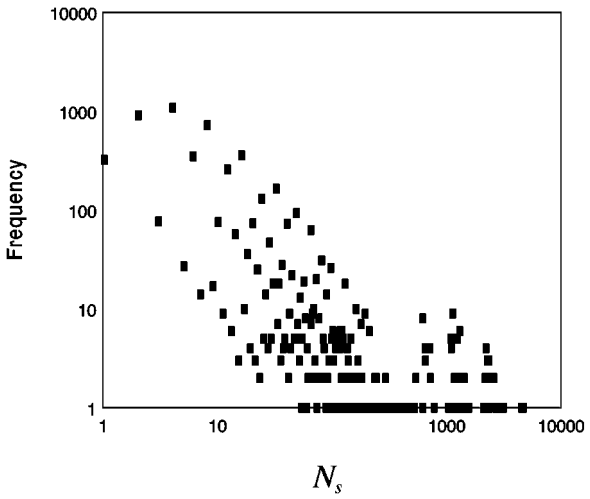

FIG. 4. Histogram of N_s for nonadditive potential.

results in a ladder energy spectrum with gaps of 2. Therefore, the energy gap for all of the structures is the same, and there is no difference between low designable and high designable structures.

In the case $\gamma \neq 0$ the potential is nonadditive. In this case we can write the energy of the αth sequence in the μth spatial configuration as

$$E_{\alpha\mu} = \langle \sigma_\alpha | G_\mu \rangle - \tfrac{1}{2}\gamma \langle \sigma_\alpha | M_\mu | \sigma_\alpha \rangle, \tag{11}$$

where σ and G are the sequence and neighborhood vectors already introduced, and M is the adjacency matrix for this configuration.

Any $|G\rangle$ has N_d different M matrices. The first term in Eq. (11) was calculated in the case $\gamma = 0$, and we need calculate only the last part. The energy spectrum for the previous case has a ladder structure with energy gap equal to 2. In this case these split to some sublevels (Fig. 3).

From the result of the additive potential we know subset G in the space of all spatial configurations which gives the minimum energy to folding for any configuration. This G subset has N_d members all of which have the same $|G\rangle$. For small γ's the ground state and the first excited state are between these N_d structures, and it is not necessary to calculate the energy for all 103 346 spatial structures for any sequence, except for sequences where their ground state is in structures with $N_d = 1$. For the $N_d = 1$ structures the value of N_s does not change, and it is not necessary to run the program, but the first excited states of these sequences are in another G subset. Thus to find the energy gap for them the program must be run over all of the 103 346 structures. We have calculated this energy spectrum, and have found the N_s for all 103 346 structures. We show the results for 51 704 configurations which are unrelated by reverse labeling symmetry in Fig. 4. We have found the energy gap for the first excited state for all sequences. You can see the diagram of mean of energy gap versus N_s in Fig. 5. This figure, in addition to a jump in energy gap for highly designable structures which was observed by Li *et al.* [5], shows that these highly designable structures are related to G subsets with one member.

In this enumeration we have calculated the energy spectrum for all of the sequences which have nondegenerate ground state for the additive potential. We had removed some of the sequences because of degeneracy of ground state in the additive potential case. It is possible that this degeneracy will be removed by the nonadditive part of the potential, and some of the sequences have unique ground state for nonadditive potential. But the energy gap for these sequences is of order of γ, and if we consider them it causes a shift in horizontal axes to bigger N_s and brings down the points nearer to the γ value in vertical direction in Fig. 5. These make this figure more similar to the results of Li *et al.* [5]. In their work the energy gaps for low designable structures are of order of γ (they choose $\gamma = 0.3$) also.

In any compact configuration in a $3 \times 3 \times 3$ cube, there are 28 nonsequential neighbor pairs. Thus the contribution of the nonadditive part of potential in energy is less than 28γ. Then if we choose $\gamma < \frac{2}{28}$ the levels are separate (Fig. 3). Of course this is a lower estimation for γ. The condition that the ground state of sequences does not change is $\delta e_0 - \delta e_1 < 2$. Our enumerations give an upper limit for γ_c equal to 1, which breaks this condition. A combinatorial approach gives a smaller region for γ_c, $0.25 < \gamma_c < 1$ [11]. This shows that there is a nonzero value for γ_c, for which, for γ less than it,

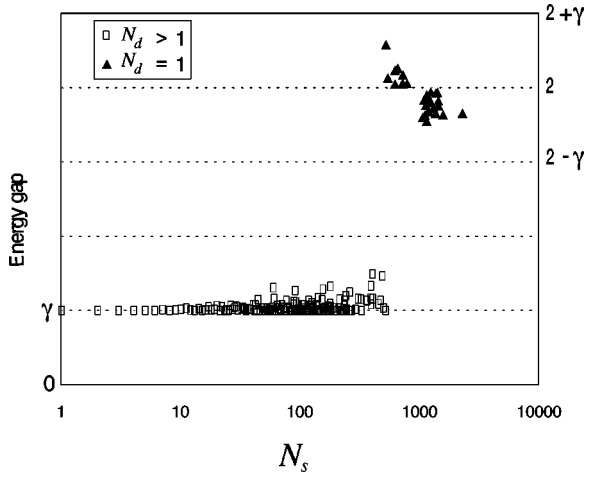

FIG. 5. The mean of energy gap (arbitrary units) vs N_s. There is a jump in energy gap for highly designable structures. All of these highly designable structures have $N_d = 1$.

the ground state structure of sequences does not change. Indeed γ_c distinguishes two phases. If $\gamma<\gamma_c$, the degree of designability of structures is independent of γ, and the change in value of γ only changes the energy gaps. On the other hand, for $\gamma>\gamma_c$, the designability of structures becomes sensitive to the value of γ, and the patterns of highly designable structures will be changed if the potential changes. Although we expect γ_c to be N dependent, its form is still uncertain to us. The upper limit for γ_c is independent of N, but the lower limit does depend on N [11].

If the designability is the answer to ''why has the nature selected a small fraction of possible configurations for folded states?'' the above discussion shows that this selection is potential independent if $\gamma<\gamma_c$, and sensitive to intermonomer interactions if $\gamma>\gamma_c$.

We would like to thank J. Davoudi for motivating the work, R. Golestanian and S. Saber for helpful comments, and S. Rouhani for helpful comments throughout the work and for reading the manuscript.

[1] L. Stryer, *Biochemistry* (W. H. Freeman and Company, San Francisco, 1988); *Protein Folding*, edited by T. E. Creighton (W. H. Freeman and Company, New York, 1992).
[2] H. F. Lau and K. A. Dill, Proc. Natl. Acad. Sci. USA **87**, 638 (1990).
[3] H. S. Chan and K. A. Dill, J. Chem. Phys. **95**, 3775 (1991).
[4] H. S. Chan and K. A. Dill, J. Chem. Phys. **90**, 492 (1989); H. S. Chan, K. A. Dill, and D. Shottle, in *Statistical Mechanics and Protein Folding*, Princeton Lectures on Biophysics, edited by W. Bialek (World Scientific, Singapore, 1992).
[5] H. Li, R. Helling, C. Tang, and N. Wingreen, Science **273**, 666 (1996).
[6] H. Li, C. Tang, and N. Wingreen, Phys. Rev. Lett. **79**, 765 (1997).
[7] K. A. Dill, Biochemistry **29**, 7133 (1990); H. S. Chan and K. A. Dill, Proc. Natl. Acad. Sci. USA **87**, 6388 (1990).
[8] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, J. Chem. Phys. **103**, 9482 (1995).
[9] J. H. Hildebrand and R. L. Scott, *The Solubility of Nonelectrolytes* (Reinhold Publishing Corporation, New York, 1950).
[10] E. Shakhnovich and A. Gutin, J. Chem. Phys. **93**, 5967 (1990).
[11] M. R. Ejtehadi, N. Hamedani, H. Seyed-Allaei, V. Shahrezaei, and M. Yahyanejad, e-print cond-mat/9710028.