Highly designable protein structures and inter-monomer interactions

M R Ejtehadi†‡§, N Hamedani†, H Seyed-Allaei†, V Shahrezaei† and M Yahyanejad†

† Department of Physics, Sharif University of Technology, Tehran, PO Box: 11365-9161, Iran
‡ Institute for Studies in Theoretical Physics and Mathematics, Tehran, PO Box: 19395-5531, Iran

Abstract. By exact computer enumeration and combinatorial methods, 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 non-additive part of the interaction potential becomes larger than a critical value, the degree of designability of structures will depend on the parameters of the potential. We also show that the existence of a unique ground state is highly sensitive to mutation in certain sites.

1. Introduction

Biologically active proteins fold into a native compact structure despite the huge number of possible configurations [1]. Although the mechanism of protein folding is not fully understood, it has been known since the refolding experiments of Anfinsen *et al* [2] that globular proteins fold in the absence of any catalytic biomolecules. From this fact, it has been established that for proteins, the three-dimensional folded structure is the minimum free energy structure, and, the information coded in the amino-acid sequence is sufficient to determine the native structure [3]. The compactness of this unique native state is largely due to the existence of an optimal amount of hydrophobic amino-acid residues [4], since these biological objects are usually designed to work in water [5]. The relation between the primary one-dimensional sequence and the final compact three-dimensional structure is the task of the protein folding problem.

In addition to the paradoxical problem of kinetics and timescale of the folding process [6], 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 do actually occur in nature. One hypothesis is that the naturally selected sequences are special because they are coded for structures that have unique and stable native states, allowing for easy folding. Thus, a central question of protein evolution is how a mutational change in the amino acid sequence leads to changes in the structure and stability.

Some efforts have been made in order to study the stability of proteins against mutation by searching the two-dimensional configuration space [7, 8]. One simple model used in these studies is the H–P model [9]. In this model there are only two types of chain monomers,

§ E-mail address: reza@theory.ipm.ac.ir

hydrophobic (H) and polar (P). Every H–H contact between topological neighbours is assigned a negative contact energy, and other contact interactions are set to zero.

Recently, Li *et al* [10], looked at this problem in three dimensions. Calculating the energy of all possible 27-mers in all compact three-dimensional configurations, they 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 [10, 11]. Other significant points of their work are that (a) only a few per cent of sequences have a unique ground state and (b) there is a jump in energy gap between highly designable and less designable structures. Thus, the highly designable structures are more stable against mutation and thermal fluctuation.

Dill and Chan [12] argued that many of the phenomena observed in proteins can adequately be understood in terms of the H–P model, but according to the work of Pande *et al* [13] 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 [14].

In the first part of our work we study this problem for an additive potential. We will show that there are some highly designable structures for this potential, but the low designable structures will disappear because of degeneracy of ground state. We will show that there is a ladder structure for energy levels for this form of potential. We then contribute a non-additive part to the energy, then the ground-state degeneracy of low-designable structures will be removed. We show that there is a critical value for non-additive part of potential, where below the critical value the patterns of highly designable structures are fixed, but above the critical value the designability of structures is sensitive to the value of non-additive part of the potential. We have published a brief report of exact enumeration results of this problem [15], which are mentioned in sections 2, 3 and 5. The additive form of potential enables us to solve some parts of problem analyticaly. Due to this form of potential and the simple geometry of the model, in section 4 we will show that by combinatorial approach we can find the number and patterns of sequences which are folded to a certian compact configuration uniquely. It also shows that the sequences which fold to highly designable structure are sensitive to mutation of certain sites. Following this calculation we give an interval for critical value of non-additive part of potential in section 6.

2. The model

We consider a H–P lattice model [9]. In this model only non-sequential nearest neighbours interact. Because the native structures of proteins are compact with the H-type monomers sitting in the core, the effective potentials which are usually used, all of the forces are attractive (negative values for potential) and the strength of the force between H–H monomers is greater than others. We can write the general form of the potential in an arbitrary energy scale as:

$$E_{PP} = 0 \qquad E_{HP} = -1 \qquad E_{HH} = -2 - \gamma. \tag{1}$$

Where γ gives the energy change due to the mixing of two types of amino acids [16]. The most usual choice of H–P model potential corresponds to the limit $\gamma \gg 1$ [7–9, 12], however, physical arguments are consistent with a smaller value for γ, for instance $\gamma = 0.3$ was used by Li *et al* [10]. They have calculated the energy of all 2^{27} sequences in 103 346 compact configurations for a 27-sites 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 is zero [11, 16].

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* [10], we consider only compact structures of sequences with length 27, occupying all the sites of a $3 \times 3 \times 3$ cube [17]. There are 103 346 compact configurations which are not related to each other by rotation and reflection symmetries. Let us call the set of all compact structures, the structure space.

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}. Let us call the set of $|\sigma\rangle$, the sequence space.

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 non-sequential neighbours of the ith monomer, or by introducing the neighbourhood vector $|G\rangle$,

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

The vector $|G\rangle$ has 27 components and at its ith component, it has the number of neighbours of the ith monomer in the structure. Due to the shape of $|G\rangle$ the type of neighbours is not relevant and all we have to do is count the non-sequential neighbours. This gives us an additional symmetry for the energy that is different from spatial symmetries. For example any of the sites in a two-dimensional 5×5 square for two spatial configurations which are shown in figures 1(*a*) and (*b*), have equal number of neighbours, but the labels of their neighbours are not the same. Visualization of the same effect in three dimensions is a bit harder, but it does exist.

The space of all three-dimensional structures has 103 346 members for all compact fulfilled structures in a $3 \times 3 \times 3$ cube. Due to this additional symmetry this space 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 2 shows that the frequency of large values of N_d, is low. Interestingly there are many $|G\rangle$'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 the space of $|G\rangle$'s. One can see the distribution of a number of ground-state degeneracies g, for all 2^{27} sequences in figure 3. There are only a few sequences in which their energy is minimized in a $|G\rangle$ uniquely, and this corresponds to the 8.47% of sequences at $g = 1$. If the energy of one sequence is minimized in a $|G\rangle$ with N_d greater than 1 it has degenerate ground state. According to the definition of designability, such sequences should not be considered. The distribution of N_s is presented for $\gamma = 0$ in figure 4. Comparing this figure with figure 2 of Li *et al* [10], we observe that there is no similarity. This suggests that designability (N_s), is sensitive to the value of γ, which is $\gamma = 0$ in our work, whereas Li *et al* chose $\gamma = 0.3$. However as we shall see later, the fact that at $\gamma = 0$, we have the additive potential plays an important role. In fact a small value of γ radically changes the picture.
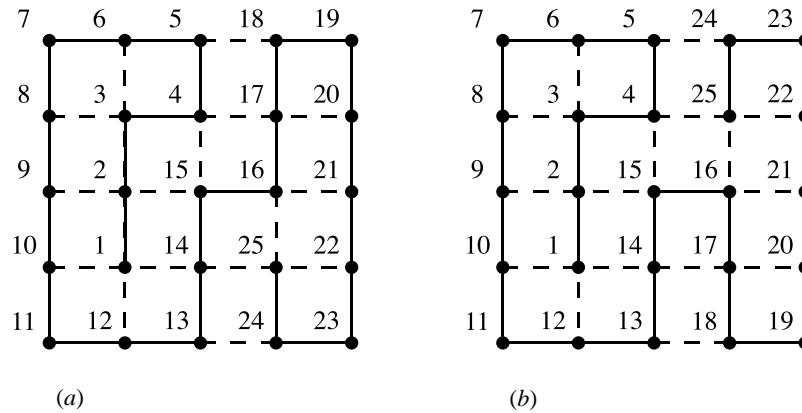
Figure 1. The number of neighbours for corresponding sites in these two configurations are the same, but the neighbours are not, for instance, site 18 in (*a*) is the neighbour of 5, but they are not adjacent in (*b*).
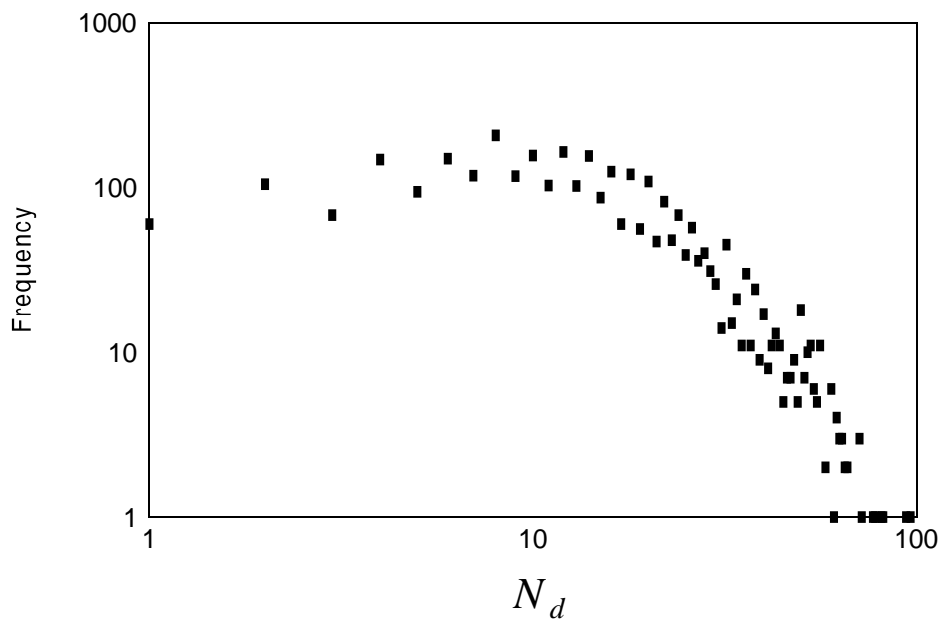

Figure 2. Histogram of N_d for members of structure space. It is interesting that there are some G sets with $N_d = 1$.

If we consider all of the sequences which have non-degenerate ground state in space of $|G\rangle$'s, 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 the designability of $|G\rangle$'s, by N'_s. Figure 5 shows the distribution of N'_s. Many of points in figure 5 are related to some $|G\rangle$'s with $N_d \neq 1$. We shall use this picture to express the nature of the energy gap in the case $\gamma \neq 0$ in section 5.
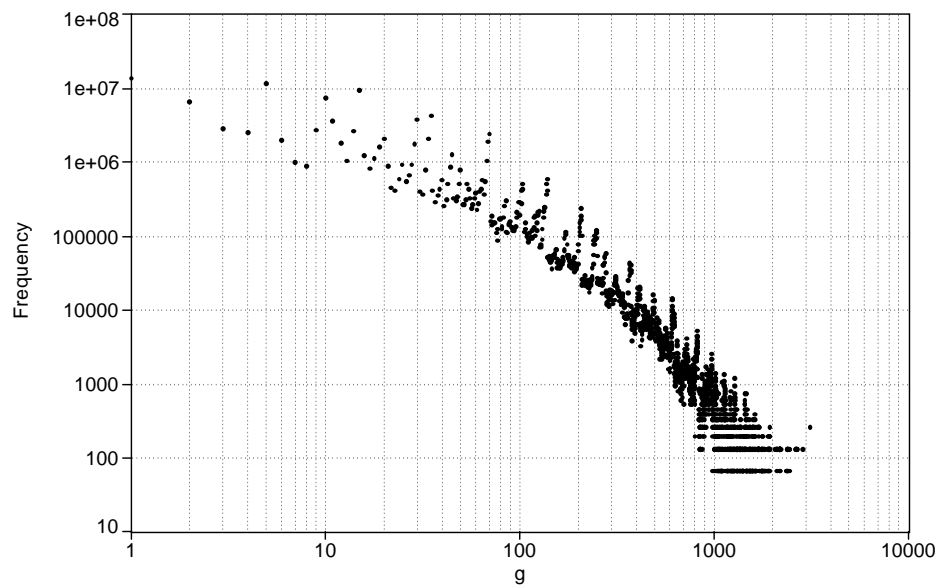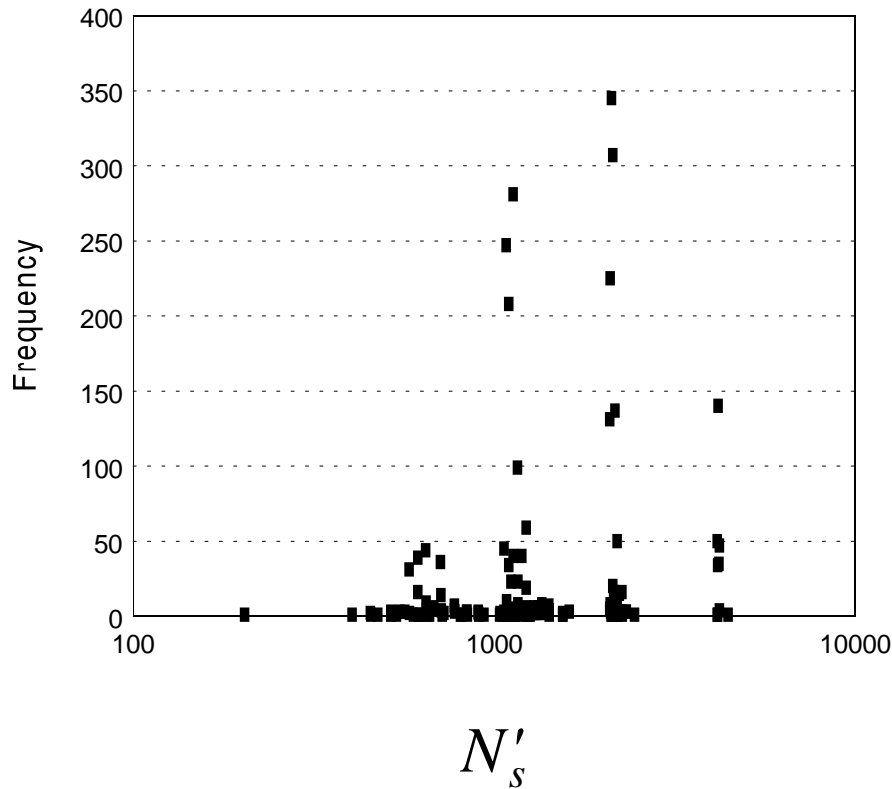
Figure 3. Histogram of degeneracy of ground state. The sequences which have non-degenerate ground state, correspond only to $g = 1$, in this diagram.

$$N_s$$
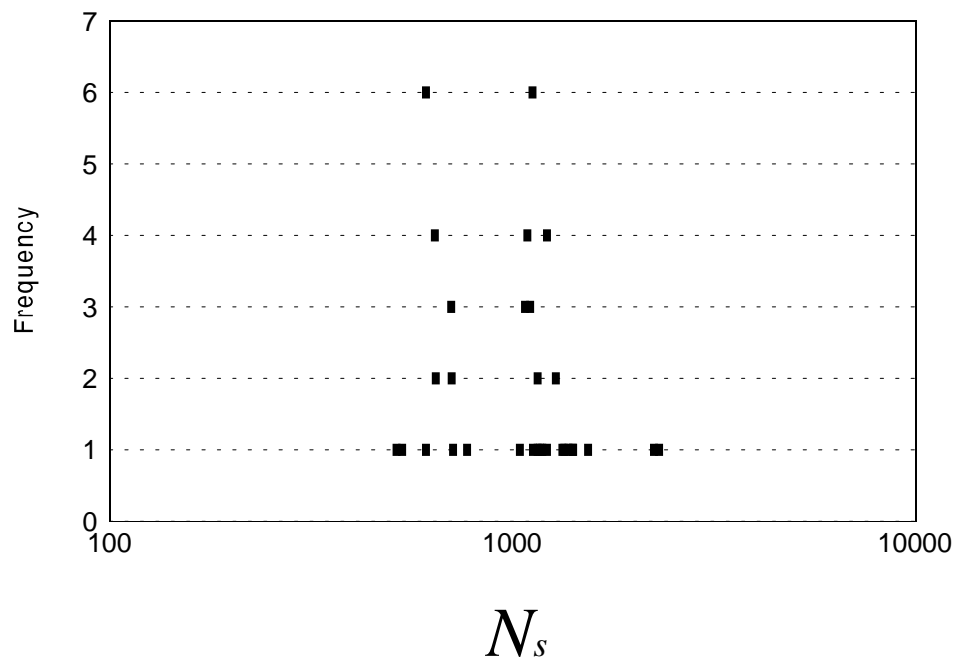
Figure 4. Histogram of N_s for additive potential.

In our enumeration we have calculated the energy of any sequence in all 6291 $|G\rangle$'s, but in figure 5 we show the results for 3153 $|G\rangle$'s which are not related to each other

Figure 5. 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.

by reverse labelling. We cannot reduce the structure space according to this symmetry before enumeration. Reverse labelling for a non-symmetric sequence gives two different configurations which may have different energies.

3. Energy levels

The number of non-sequential neighbours is related to type of site. A $3 \times 3 \times 3$ cube has eight corner sites (C), 12 link sites (L), six face sites (F), and one centre site (O) (figure 6). C sites have three neighbours, where two of them are connected by sequential links and there is only one non-sequential neighbour. Similarly L, F and O sites have 2, 3 and 4 non-sequential neighbours respectively. We must add 1 to these numbers for two ends of chain. This 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 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 1st and 27th components which are like even components). Thus,

$$|G\rangle = |g_1, \ldots, g_{27}\rangle \tag{6}$$
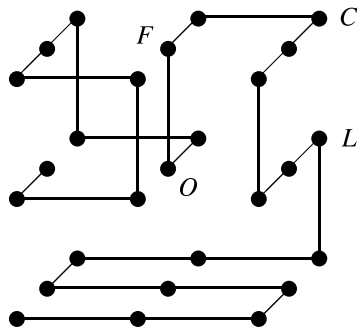
Figure 6. A $3 \times 3 \times 3$ cube has eight corner sites, 12 link sites, six face sites and one centre site.

where,

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

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

$$\begin{aligned} E_{\alpha\mu} &= \langle \sigma_\alpha | G_\mu \rangle \\ &= \sum_{i \in \text{odd}} (g_{\mu i} - 1)\sigma_{\alpha i} + \sum_{i \in \text{even}} (g_{\mu i} - 2)\sigma_{\alpha i} + \sum_{i \in \text{odd}} \sigma_{\alpha i} + 2 \sum_{i \in \text{even}} \sigma_{\alpha i}. \end{aligned} \tag{8}$$

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 last two terms in equation (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 equation (9) is an integer times two, thus it results in a ladder energy spectrum with gaps of 2. Therefore the energy gap for all of structures is the same, and there is no difference between low- and high-designable structures. Of course it is possible that energy gap would be a muliple of 2, but our enumeration shows that in this model it does not occur. However, the reason might be seen by the combinatorial approach which will be introduced in the next section, but we have not a proof for larger polymers.

4. Combinatorial approach

Our aim is to find the N'_s for any spatial configuration, determined by a vector $|G\rangle$. Because $|X\rangle$ has a simpler structure, than $|G\rangle$, we shall use $|X\rangle$ instead of $|G\rangle$. Any vector $|X\rangle$, has seven 1's and twenty 0's. One of the 1's is in the even sites, and the others are in odd sites. Energy could be calculated by performing a 'logical and' of two binary numbers ($|\sigma\rangle$ and $|X\rangle$). For example, a typical $|G\rangle$ is,

4	1	3	3	1	1	1	1	3	1	3	1	3	2
	2	2	4	2	2	2	2	2	2	2	2	2	2

$G.$

To recognize odd and even components of the vectors, we show them in the above form, writing the even sites below. The vector $|X\rangle$ corresponding to the above $|G\rangle$ is,

1	0	1	1	0	0	0	0	1	0	1	0	1	0
	0	0	1	0	0	0	0	0	0	0	0	0	

$X.$

On the other hand $|\sigma\rangle$'s have a similar form:

H	0	H	H	H	0	0	H	0	H	H	0	H	H
	0	H	0	H	0	0	H	H	H	0	0	0	H

$\sigma.$

Where we show P monomers by numerical equivalence of them. Recall that numeric equivalence for H monomers is -1. Energy of any sequence in any spatial configuration is calculated by inner product of its $|\sigma\rangle$ to corresponding $|X\rangle$. For the above $|\sigma\rangle$ and $|X\rangle$ the energy is 5H. This value is related to exact value of energy according to equation (9) by a factor of 2 and two sequence-dependent additional terms, since we are interested in the ground state and the energy gap of a sequence, the sequence-dependent term may be ignored, as the structure determines these quantities alone.

By construction any $|X\rangle$ has six 1's in odd sites, and one 1 in even sites. If we do not consider any other constraint for $|X\rangle$, we obtain an upper limit for number of $|X\rangle$'s.

$$n = \binom{14}{6} \times \binom{13}{1} = 39\,039. \tag{11}$$

This is far larger than the number of possible $|X\rangle$'s which we have obtained by enumeration, that is 6291. The fact that all $39\,039$ possible configuration do not exist points to extra constraints which are yet to be discussed. If all $39\,039$ of $|X\rangle$'s were to exist each of them would have to be a unique ground state of only one sequence, thus removing all interest! To see this, it is enough to insert an H into $|X\rangle$ wherever one finds a 1, and P for zeros. Indeed the absence of some of these vectors in the real world makes some of the other more preferable in nature.

The connectivity of a self-avoiding walk, further constraints the $|X\rangle$. For example to pass through the centre site, the walk has to pass through two face sites. This means that only 1 (corresponding to the centre site) in even sites must be sandwiched between two 1's in odd sites (face sites). This constraint reduces the number of possible $|X\rangle$'s. Two 1's in odd sites are fixed by even 1, and only 12 sites remain for four other 1's. Then there are,

$$n = \binom{12}{4} \times \binom{13}{1} = 6435 \tag{12}$$

vectors. This number is still larger than exact number of $|X\rangle$'s by 144. Although due to our enumeration we know these 144 vectors, we cannot find the complex constraints which prune them out, and we shall continue our calculation as though these 144 vectors were correct. Of course the values are different from exact enumeration, however it can be seen that this difference is not too large, and it may be considered as an approximation to the exact solution. Also we aid a computer enumeration including the extra 144 vectors and compared the results with the combinatorial calculation. This has served as a check on our code.

We now proceed to calculate N'_s for the following example:

1	0	1	1	1	0	0	0	1	0	0	0	1	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0

X_0.

First let us introduce some new parameters and notations. We will show the energy of a $|\sigma\rangle$ in an $|X\rangle$ as:

$$E = E(a, b, c) = (a + b + c)H \tag{13}$$

where a, b and c are related to energy parts which come from centre (1 in lower row), faces which are connected to centre, and energy of other parts, respectively. For example energy of following $|\sigma\rangle$:

H	0	H	H	0	H	0	0	H	H	0	0	H	H
0	H	0	0	H	H	H	0	H	0	0	H	0	

σ

in $|X_0\rangle$, is $E(0, 2, 3) = 5H$.

Besides, we name the number of pairs of 1's in the upper row of $|X\rangle$ as z and the number of 1's in two ends of vectors as y. For $|X_0\rangle$, $z = 2$, and $y = 1$.

Now we try to count the number of all polymers which have their energy minimized in $|X_0\rangle$ and, there is no other $|X\rangle$ with energy equal to the ground state for them. To do this we discuss all possible cases.

4.1. Case (i): E(1,2,4)

Such polymers have at least seven H sites corresponding to 1's of X_0. These polymers have minimum possible energy, thus X_0 is a minimum energy configuration for them. However, it must be checked whether it is a unique ground state or not. First consider polymers which in addition to these seven H's have another H monomer in their upper row sites,

H	0	H	H	H	0	0	0	H	0	H	0	H	0
0	0	H	0	0	0	0	0	0	0	0	0	0	

σ_1.

The energy of this sequence in following $|X\rangle$ is also 7H.

1	0	1	1	1	0	0	0	1	0	1	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0

X_1.

Then the ground state of polymers which have additional H monomers in corresponding to upper row 0's of X_0, is degenerate, and they do not count in N'_s of X_0. The above discussion is independent of value of a and b in $E(a, b, 4)$, and degrees of freedom to choose sites for H monomers is limited to lower row sites.

For the $|X\rangle$ with $z \neq 1$ (like X_0) polymers cannot have H monomers in the lower sites between two upper row 1's. For example, the following sequence,

H	0	H	H	H	0	0	0	H	0	0	0	H	0
	0	0	H	H	0	0	0	0	0	0	0	0	0

σ_2

has energy 7H in the following $|X\rangle$ too

1	0	1	1	1	0	0	0	1	0	0	0	1	0
	0	0	0	1	0	0	0	0	0	0	0	0	0

X_2.

Then the contribution of polymers with $E(1, 2, 4)$ in N'_s is:

$$N'_s(\mathbf{i}) = 2^{12-(z-1)} = 2^{13-z}. \tag{14}$$

4.2. Case (ii): E(0,2,4)

In this case if $z > 1$ (such as X_0) the ground state is degenerate. It can be seen that any sequence with energy $E = (0, 2, 4)$ in X_0 state has the same energy in X_2 state. In the case $z = 1$, only the sites in lower row by condition that they are not a neighbour of corresponding upper 1's of X, have freedom to be an H or P monomer. There are $2 \times 6 - z - y$ sites which do not have this freedom in the lower row. Then,

$$N'_s(\mathbf{ii}) = \begin{cases} 2^{2+y} & z = 1 \\ 0 & z > 1. \end{cases} \tag{15}$$

4.3. Case (iii): E(1,0,4)

In this case there is only one sequence with a non-degenerate ground state. For our example, X_0, this sequence is,

H	0	0	0	H	0	0	0	H	0	0	0	H	0
	0	0	H	0	0	0	0	0	0	0	0	0	0

σ_3.

In the above sequence changing any P monomer to H type, will cause the ground state to becomes degenerate. Then,

$$N'_s(\mathbf{iii}) = 1. \tag{16}$$

4.4. Case (iv): E(1,1,4)

For this case b is 1, and if this 1 comes from right or left neighbour of lower 1, it has different solutions. Then we introduce new parameters (z_R, y_R) and (z_L, y_L), which are similar to old z and y, when right or left neighbour 1 of lower 1 will be omitted. For X_0 we have $z_R = 0$, $z_L = 1$ and $y_R = y_L = y = 1$. By introducing this new parameters this case is very similar to case (**ii**), and the difference comes from number of corresponding 1's in upper row (five instead six), and no restriction in value of z. Then,

$$N'_s(\mathbf{iv}) = 2^{3+z_R+y_R} + 2^{3+z_L+y_L}. \tag{17}$$

4.5. Case (v): Other cases

All of the other cases for ground-state energy are degenerate, and need not be considered.

With this analysis it is possible to find N'_s for any $|X\rangle$. For our X_0 example it is,

$$N'_s(\mathbf{i}) = 2^{11}$$
$$N'_s(\mathbf{ii}) = 0$$
$$N'_s(\mathbf{iii}) = 1$$
$$N'_s(\mathbf{iv}) = 2^4 + 2^5$$

that gives,

$$N'_s(X_0) = 2097.$$

By this calculation all of the values of N'_s's can be calculated. In addition to the number of sequences which have a unique ground state, in this way monomer coding of these sequences is known. Had the 144 additional structures been taken out, the calculation of N'_s for the problem would correspond to enumeration exactly. However, taking these structures out is too complex and would have to be done case by case. Besides the value of N'_s, this calculation shows that the sequences with non-degenerate ground state have between 4 and 6 H-type monomers in face sites and none in corner sites. Indeed in our model the stability of polymers is very sensitive to mutation in corner sites.

5. Non-additive potentials

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

$$E_{\alpha\mu} = \langle \sigma_\alpha | G_\mu \rangle - \tfrac{1}{2}\gamma \langle \sigma_\alpha | M_\mu | \sigma_\alpha \rangle \tag{18}$$

where σ and G are the sequence and neighbourhood vectors, that introduced in previous sections. M is the adjacency matrix for this configuration

$$M_{ij} = \begin{cases} 1 & \text{if the } i\text{th and } j\text{th monomers are adjacent} \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

Any $|G\rangle$ has N_d different M-matrices. The first term in equation (18) was calculated in the case $\gamma = 0$, and we only need to calculate the last part. The aim of our calculation is to find the ground state. In any compact configuration in a $3 \times 3 \times 3$ cube, there are 28 non-sequential neighbour pairs. Thus, the contribution of the last term in energy is less than 28γ. We have shown that energy spectrum for the previous case has a ladder structure with energy gaps equal to 2. In this case these split into some sublevels (figure 7). Then if we choose $\gamma < \frac{2}{28}$ the levels are separate. Of course this is a lower estimation for γ_c. In the next section we will obtain a better estimate for lower and upper limits of the critical value of γ.

From the result of the additive potential we have a subset G in the space of all the spatial structures which gives the minimum energy to folding. This G subset has N_d members which all of them 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 of 103 346 spatial structures for any sequence, except for sequences which 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. The first excited state of these sequences are in another G subset. Thus, to find the energy gap for them the program
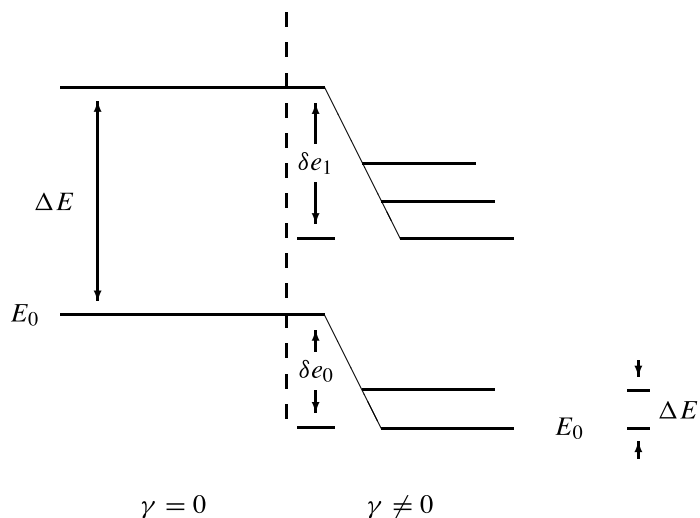
Figure 7. Energy levels of additive potential split to sublevels for non-additive potential.
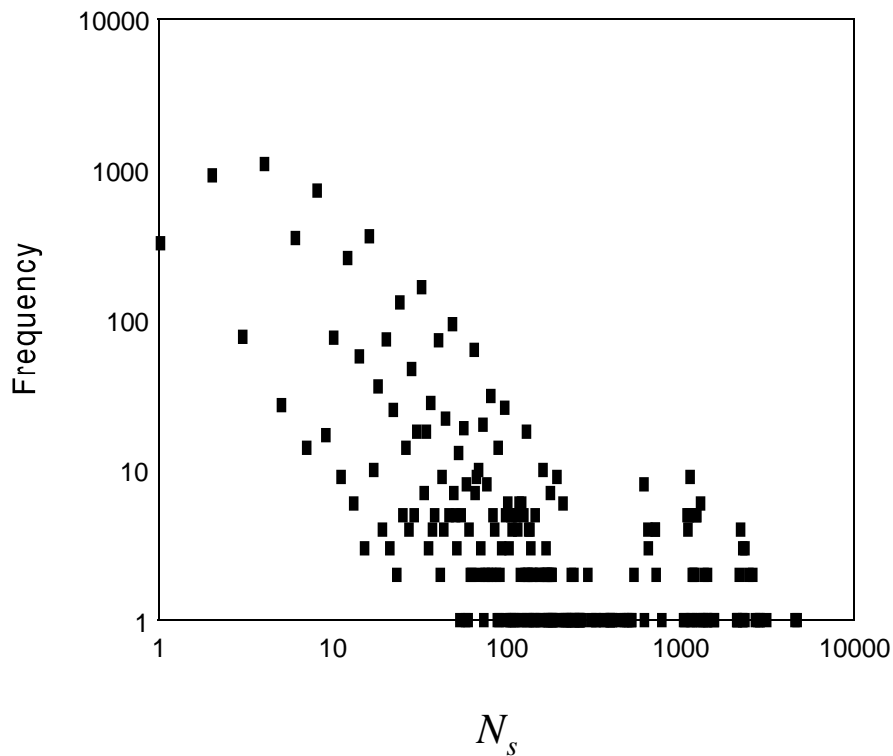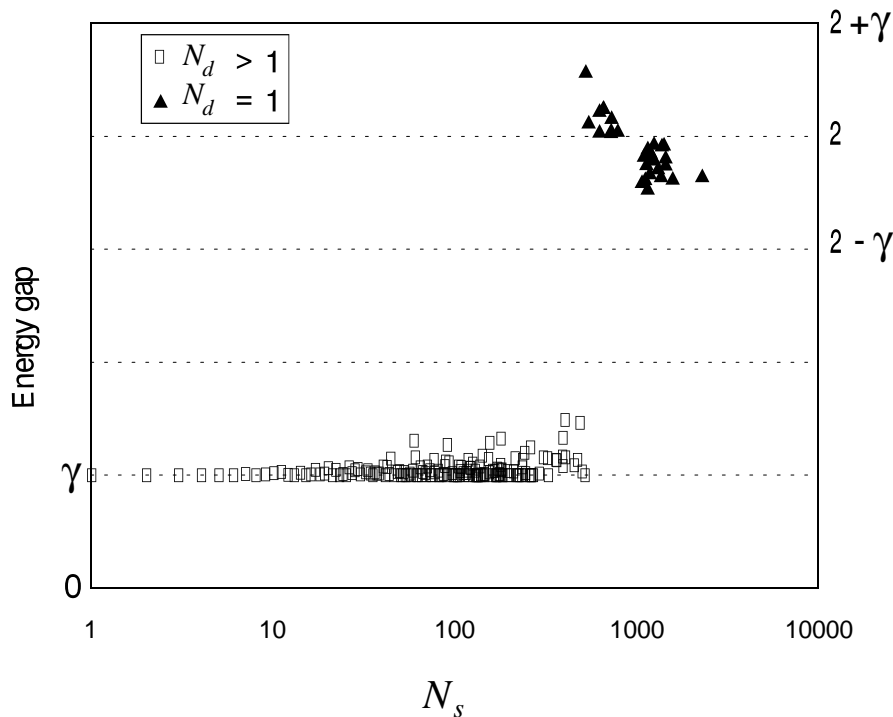

Figure 8. Histogram of N_s for non-additive potential.

must be run over all of the 103 346 structures. We have calculated this energy spectrum, and have found the new N_s for all 103 346 structures. We show the results for the 51 704 configuration which are unrelated by reverse labelling symmetry in figure 8. We have found

Figure 9. The mean of energy gap versus N_s. There is a jump in energy gap for highly designable structures. All of these highly designable structures have $N_d = 1$.

the energy gap for first excited state for all sequences. Figure 9 shows the diagram of mean of energy gap versus N_s. This figure shows that highly designable structures which are related to G subsets with one member.

In this enumeration we have calculated the energy spectrum for all of the sequences which have non-degenerate ground state for the additive potential. We removed some of the sequences because of the degeneracy of the ground state in the additive potential case. It is possible that this degeneracy will be removed by the non-additive part of the potential, and some of the sequences have a unique ground state for non-additive potential. However, the energy gap for these sequences is of order of γ, and if we consider them it causes a shift in the horizontal axes to bigger N_s and brings the points down nearer to the γ value in vertical direction in figure 9. These make this figure more similar to results of Li *et al* [10]. In their work the energy gap for low-designable structures are of order of γ (they chose $\gamma = 0.3$) also.

6. Estimation of γ_c

The energy levels for the additive potential have a ladder structure, as it had been proven in previous sections. The energy gaps between the levels is 2 in our arbitrary energy unit.

In the case of $\gamma \neq 0$ the energy has two parts (equation (18)). The first part comes from the additive part of potential and does not change. The second part comes from the non-additive part of potential, and is equal to the number of H–H non-sequential neighbours

in spatial configuration. Because of this non-additive part, the energy spectrum is changed, and a level is split into some sublevels (figure 7).

If the contribution of the second part to energy is less than 2, for all structures, the ground state and excited state of any polymer is between $|G\rangle$ partners of its ground state for additive potential, except for $|G\rangle$ with $N_d = 1$, where there is only ground state.

Let δe_0 be the difference in the ground-state energies of additive and non-additive potentials, and δe_1 be difference energy of first excited state in the case of $\gamma = 0$ with a minimum of new energies for the sequence in the structures corresponding to these excited states (there is no uniqueness constraint for excited states). If $\delta e_0 - \delta e_1 < 2$ the ground state does not change and the values of N_s for structures that we presented in the previous section do not change. By increasing γ, the absolute values of δe_0 and δe_1 increase.

To find the difference between δe_0 and δe_1 one has to calculate the difference in H–H contacts in ground-state structure and maximum of H–H contacts in excited level structures.

This difference has two sources. Because the energy levels in the case $\gamma = 0$ are separated by 2, then the difference between them comes from replacing a H monomer from O site to an L site, or from an F site to a C site. Both of them cause the energy to increase by 2. But it is possible that these replacing decrease the energy by 2γ. For example, consider that one F site with no H neighbour will go to one C site with two non-sequential H neighbours (this monomer must be an end residue), then this gives an upper limit for γ_c, which is 1.

The other source for increasing the H–H contacts comes from replacing H monomers in L and F sites by the same type sites. These changes are only relevant in the case $\gamma \neq 0$. The maximum increase in H–H contacts due to this replacements is 6γ, related to the sequences which have four H monomers in the F sites and five to seven in L sites in their ground-state structures. We must add to this the most probable maximum of energy decrease, corresponding to the change of site-type of one H monomer, which we have shown to be equal to 2 (the first source). This gives the most probable change of energy due to the non-additive part of the potential, $(2 + 6)\gamma$. Thus, the lower limit for γ_c is $\frac{2}{2+6} = 0.25$. Therefore, we have:

$$0.25 < \gamma_c < 1. \tag{20}$$

However, this calculation is highly dependent on the geometry, but designability is a geometry-dependent parameter itself [18]. Although we expect γ_c to be N dependent, its form is still uncertain to us. The upper limit for γ_c comes from the odd–even problem in the cubic lattice and is N independent, but the lower limit can depend on N. The existence of a non-zero value for 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.

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

The fact that the ground-state configuration for a certain sequence does not depend on the value of γ for $\gamma < \gamma_c$, suggests that it can be used as a control parameter for energy gap in Monte Carlo simulation. This shows that by using the additive potential one can reduce the folding time of highly designable proteins in such simulations.

Acknowledgments

References

[1] Stryer L 1988 *Biochemistry* (San Francisco, CA: Freeman)
Stryer L 1992 *Protein Folding* ed T E Creighton (New York: Freeman)
[2] Anfinsen C B, Haber E, Sela M and White F H 1961 *Proc. Natl. Acad. Sci., USA* **47** 1309
[3] Anfinsen C B 1973 *Science* **181** 223
[4] Dill K A, Bromberg S, Yue K, Fiebig K M, Yee K P, Thomas P D and Chan H S 1995 *Protein Sci.* **4** 561
[5] Garel T, Orland H and Pitard E 1997 Protein folding and Heteropolymers *Spin Glasses and Random Fields* ed A P Young (Singapore: World Scientific)
[6] Levinthal C 1968 *J. Chem. Phys.* **65** 44
[7] Lau H F and Dill K A 1990 *Proc. Natl. Acad. Sci., USA* **87** 638
[8] Chan H S and Dill K A 1991 *J. Chem. Phys.* **95** 3775
[9] Chan H S and Dill K A 1989 *J. Chem. Phys.* **90** 492
Chan H S, Dill K A and Shottle D 1992 Statistical mechanics and protein folding *Prinston Lectures on Biophysics* ed W Bialek (Singapore: World Scientific)
[10] Li H, Helling R, Tang C and Wingreen N 1996 *Science* **273** 666
[11] Li H, Tang C and Wingreen N 1997 *Phys. Rev. Lett.* **79** 765
[12] Dill K A 1990 *Biochemistry* **29** 7133
Chan H S and Dill K A 1990 *Proc. Natl. Acad. Sci., USA* **87** 6388
[13] Pande V S, Grosberg A Yu and Tanaka T 1995 *J. Chem. Phys.* **103** 9482
[14] Kardar M 1996 *Science* **273** 610
[15] Ejtehadi M R, Hamedani N, Seyed-Allaei H, Shahrezaei V and Yahyanejad M 1998 *Phys. Rev.* E **57** 3298
[16] Hildebrand J H and Scott R L 1950 *The Solubility of Nonelectrolytes* (New York: Reinhold)
[17] Shakhnovich E and Gutin A 1990 *J. Chem. Phys.* **93** 5967
[18] Irback A and Sandelin E 1998 *J. Chem. Phys.* **108** 2245