Energy Profile of the Space of Model Protein Sequences

G. TIANA^{1,2}, R.A. BROGLIA^{1,2} and E.I. SHAKHNOVICH³

¹*The Niels Bohr Institute, University of Copenhagen, Blegdamsvej 16, 2100 Copenhagen, Denmark*

²*Dipartimento di Fisica, Università di Milano and INFN, sezione di Milano, via Celoria 16, 20133 Milano, Italy*

³*Department of Chemistry, Harvard University, 16 Oxford st., Cambridge, MA, USA*

Abstract. A numerical study of the energy landscape of the space of model protein sequences is carried out. As a consequence of the heterogeneity of the contact energies among amino acids, the energy landscape displays a very rough profile, a behaviour typical of frustrated systems. This gives rise to a hierarchical clustering of low-energy sequences and can have evolutionary consequences.

Key words: energy landscapes, model protein, space of sequences

1. Introduction

Native conformations of proteins are highly degenerate, in the sense that several sequences are known to target to each natural fold [1]. Such sequences can have very little similarity, like in the case of analogs, which are sequences folding to the same three dimensional conformation and having less than 20% similarity.

In order to be biologically relevant, protein sequences need to encode a unique, stable and kinetically reachable native conformation. A necessary and sufficient condition to satisfy these requirements has been shown to be a low value of the energy of the sequence in the native conformation, with respect to the energies of the unfolded states, or, more accurately, to the lowest energy of the structurally dissimilar compact conformations [2]. Consequently, it is interesting to study the energy landscape of sequences in a given, fixed native conformation. Due to the low native energy of good folder sequences, the analysis of the lowest part of the energy landscape gives informations about the properties of sequences folding to a given three dimensional structure, such as their similarity and the conservation of some residues along the sequence.

Making use of a simple lattice model [3–5] which takes into account the polymeric character of proteins and the heterogeneity of the interactions among different residues, and which has been shown to describe well a number of features of real proteins [6], we shall show that the energy landscape in sequence space is remarkably rough and is organized in clusters and super-clusters [7] (cf. Figure 1).

Figure 1. A sketch of the energy landscape of the space of sequences. Its roughness gives rise, at low energy ($\approx E_{min}$), to clusters (dark grey) and super-clusters (light grey). At higher temperature ($\sim E_c$) the division in clusters disappears, leaving only the super-clusters. Barriers among clusters are lower than E_c , while barriers between different superclusters exceed E_c .

Defining the distance between two sequences as the ratio between non-matching residues and the total number of monomers, we have found that best folders (those with the lowest energy in the common, native conformation) cluster together in a hierarchical way (cf. horizontal planes in Figure 1).

In other words, each low-energy sequence is surrounded by other very similar low-energy sequences, giving rise to a 'cluster'. Each cluster is surrounded by other clusters, the distance among clusters being approximately 0.5 (meaning that half of the residues of a representative sequence of a cluster match the residues of a representative sequence of another cluster), giving rise to 'superclusters'. Sequences belonging to different superclusters have very little similarity (≈ 10 – 20% of residues match). Some residues (approximately 5 – 10% of the total number of residues of the chains) are 100% conserved within a supercluster and are partially conserved across superclusters. These few 'hot' sites¹ have been shown to be momentous in determining the thermodynamical and kinetical properties of model proteins [8, 9].

The roughness of the energy landscape in the space of sequences of model proteins resembles that of other disordered physical systems, such as spin glasses [10], having a similar origin [11, 12]: frustration². In the case of spin glasses, that associated with the disordered interaction acting among spins. For model proteins that caused by the heterogeneity of contact interactions. In fact, as discussed below, the use of different heterogeneous interaction matrices (e.g. MJ [14], Skolnick [15], random 20 letters) produces the same kind of rough energy landscape and hierarchical organization of low-energy sequences, while using simpler interaction matrices (e.g., HP-model, whose space of sequence is not frustrated [18]) gives a funnel-like energetic profile.

¹ The definition of 'hot' sites in the protein is taken from ref. [8] and indicates those sites which are highly sensitive to point mutations.

² I.e., the inability of a system to satisfy all interactions at the same time in any state [13].

2. Methods and Results

The model employed to study the energy landscape of sequences pictures each protein as a chain of beads seating on a cubic lattice [2–4]. Each residue (i.e. each bead) interacts with its nearest neighbours, except with those which are consecutive along the chain, and contributes to the total energy with a term which depends on the kind of residues involved through a given interaction matrix describing the contact energies among the different amino acids. It has been shown that, provided the interaction matrix is enough heterogeneous³ (e.g. MJ [14], Skolnick [15], random 20 letters), the model describes realistically the essential thermodynamical and kinetical features of protein folding [16].

The use of lattice models in the analysis of designed sequences, in particular in the study of the energy landscape of the space of sequences has a number of attractive properties, all of them associated with the simplicity of the model. In particular, it is possible to generate and to design as many sequences as one needs, all of them fulfilling the condition of being good folders, by varying the selective temperature T_s , of the corresponding simulations. Furthermore, it is possible to study the dynamical properties of each of these sequences with excellent statistics, by conveniently adjusting the temperature (in conformation space⁴) so as to ensure a fast ($\approx 10^6$ MC steps) folding time and at the same time a nonvanishing population of the native conformation ($>10\%$).

To analyze the space of sequences associated with a given conformation, we will use a canonical ensemble approach, by comparing sequences associated with different selective temperatures $T_s \equiv (\partial S / \partial E)^{-1}$. Naively, this means studying ‘slices’ of the space of sequences characterized by different average energies. We show that sequences folding to the same native conformation cluster together in a hierarchical fashion. This cluster organization is rather articulated at the bottom of the energy landscape (low T_s) becoming simpler as the energy gets larger (higher values of T_s). The threshold energy⁵ E_c , dependent only on the interaction matrix, on the composition of the chain and on its length, has been shown [2] to separate good folder sequences (those with $E < E_c$) from bad folders (those with $E > E_c$)⁶. Consequently, there is a temperature $T_c \equiv (\partial S / \partial E |_{E_c})^{-1}$ such that sequences selected at temperatures $T_s < T_s^c$ are in average good folders while those selected at higher temperatures are not.

³ That is, its standard deviation is large compared to the absolute value of its average.

⁴ Note that when comparing T with T_s one has to keep in mind that the number of dimensions of the space of sequences is larger than that of the space of conformations.

⁵ That is the lowest energy of the conformations structurally dissimilar to the native conformations.

⁶ In a random energy approximation [17] E_c can be written as $E_c = NB_0 - N\sigma_B(2\log\gamma)^{1/2}$, where B_0 and σ_B are the average and standard deviation of the interaction matrix weighted on the amino acid composition, N is the number of contacts and γ is the effective coordination number of the lattice. Typical values of γ for compact conformations lie in the range 1.8–2.2.

The main tools used here to describe the space of sequences are the parameters (connected to the distance between sequences)

$$q_s^{\alpha\beta} = \frac{1}{N} \sum_{i=1}^N \delta(\sigma_i^\alpha, \sigma_i^\beta), \quad (1)$$

which specifies the relative number of matching residues between any two sequences σ^α and σ^β , and its distribution

$$P(q_s) = \langle \delta(q_s - q_s^{\alpha\beta}) \rangle, \quad (2)$$

where $\langle \rangle$ indicates the thermal average, calculated from all possible pairs of sequences σ^α and σ^β . The parameter q_s serves as a metric in sequence space and its distribution (like the distribution of magnetization in ferromagnetic systems, or the distribution of the replica order parameter in spin glasses) contains information about the distances between the states of the system populated at a given temperature. In the simplest physical systems, characterized by a single free energy minimum, the distribution displays a single peak centered around 1. If, increasing the temperature, the peak gets broadened in a regular fashion, one deduces that the energy landscape is funnel-like⁷.

To study the properties of the distribution $P(q_s)$ we have chosen the 48-mer configuration shown in the inset to Figure 2 and used the Metropolis algorithm [19] at a fixed temperature T_s to generate the sequences in order to find $P(q_s)$. That is, we start from a random sequence with a composition corresponding to the average composition for natural proteins [1]. Basic moves in the space of sequences are swaps between couples of residues in the chain, which keep the composition of the chain constant. After each move, the new energy is calculated on the target structure and the move is accepted or rejected according to the Metropolis criterion. During the calculation, 1000 sequences separated by 1000 MC steps (sampling rate) have been recorded and from these results the distribution $P(q_s)$ has been calculated. For very low values of T_s , calculations with different starting sequences have been carried out to check the efficiency of the Metropolis algorithm. The ‘selective’ temperature T_s can be interpreted as the degree of evolutionary pressure towards sequences having larger energy gap. This means that at low temperatures, only sequences with high stability and, consequently, short folding times are selected [9, 20].

We start by analyzing the very bottom of the energy landscape (sequences with the largest gap δ), which corresponds to the states associated with $T_s = 0$. Since the Metropolis algorithm fails at this temperature, we had to use an *escamotage*.

⁷ In a ferromagnetic system in the low-temperature phase, the magnetization (which is its order parameter) displays two symmetric peaks, at +1 and -1, indicating two equivalent ground states (all spins up and all spins down, respectively). In the case of spin glasses the distribution of replica order parameter can have, below the glass temperature, more complicated shapes, due to the manifold of low energy states, related by non-trivial symmetries [10].

Figure 2. The probability distribution $P(q_{\alpha\beta})$, calculated over 1000 sequences generated by the Metropolis algorithm at $T_s = 0.01$ and with energy $E < -25.50$. The insets show the target conformation, highlighting in black the two possible sets of 'hot' sites [8].

Namely, the calculation is carried out at the lowest temperature at which the Metropolis algorithm is reliable ($T_s = 0.01$, to be compared to the standard deviation of the interaction matrix $\sigma_B = 0.30$), and a threshold is set in such a way that all sequences with energies higher than the threshold energy are not recorded. For the conformation displayed in the inset to Figure 2, the optimal sequence found has an energy $E_{opt} = -26.97$, while $E_c = -21.5$. The value of the threshold, which we have set at -25.50 , is somehow arbitrary. We have checked that the outcome does not depend qualitatively on this choice.

The distribution $P(q_s)$, shown in Figure 2, displays three well defined peaks⁸ at $q_s = 0.1$, $q_s = 0.55$ and $q_s = 0.95$. The peaks centered at 0.55 and 0.95 indicate that the low-energy part of sequence space can be divided into clusters, each cluster containing sequences that differ only for a few mutations. The peak close to unity arises from the comparison of sequences belonging to the same cluster. This means that, starting from a given sequence, it is possible to introduce mutations and still obtain a good folding sequence. This continuous modification can be continued until about 15% of the monomer of the chain have been swapped. If this limit is exceeded, a wider cooperative rearrangement of residues (between 40% and 60%, corresponding to the second peak in Figure 2) is needed to obtaining a low-energy sequence. Such rearrangements are associated with moves in the space of

⁸ A dependence of the position of the peaks on the sampling rate is found if the latter is $\gg 10^3$. This is a consequence of the entropic term in the free energy, which is present because the calculation is performed at a non-zero temperature. Since the meaning of this calculation is to simulate the behaviour at $T_s = 0$, we discarded the results at large sampling rates as an artifact of the calculation technique.

Figure 3. The distribution $P(q_s)$ obtained making use of an interaction matrix whose element, randomly generated, have a Gaussian distribution with the same average and standard deviation as the MJ matrix. The shape of the distribution is the same as the one displayed in Figure 2, except for the lack of the peak at $q_s \approx 1$.

sequences from one cluster to another one. A further analysis shows that, within each cluster, ‘hot’ sites are 100% conserved and, although the size of each cluster varies to a considerable extent, no sequence is isolated.

The peak at $q_s = 0.1$ in Figure 2 corresponds to a higher-level organization of clusters, giving rise to ‘super-clusters’. Clusters within the same superclusters are associated with a similarity parameter $q_s \approx 0.55$, while clusters belonging to different super-clusters display $q_s \approx 0.1$. For the conformation displayed in the inset of Figure 2 we could find two superclusters.

Within each super-cluster, ‘hot’ sites are in the same position and the type of residues occupying each of these sites is conserved in 100% of the sequences. This fact emphasizes the importance of ‘hot’ sites to govern the low-energy properties of protein sequences [8, 9]. Furthermore, half of the ‘hot’ sites of sequences belonging to the first super-cluster are in the same positions as the ‘hot’ sites of sequences of the second super-cluster and the other half move to neighboring sites. The common ‘hot’ sites occupied by residues which are very similar and again strongly interacting (e.g. exchange of D with E and Q or K with R). This result is intimately connected with the similarity the columns of the MJ matrix describing the corresponding interactions have among each other. In other words, this result suggests that the presence of the two super-clusters are likely to be due to the quasi-degeneracy of these matrix elements. The substitution of a monomer in a ‘hot’ site with a very similar one causes a small extent a rearrangement of the other ‘hot’ sites, and a much larger degree of rearrangement of the remaining residues (called ‘warm’ and ‘cold’ in ref. [8]).

To substantiate this scenario, we have repeated the calculation with a random-generated matrix having the same distribution of the elements of the MJ matrix. Because in this case all the columns of the matrix are different, we expect no super-

Figure 4. The distribution $P(q_s)$ for sequences targeting to the conformation displayed in the inset to Figure 2 and generated at $T_s = 0.05$ (solid line), $T_s = 0.08$ (dotted line), $T_s = 0.10$ (dashed line) and $T_s = 0.15$ (dot-dashed line). These distributions display a single peak at low q_s while the self-overlapping peak ($q_s \approx 1$) cannot be seen, being at a much smaller scale.

clusters. In fact, the distribution $P(q_s)$ at temperatures $T_s \gg T_s$ is, in this case, very similar to the distribution shown in Figure 2, except for the absence of the peak at $q_s \approx 0.1$ (Figure 3).

So far, we have only dealt with the lowest section of the energy landscape. By increasing the temperature T_s , it is possible to study how sequences with higher energy are distributed and whether they still group into clusters and super-clusters. The distribution $P(q_s)$ at different temperatures is displayed in Figure 4. Unlike the case at $T_s \approx 0$, these curves show a single⁹ peak, whose centroid decreases as temperature increases. The limiting case ($T_s = +\infty$) is a Gaussian centered at $q_s = 0.05$ ($= 1/20$), which corresponds to random alignments of 20 types of amino acids, each of them appearing with approximately the same probability.

While the lack of peaks at large q_s indicates that the division in clusters has disappeared, it says nothing about superclusters. A detailed analysis of the sequences generated in the range of temperatures $0 < T_s < T_s^c$ show that super-clusters are still present, each of them being characterized by a unique choice of ‘hot’ sites.

3. Discussion

The energy landscape we found associated with the space of sequences is composed of a multitude of valleys, separated by barriers of different height (cf. Figure 1). This is the typical energy landscape of frustrated systems and gives rise to a hierarchical organization of thermodynamical states [10].

⁹ The absence in Figure 4 of a peak at $q_s \approx 1$ is an entropic effect. A careful analysis shows a small $q_s \approx 1$ peak at a much smaller scale.

Figure 5. The monomer entropy per site at different temperatures.

As a consequence, the picture that emerges at $T_s \approx 0$ is that, for each good folder sequence it is possible to mutate amino acids in a few sites (except ‘hot’ sites), up to 15% of them and still obtain a good folding sequence. This process produces a cluster. If amino acids in more sites are changed, there is a partial rearrangement of amino acids in many other sites to keep the energy low, leading to different clusters. When a strongly interacting amino acid occupying a ‘hot’ site is mutated with a similarly strongly interacting residue (if the interaction matrix allows for such a possibility), the structure of the other ‘hot’ sites needs to be only slightly rearranged, while the amino acids occupying the other cold sites are heavily rearranged, the net result being a new super-cluster.

Clusters are composed of almost degenerate sequences, without barriers among them (at least, on an energy scale down to 0.01, the standard deviation of the contact energies being $\sigma_B = 0.3$). Among different clusters there are barriers characterized by a wide range of energies. While sequences belonging to the same super-cluster are connected through pathways lying below E_c , to go from a super-cluster to another one, one has to cross barriers whose top is above E_c . This means that, within a super-cluster, it is possible to go from a folding sequence to another folding sequence through a pathway of folding sequences made of single-mutation steps. On the other hand, to go from one super-cluster to another one, it is necessary to go through sequences with energy higher than E_c which, at a first approximation, do not fold.

Super-clusters, unlike clusters, are defined by residues in few specific sites, which have very little variability. This result agrees with the interpretation of model proteins as composed by a well-optimized nucleus surrounded by a low temperature state of the other, frustrated residues [9]. The nucleus undergoes a sharp transition between the high- and the low-energy states. The low energy state can be realized only in few different ways, giving rise to the few observed super-clusters.

Figure 6. The distribution of entropy, as found from Figure 5.

The topography of the space of sequences highlights once again the fundamental role of ‘hot’ sites in determining the thermodynamical properties of proteins. In Figure 5 it is shown the residue entropy per site, defined as

$$S(i) = - \sum_{\sigma=1}^{20} p_{\sigma}(i) \ln p_{\sigma}(i), \quad (3)$$

where $p_{\sigma}(i)$ is the probability that the amino acid of kind σ occupies the i th site of the chain. Of course, the entropy of a site is zero if all sequences selected at a given temperature have the same kind of amino acid in that site, while it is $\ln 20 \approx 3$ if all amino acids appear with the same probability $p_{\sigma}(i) = 1/20$. The plot displays deep minima corresponding to ‘hot’ sites. The curves associated with temperatures below or equal to 0.08 display minima with approximately the same entropy, meaning that the folding nucleus is ‘frozen’ (cf. ref. [9]). The curves obtained at $0.10 \leq T_s \leq 0.20$ display minima at the same sites as before, but the associated entropy is somewhat larger. At $T_s = 0.40$ the behaviour is apparently random.

The histogram of sequence entropy per site corresponding to these data is displayed in Figure 6. At low temperature ($T = 0.01$, solid line), the histogram shows two distinct peaks, corresponding to highly (11 monomers, among which there are the ‘hot’ sites) and poorly (37 monomers) conserved sites. At $T_s^c = 0.08$ a transition regime takes place towards a mono-modal distribution and, at still higher temperatures, the entropy approaches the equiprobability behaviour $S(i) = 3$ for each site. The bimodal distribution observed at low temperature is related to the complicated structure of the energy landscape in sequence space, which simplifies when the threshold $T_s^c = 0.08$ is overcome. Figure 7 displays the distribution of entropy for the Immunoglobulin fold, obtained from the analysis of real sequences.

Figure 7. Distribution of entropy for the Immunoglobulin fold, obtained from real sequences.

Figure 8. Distribution of $p(q_s)$ at low temperature calculated for a different structure than that of Figure 2.

One can notice that the distribution has the same bimodal shape as that found at low temperature, making use of model sequences.

We have analyzed the distribution $P(q_s)$ for ten different compact conformations of a 48 monomers chain. In all cases the distribution $P(q_s)$ taken at low temperature $T = 0.01$ exhibits three peaks and the last of them is centered at $q_s \approx 0.9$ (see e.g. Figure 8). The two peaks observed at lower values of q_s are always well separated from this peak by a minimum at $q_s \approx 0.7$. The position and width of the other peaks seem to be dependent on the specific conformation. This is in agreement with the interpretation we gave of the existence of these peaks. If they are caused by rearrangements of a substantial part of the sequence, the number of amino acids involved in these rearrangements can depend on the details of the

geometry of the system and thus the degree of similarity that the mutated sequences will display with each other.

To study the dependence of $P(q_s)$ on the interaction matrix, we repeated the same calculations with a random-generated matrix and with the matrix published by Kolinski et al. in ref. [15]. Unlike the MJ matrix, these do not display any similar columns and consequently the distribution $P(q_s)$ shows only two peaks. The peaks are located at $q_s = 0.55$ and $q_s = 0.95$ when simulations are run at $T = 0.01$. At higher temperature the ‘left’ peak moves to lower values of $q_s \approx 0.2$. Importantly, in this case for each structure, the type of residue occupying ‘hot’ sites is unique because with that set of parameters only one type of interactions can provide substantial stabilization to the hot sites. Hence in this case the superclusters probably do not exist in the energy landscape of sequence space.

In the case of models containing only two kinds of residues (Hydrophobic and Polar), the assertion that E_c , the lowest boundary energy of states that are structurally different from the native one, depends only on the composition of the chain, may not hold [22, 23]. Therefore in this case it may not be possible to describe in a simple fashion the space of large-gap sequences in terms of ground-state properties. Nonetheless, it is still necessary (but now not sufficient) for designed sequences to have, in the native conformation, low energy as compared, for example, to the average energy of compact configurations. Calculations performed with the matrix elements used in ref. [21] show that $P(q_s)$ has only one peak close to $q_s = 1$. In fact, the space of sequences of a two-letter model is not a frustrated system [18].

We also analyzed sequences composed of three kinds of residues. The behavior in this case is more similar to that found for 20 kinds of residues. Three seems then to be the minimum quantity of different residues necessary to have a non-trivial shape of $P(q_s)$.

Furthermore, we repeated the calculations optimizing the sequences with an algorithm that does not require the constraint of constant amino acid composition of the chain. This algorithm [24] maximizes the stability of the sequences, approximating the free energy of non-native states with a high temperature expansion. The resulting shape of $P(q)$ is very similar to the one shown in Figure 2, except for the fact that the low- q peak is very small. This last feature seems to be a non physical consequence of the high temperature approximation¹⁰.

The main result found about the non-connectivity of the space of good folders do not agree with the results of Bastolla and coworkers [25]. In this work, the authors, starting from a set of random sequences, span sequence space by single

¹⁰ While designing sequences with this technique, the use of residues D, E, K and R (which are usually very abundant in the folding nucleus) is unfavourable due to the fact that their columns, according to the MJ matrix contain two strongly attractive elements (i.e. E has $B_{EK} = -0.97$ and $B_{ER} = -0.74$, instead of one, i.e. C has only $B_{CC} = -1.05$) which decreases $\log Z$ in the mean field. So the designing algorithm prefers in the hot sites residues which have only one favourable counterpart (C, M, Q and W). This inhibits the mechanism which is responsible for the peak at low- q (the substitution of a core residue with a similar one). The small peak found at low- q , corresponding to the substitution of two similar residues agrees with this idea.

neutral mutations, reaching sequences which are 95% different than the starting one. From this they draw to the conclusion that the space of good folders is fully connected. The algorithm used makes a mutation, simulate the folding dynamics and accept the mutation only if the mutated chain still folds to the native conformation. While this algorithm is very effective in searching through a connected network of good folders, it is not useful to establish if there are non-connected zones, because, in any case, it could not cross barriers of non-folding sequences. Moreover, having linked with a neutral path two sequences which are 95% dissimilar is not a guarantee that the whole space is connected. In fact, we have shown that what defines a super-cluster is just the 5–10% of the residues (those in the ‘hot’ sites).

On the other hand, one has to keep in mind that, in the present work, we have made a drastic approximation, assuming that only sequences with energy below E_c are able to fold. This is not completely true [9] leaving the possibility that there do exist paths of folding sequences among super-clusters, characterized by energies somewhat above E_c . Even if this is true, they are likely to give raise to connecting paths which are so narrow that may not be operative from the evolutionary point of view. Furthermore, sequences building such paths are very unstable and prone to destructive mutations, so one should carefully consider if they are useful to evolution at all.

4. Conclusions

The energy landscape of the space of model protein sequences is extremely rough, as a consequence of the heterogeneity of the interaction among residues. This gives raise to a hierarchical clustering of low-energy sequences. Sequences gather together into cluster, which depend on the selective temperature T_s , while clusters are grouped into super-clusters at any selective temperature which produces folding sequences. Super-clusters are defined by residues in few ‘hot’ sites and are separated by high energy barriers, so that it is not possible to go from one super-cluster to another one by point mutations through a pathway of good folder sequences. This organization of the space of sequences can have evolutionary consequences. Sequences belonging to the same super-cluster are likely to have undergone divergent evolution, being the scions of a common ancestor which evolved without being able to cross the barrier to another super-cluster. Vice versa, sequences in different super-clusters can be the result of convergent evolution, that is sequences which have evolved independently on each other till they have built the nucleus (one among the few possible) which determines their fold. Another evolutionary conclusion we can draw is about dynamics. A non-trivial distribution of barriers translates into complex dynamics with multiple time scales. Moreover, we have found that evolutionary dynamics may proceed with distinctly different rates for different parts of the same protein.

References

1. Creighton, T.E.: *Proteins: Structures and Molecular Properties*. J. Freeman and co., New York (1993).
2. Shakhnovich, E.I.: *Phys. Rev. Lett.* **72**, 3907 (1994).
3. Go, N.: *Int. J. Peptide Prot. Res.* **7**, 313 (1975).
4. Lau, K.F. and Dill, K.: *Macromolecules* **22**, 3986 (1989).
5. Shakhnovich, E.I. and Gutin, A.M.: *J. Chem. Phys.* **93**, 5967 (1989).
6. Karplus, M. and Shakhnovich, E.I.: In: T.E. Creighton (ed.), *Protein Folding*. W.E. Freeman and Co., New York (1992).
7. Tiana, G., Broglia, R.A. and Shakhnovich, E.I.: *Prot. Struct. Funct. Gen.* **39**, 244 (2000).
8. Tiana, G., Broglia, R.A., Roman, H.E., Vigezzi, E. and Shakhnovich, E.I.: *J. Chem. Phys.* **108**, 757 (1998).
9. Broglia, R.A. and Tiana, G.: *J. Chem. Phys.* (in press).
10. Mezard, M., Parisi, G. and Virasoro, J.: *Spin Glasses and Beyond*. World Scientific, New York (1988).
11. Frauenfelder, H. and Wolynes, P.G.: *Physics Today* **February** (1994).
12. Shakhnovich, E.I. and Gutin, A.M.: *J. Phys. A* **22**, 1647 (1989).
13. Toulouse, G.: *Commun. Phys.* **2**, 115 (1977).
14. Miyazawa, S. and Jernigan, R.: *Macromolecules* **18**, 534 (1985), Table VI.
15. Kolinski, A., Godzik, A. and Sklonik, J.: *J. Chem. Phys.* **98**, 7420 (1993).
16. Shakhnovich, E.I. and Gutin, A.M.: *Biophys. Chem.* **34**, 187 (1989).
17. Derrida, B.: *Phys Rev. B* **24**, 2613 (1981).
18. Shakhnovich, E.I. and Gutin, A.M.: *Prot. Engin.* **6**, 793 (1993).
19. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E.: *J. Chem. Phys.* **21**, 1087 (1953).
20. Gutin, A.M., Abkevich, V.I. and Shakhnovich, E.I.: *Proc. Natl. Acad. Sci. USA* **92**, 1282 (1995).
21. Li, H., Helling, R., Tang, C. and Wingreen, N.: *Science* **273**, 666 (1996).
22. Sfatos, C., Gutin, A.M. and Shakhnovich, E.I.: *Phys. Rev. E* **48**, 465 (1993).
23. Deustch, J.M. and Kurosky, R.: *Phys. Rev. Lett.* **76**, 323 (1996).
24. Morressey, M. and Shakhnovich, E.I.: *Folding and Design* **1**, 391 (1996).
25. Bastolla, U., Roman, H.E. and Vendruscolo, M.: *Eur. Phys. J.* **B15**, 385 (2000).