INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **18** (2006) S297–S306 doi:10.1088/0953-8984/18/14/S13

Marginal compactness of protein native structures

Trinh X Hoang[1], Antonio Trovato[2], Flavio Seno[2], Jayanth R Banavar[3] and Amos Maritan[2]

[1] Institute of Physics and Electronics, VAST, 10 Dao Tan, Hanoi, Vietnam
[2] Dipartimento di Fisica 'G Galilei', Università di Padova, and INFN, Sezione di Padova, Via Marzolo 8, 35131 Padova, Italy
[3] Department of Physics, 104 Davey Lab, The Pennsylvania State University, University Park, PA 16802, USA

Received 22 September 2005, in final form 18 November 2005
Published 24 March 2006
Online at stacks.iop.org/JPhysCM/18/S297

Abstract

Globular proteins are a critically important component of the network of life. We recently proposed a simple coarse-grained model (Hoang *et al* 2004 *Proc. Natl Acad. Sci. USA* **101** 7960, Banavar *et al* 2004 *Phys. Rev.* E **70** 041905), which incorporates in a minimal way several physico-chemical features of globular proteins (GPs), such as the inherent anisotropy of a chain molecule and the geometrical and energetic constraints at the C^{α} description level imposed by hydrogen bonds, sterics, and hydrophobicity. Here, we provide a detailed description of the phase diagram of a 48-bead homopolymer chain, showing the existence and the robustness of a marginally compact phase in which the amount of degeneracy in low energy conformations is greatly reduced and the corresponding conformations exhibit a high amount of secondary structure content, thus resembling the native state folds of GPs. These results are obtained for chain lengths comparable to small real GPs and give further support to the hypothesis that GPs lie in a marginally compact phase in which the free energy landscape is pre-sculpted by geometry and symmetry.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

Proteins are well-tailored chain molecules employed by life to store and replicate information, to carry out a dizzying array of functionalities and to provide a molecular basis for natural selection. A protein molecule is a large and complex physical system with many atoms. In addition, the water molecules surrounding the protein play a crucial role in its behaviour. At the microscopic level, the laws of quantum mechanics can be used to deduce the interactions, but the number of degrees of freedom are far too many for the system to be studied in all its detail. When one attempts to look at the problem in a coarse-grained manner [3] with what one hopes are the essential degrees of freedom, it is very hard to determine what the effective potential

energies of interaction are. This situation makes the protein problem particularly daunting and no solution has yet been found. Nevertheless, proteins fold into a limited number [4, 5] of evolutionarily conserved structures [6, 7]. The same fold is able to house many different sequences which have that conformation as their native state, and is also employed by nature to perform different biological functions, pointing towards the existence of an underlying simplicity and of a limited number of key principles at work in proteins.

We have recently shown that a simple model which encapsulates a few general attributes common to all polypeptide chains, such as steric constraints [8–10], hydrogen bonding [11–13] and hydrophobicity [14], gives rise to the emergent free energy landscape of globular proteins [1, 2]. The relatively few minima in the resulting landscape for a short 24-bead homopolymer chain correspond to distinct putative marginally compact native-state structures of proteins, which are tertiary assemblies of helices, hairpins and planar sheets. A superior fit [15, 16] of a given protein or sequence of amino acids to one of these pre-determined folds dictates the choice of the topology of its native-state structure. Instead of each sequence shaping its own free energy landscape, we find that the overarching principles of geometry and symmetry determine the menu of possible folds that the sequence can choose from.

Sequence design would favour the appropriate native state structure over the other putative ground states, leading to an energy landscape conducive for rapid and reproducible folding of that particular protein. Nature has a choice of 20 amino acids for the design of protein sequences. A pre-sculpted landscape greatly facilitates the design process, and the introduction of sequence heterogeneity using a simple scheme of just two types of amino acids, hydrophobic (H) and polar (P), allows a selected putative native fold to become the free energy minimum at low temperature, as we showed in a specific case for a 24-bead chain [17]. We also showed how a longer HP sequence composed of regularly repeated patterns yields as a ground state a β-helical structure, remarkably similar to a known architecture in the Protein Data Bank [18], leading to the formation of a hydrophobic core characterized by a high degree of geometric regularity, as is the case for a broad class of proteins known as repeat proteins.

In this paper we will investigate the properties of the pre-sculpted free energy landscape by stepping back to the case of a homogeneous chain (with just one kind of amino acid, characterized by its own hydrophobicity and bending penalty) and analyzing in detail the properties of the phase diagram of a 48-bead chain. Similar chain lengths are shared by some of the smallest proteins and, if the scenario of a pre-sculpted landscape is confirmed, we expect to recover a rich variety of putative native topologies having their counterpart in real protein structures from the Protein Data Bank. We will then provide a quantitative analysis of the lowest energy conformations found by means of several simulated annealing simulations showing that the phase, in which protein-like conformations with a high amount of secondary structure content are found, lies at the boundary between the swollen and the compact phase. The phase is also characterized by a marked decrease in the degeneracy of the corresponding energies, strongly supporting the notion of protein structures belonging to a *marginally compact* phase in which the free energy landscape is pre-sculpted by considerations of geometry and symmetry.

2. Methods

We model a protein as a chain of *identical* amino acids, represented by their C^α atoms, lying along the axis of a self-avoiding flexible tube. The preferential parallel placement of nearby tube segments approximately mimics the effects of the anisotropic interaction of hydrogen bonds, while the space needed for the clash-free packing of side chains is approximately captured by the non-zero tube thickness [19–21]. A tube description places constraints on the

radii of circles drawn through both local and non-local triplets of C^α positions of a protein native structure [20, 22].

Unlike unconstrained matter, for which pairwise interactions suffice, for a chain molecule it is necessary to define the context of the object that is part of the chain. This is most easily carried out by defining a local Cartesian coordinate system whose three axes are defined by the tangent to the chain at that point, the normal that is perpendicular to the tangent and pointing to the centre of the circle which defines the local radius of curvature, and the binormal, which is perpendicular to both the other two vectors. A study [1, 2] of the experimentally determined native state structures of proteins from the Protein Data Bank reveals that there are clear amino acid aspecific geometrical constraints on the relative orientation of the local coordinate systems due to sterics and also associated with amino acids which form hydrogen bonds with each other. Similar geometrical constraints had already been introduced in off-lattice polymer models [23, 24] in order to model hydrogen bond formation.

The geometrical constraints associated with the formation of hydrogen bonds and with the tube description within the C^α representation of our model are described in detail elsewhere [1, 2]. In our representation of the protein backbone, local hydrogen bonds form between C^α atoms at distance three along the sequence with an energy defined to be -1 unit, whereas non-local hydrogen bonds are those that form between C^α atoms separated by more than 4 along the sequence with an energy of -0.7. This energy difference is based on experimental findings that the local bonds provide more stability to a protein than do the non-local hydrogen bonds [25]. Cooperativity effects [26, 27] are taken into account by adding an energy of -0.3 units when consecutive hydrogen bonds along the sequence are formed. There are two other ingredients in the model: a local bending penalty e_{R} which is related to the steric hindrance of the amino acid side chains and a pair-wise interaction e_{W} of the standard type mediated by the water [14]. Note that whereas the geometrical constraints associated with the tube and hydrogen bonds are representative of the typical *aspecific* behaviour of the interacting amino acids, the latter properties clearly depend on the *specific* amino acids involved in the interaction.

We carried out Metropolis Monte Carlo simulations for a linear chain molecule made up of 48 amino acids each represented by its C^α atom. We used pivot and crankshaft move sets, commonly used in stochastic chain dynamics [28], with a probability of 0.1 for selecting a pivot move and 0.9 for the crankshaft ones. In a pivot move, an amino acid is selected at a random position i along the chain and a random axis of rotation is generated. An attempt is made to perform the rotation about this axis for all amino acids from $i + 1$ to N by an angle randomly drawn from a Gaussian distribution with zero mean and a dispersion of 4°. In a crankshaft move, two amino acids i and j are chosen randomly with the requirement $i + 1 < j < i + 6$ and the rotation is attempted for all the residues between i and j about the axis joining the ith and jth amino acids, again drawing the rotation angle randomly from a Gaussian distribution with zero mean and a dispersion of 4°. The new conformation is rejected if either steric clashes occur or the tube constraint is violated, otherwise it is accepted with a probability $P = \min\{1, \mathrm{e}^{-\Delta E/T}\}$, where T denotes a fictitious temperature and ΔE is the change in energy.

The simulations start from a randomly chosen open conformation and the starting temperature is $T = 0.3$ (the Boltzmann constant has been fixed to unity). An annealing scheme is adopted in which the temperature is slowly decreased from 0.3 to about 0.1 in 36 steps according to the recursive formula $T_{k+1} = 0.97T_k$, after which a quenching to $T = 0$ is performed. For each T, the number of attempted moves is 20 million. The conformation obtained after the quenching is saved as a local energy minimum. About 1200 such conformations have been generated from independent runs with the energy parameters e_{R} and e_{W} varied in a region including the marginally compact phase of the homopolymer.

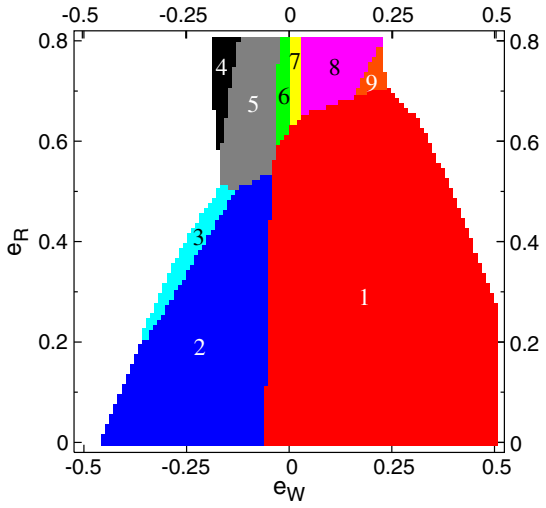

Figure 1. Ground state phase diagram obtained from simulations. The different colours correspond to different ground state conformations: single helix (1-red) (figure 2(a)), three-helix bundle (2-blue) (figure 2(c)), a conformation comprised of two helices and a sheet (3-cyan) (figure 2(d)), a helix of strands (4-black) (figure 2(h)), a sandwich of sheets (5-grey) (figure 2(i)), three kinds of four-stranded sheet (6-green, 7-yellow and 8-magenta) (shown in figure 2(j), (k) and (l) respectively), and a five-stranded sheet (9-orange) (figure 2(m)).

Specifically, e_R is varied between 0 and 0.8 energy units and e_W is varied between -0.5 and 0.5 energy units in steps of 0.05. A large number of runs (around 400) were carried out for $e_R = 0.15$ and $e_W = -0.1$, which corresponds to a ground state of a three-helix bundle (see figure 1 for the ground state phase diagram). For other values of the parameters, the number of runs is typically 20. The runs with $e_W < -0.2$ typically yield conformations which are very compact, while those with $e_W > 0.2$ lead to open conformations. In both cases, the structures obtained are quite disordered and contain little secondary structure. Protein-like conformations, predominantly modular structures built of helices and sheets, occur in the vicinity of the swollen phase when the hydrophobic interaction energy parameter, e_W, is negative but close to zero. On occasion, several runs yield very similar conformations, e.g. the single helix or the three-helix bundle. By removing similar structures whose root mean square deviation (rmsd) is smaller than 2 Å, we obtain a set of 1079 distinct local energy minima constituting the ensemble which will be analysed in the next section.

3. Results and discussion

The model confirms the existence of a pre-sculpted free energy energy landscape (figure 1) for a homopolymer made up of just one kind of amino acid (compare with figure 3 of [1] for a 24-bead chain). On varying the two parameters, e_W and e_R, one obtains a set of ground states and low-lying energy minima (figure 2), some of which bear a striking resemblance to native state protein structures. The low energy structures are commonly made up of helices and sheets, the building blocks of proteins, which are both characterized by the nearly parallel placement of nearby chain segments as expected from both packing considerations and the intrinsic anisotropy of the chain molecule [21]. Note that the observed folds lie (figure 1) in the vicinity of a swollen phase without any compaction, obtained for positive enough e_W (the introduction

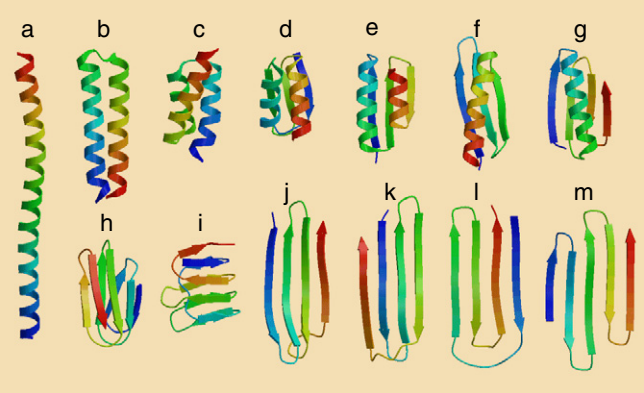

Figure 2. Gallery of some of the low energy conformations of a homopolymer of 48 residues. Conformations (a), (c), (d), (h), (i), (j), (k), (l) and (m) are the ground states for certain ranges of the energy parameters (see figure 1), whereas other conformations ((b), (e), (f) and (g)) are representative low-lying energy minima. The structures are shown in the Molscript presentation in which helices and strands are assigned based on the formation of local or non-local hydrogen bonds according to the rules defined in the model.

of intra-chain hydrogen bonds without the possibility of explicit hydrogen bonding to water molecules introduces an effective attraction which is also present when $e_W = 0$), whereas for negative enough e_W one finds compact disordered low energy conformations. The proximity of the marginally compact phase to a phase transition suggests that the structures ought to be sensitive to the right type of perturbations akin to the remarkable conformational flexibility exhibited by real proteins.

We quantify the similarity of the structures corresponding to the local minima to real protein native folds by looking at the amount of secondary structure present in the conformations. The secondary structure assignment is made according to the rules associated with the hydrogen bonds in the model. Specifically, if there are two cooperative local hydrogen bonds $(i, i+3)$ and $(i+1, i+4)$ then all amino acids from i to $i+4$ are considered to be in a helix conformation. If there are two cooperative non-local hydrogen bonds (i, j) and $(i \pm 1, j \pm 1)$ then all amino acids participating in these hydrogen bonds are considered to be in a strand conformation.

The secondary structure content is defined as the total number of residues in a helix or a strand divided by 48, the total number of amino acids in the chain molecule. Figure 3, the histogram of the number of local minima having a given secondary structure content, exhibits two main peaks—one near the origin corresponding to structures in the swollen phase and the other with a secondary structure content of about 0.7. Interestingly, when maximally compact local minima are discarded by imposing a lower bound on the radius of gyration, the distribution of secondary structure content is shifted towards higher values and is significantly lower in the intermediate range, from 0.1 to 0.5. This provides a clear hint that protein-like conformations can be characterized as *marginally compact* structures. In fact, figure 4 shows a sample of randomly chosen structures, which are neither too compact nor open (7.6 Å $< R_g <$ 14 Å), with secondary structure content between 0.4 and 0.7. While these structures contain some disordered loops, they nevertheless show a high resemblance to real protein structures.

In order to assess the role of energy in the selection of protein like structures, we rank the local minimum structures by their energies in the following way. e_R and e_W are independently changed in steps of 0.01 in the ranges $0 \leqslant e_R \leqslant 0.8$ and $-0.2 \leqslant e_W \leqslant 0$. This part of the

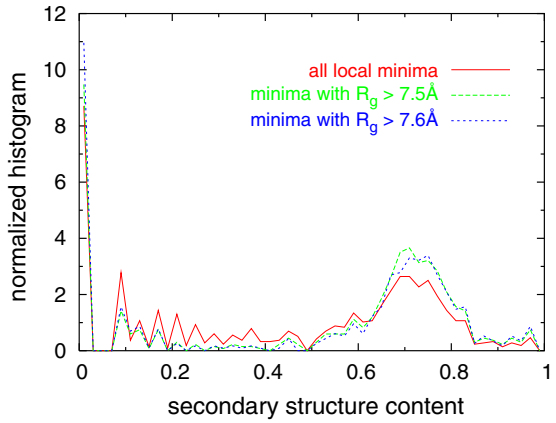

Figure 3. Histograms of the number of structures with a given secondary structure content. The area under each of the curves has been normalized to unity. The lines correspond to different constraints on the radius of gyration, R_g, as indicated.

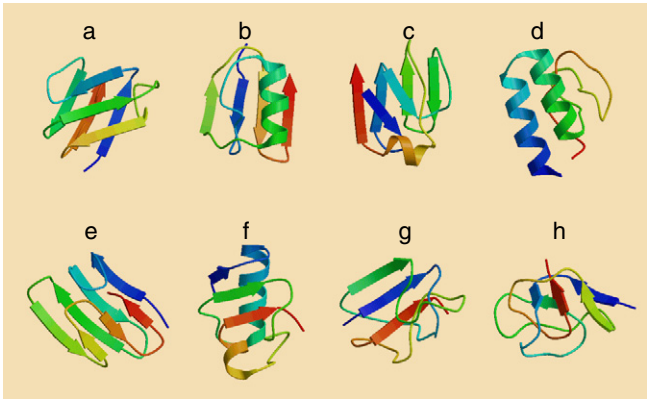

Figure 4. Molscript representation of randomly selected local minima with the radius of gyration, R_g, in the interval 7.6 Å $< R_g <$ 14 Å. The secondary structure contents of the conformations are 0.646 (a), 0.687 ((b), (c), (d)), 0.666 ((e), (f)), 0.542 (g), and 0.458 (h).

parameter space is expected to be most favourable for protein-like structures. For each pair of values of the parameters, we compute the energies of the local minima and rank them in energy so that the lowest energy minimum has rank 1. We then assign each decoy its final rank given as the lowest rank obtained among all the sets of energy parameters. Every structure in the gallery shown in figure 2 is ranked better than 20 (of course, all the ground states have rank 1). The structures in the gallery shown in figure 4 have ranks that vary between 30 and 500. Figure 5(a) shows a histogram of the number of structures with a given normalized rank defined as the rank divided by the total number of structures (we have sorted the ranks into 20 equal bins). The histogram exhibits two clear peaks which suggest that the set of structures can be divided into two groups—one is a group of structures that are competitive in energy having the normalized ranks lower than 0.5 and another group, having the normalized rank greater than 0.5, corresponds to structures that are not competitive in energy. Figure 5(b) shows very clearly that structures in the first group have a secondary structure content much higher than those in the second group. Furthermore, the secondary structure content is a decreasing

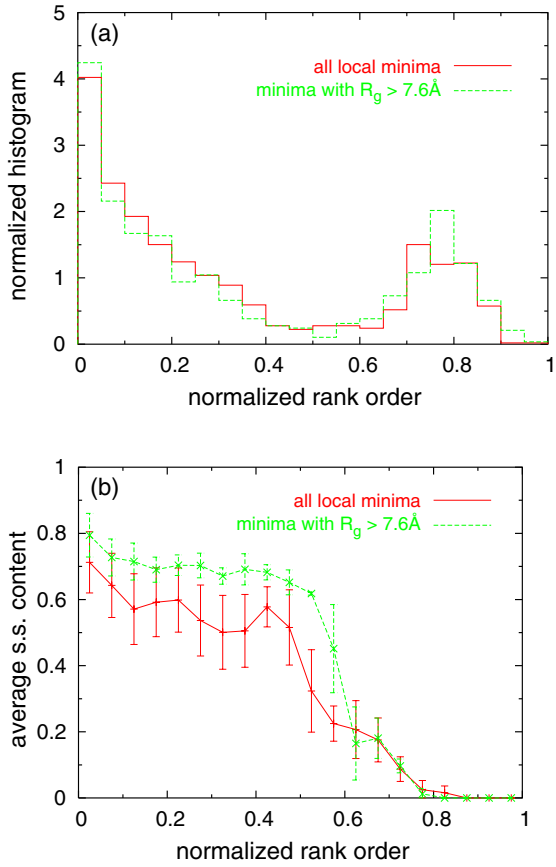

Figure 5. (a) Histograms of the number of structures as a function of their normalized rank (rank divided by the total number of structures) for all local minima and for only those local minima with $R_g > 7.6$ Å. (b) Plot of the average secondary structure content as a function of the normalized rank for the two sets of local minima.

function of the rank. When one considers structures which are not too compact (i.e. with $R_g > 7.6$ Å), one obtains similar trends with a general increase in the overall secondary structure content. Our results highlight the importance of both compactness and energy on the selection of protein-like structures.

We turn now to a more quantitative assessment of how different properties of our set of top-ranking local minima depend on the strength of the hydrophobic attraction e_W. We compute the average radius of gyration, R_g, and secondary structure content as a function of e_W in the following way: we first average over the 100 local minima having the lowest energies for a given (e_W, e_R) point in the phase diagram and then average the resulting averages over different e_R values, separated by the same 0.01 step as above, with e_W fixed. As seen in figure 6(a), the average gyration radius decreases from values typical of open extended conformations at positive e_W, when the interaction is repulsive and one has a swollen phase, to the lowest values typical of maximally compact conformations at sufficiently negative e_W. Such a behaviour is *indistinguishable* from that expected for the usual θ-transition for conventional homopolymers, as solvent quality is reduced, when the geometric constraints enforced by hydrogen bonding and the tube symmetry are not present. The signature of the geometry and symmetry of proteins

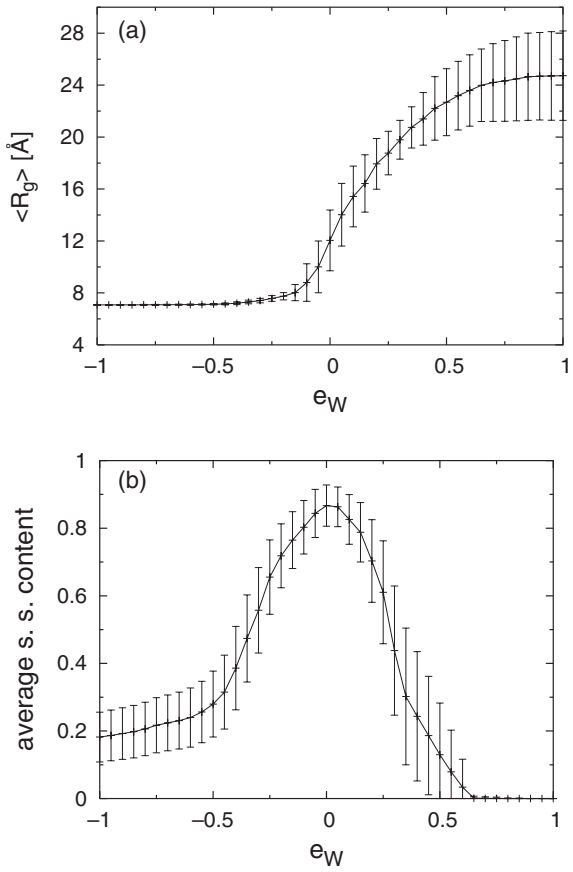

Figure 6. (a) Mean radius of gyration, averaged over different values of the local bending penalty e_R, and over the 100 lowest energy local minima at each (e_W, e_R) pair, as a function of the strength e_W of the hydrophobic attraction. (b) Average secondary structure content computed in the same way as a function of the strength e_W of the hydrophobic attraction. Bars in both plots correspond to standard deviations.

becomes evident on monitoring the secondary structure content. Figure 6(b) exhibits a clear peak of the secondary structure content in an intermediate window of e_W values, corresponding to the radius of gyration just above its lowest value for maximally compact structures. Notice that the width of the peak nicely matches the width of the marginally compact phase in figure 1.

The key point is that the combined action of hydrogen bond geometry and tube symmetry not only selects protein-like conformations in the proper region of parameter space, but also sculpts the energy landscape in such a way that the resulting degeneracy of low energy local minima is much lower than either in the swollen or in the maximally compact phase. This crucial result can be quantitatively appreciated in figure 7, where we show the contour plot of the standard deviation computed from the energy distribution for the 100 local minima with the lowest energies. A low degeneracy of the local minima in the energy landscape is evidently associated with a high standard deviation of the corresponding energy distribution. In the marginally compact region of the parameter space, at intermediate e_W values, protein-like conformations are selected as ground states and local energy minima (see figure 1), and, furthermore (see figure 7), one has a greatly reduced degeneracy. Interestingly, the shape of the

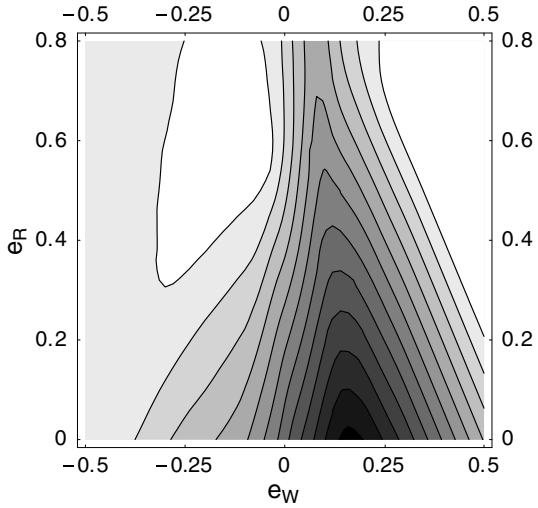

Figure 7. Contour plot of the standard deviation of the energy distribution for the 100 local minima with the lowest energies in the (e_W, e_R) parameter space. The darker the grey the higher the standard deviation, and the less degenerate the low energy structures. Neighbouring contour levels are separated by one energy unit.

standard deviation level sets in figure 7 closely follows the borders of the marginally compact phase in figure 1.

The degeneracy is reduced to its lowest level in the positive e_W region where the single helix (see figure 2(a)) is most stable. Since both diversity and stability are essential goals of any evolutionary process, most of the protein-like conformations shown in figure 2 are found for $e_W \sim -0.1$. In this region, the degeneracy is higher than in the single-helix region, allowing for a broad menu of putative ground states, but it is nevertheless lower than in the maximally compact region. When the attractive hydrophobic energy parameter is strengthened, one obtains more compact structures (as measured by the radius of gyration) with lower amounts of secondary structure.

4. Conclusions

In summary, we have shown within a simple framework that protein-like structures can arise from considerations of symmetry and geometry associated with the polypeptide chain. The sculpting of the free energy landscape with relatively few broad minima is consistent with the fact that proteins can be designed to enable rapid folding to their native states. Our results underscore the essential differences in the phases of matter associated with conventional polymers and biological molecules: the generic compact phase of polymers has a large number of compact ground states with little secondary structure and these structures are not characterized by flexibility and versatility; the marginally compact phase has a vastly simpler energy landscape with relatively few putative ground states and a greatly reduced degeneracy in the low-lying energy minima. The putative folds presented by the low energy states are dynamically accessible, allowing proteins to fold reproducibly and rapidly, and the vicinity of the marginally compact phase to the swollen phase provides the sensitivity and flexibility associated with protein native state structures.

It has been suggested that proteins must follow the maximal consistency [29] and the minimal frustration principles [15]. Our idea of the pre-sculpted free energy landscape is not

in conflict with these principles. Rather, our work suggests that sequence design follows in a much simpler way than envisioned before. Instead of the sequence shaping the free energy landscape from scratch, we suggest that the role of the sequence is merely to select from a pre-determined menu of possible structures in such a way that one obtains a good sequence–structure fit yielding a smooth folding funnel with a large basin of attraction. We have found that the number of marginally compact ground states in the $N = 48$ homopolymer case does not increase substantially compared to the $N = 24$ case. Likewise, the number of low-lying energy minima increases as the chain length grows only modestly due to the persistence lengths provided by the secondary structures. Thus it is a reasonable guess that for chain lengths smaller than 200 the number of distinct folds is of the order of a few thousand. Our model does not predict the favourable topologies of tertiary compaction because this depends on sequence specificity and details. In fact, for the chain length considered, we have found that only a few tertiary motifs in our low-lying energy minima structures are similar to those that exist in real proteins. One cannot *a priori* exclude the possibility that proteins in nature have utilized only a fraction of all the folds allowed by geometry and symmetry.

Acknowledgments

This work was supported by PRIN 2003, NASA, NSF IGERT grant DGE-9987589, NSF MRSEC at Penn State and the NSC of Vietnam (grant No 410704).

References

[1] Hoang T X, Trovato A, Seno F, Banavar J R and Maritan A 2004 *Proc. Natl Acad. Sci. USA* **101** 7960
[2] Banavar J R, Hoang T X, Maritan A, Seno F and Trovato A 2004 *Phys. Rev.* E **70** 041905
[3] Banavar J R and Maritan A 2001 *Proteins* **42** 433
[4] Chothia C and Finkelstein A V 1990 *Annu. Rev. Biochem.* **59** 1007
[5] Chothia C 1992 *Nature* **357** 543
[6] Denton M and Marshall C 2001 *Nature* **410** 417
[7] Chothia C, Gough J, Vogel C and Teichmann S A 2003 *Science* **300** 1701
[8] Ramachandran G N and Sasisekharan V 1968 *Adv. Protein Chem.* **23** 283
[9] Pappu R V, Srinivasan R and Rose G D 2000 *Proc. Natl Acad. Sci. USA* **97** 12565
[10] Baldwin R L and Rose G D 1999 *Trends Biochem. Sci.* **24** 26
[11] Pauling L, Corey R B and Branson H R 1951 *Proc. Natl Acad. Sci. USA* **37** 205
[12] Pauling L and Corey R B 1951 *Proc. Natl Acad. Sci. USA* **37** 729
[13] Eisenberg D 2003 *Proc. Natl Acad. Sci. USA* **100** 11207
[14] Kauzmann W 1959 *Adv. Protein Chem.* **14** 1
[15] Bryngelson J D and Wolynes P G 1987 *Proc. Natl Acad. Sci. USA* **84** 7524
[16] Brenner S A 2001 *Nature* **409** 459
[17] Hoang T X, Trovato A, Seno F, Banavar J R and Maritan A 2005 *Biophys. Chem.* **115** 289
[18] Trovato A, Hoang T X, Banavar J R, Maritan A and Seno F 2005 *J. Phys.: Condens. Matter* **17** S1515
[19] Maritan A, Micheletti C, Trovato A and Banavar J R 2000 *Nature* **406** 287
[20] Banavar J R and Maritan A 2003 *Rev. Mod. Phys.* **75** 23
[21] Marenduzzo D, Flammini A, Trovato A, Banavar J R and Maritan A 2005 *J. Polym. Sci. Polym. Phys.* **43** 650
[22] Banavar J R, Maritan A, Micheletti C and Trovato A 2002 *Proteins* **47** 315
[23] Kemp J P and Chen Z Y 1998 *Phys. Rev. Lett.* **81** 3880
[24] Trovato A, Ferkinghoff-Borg J and Jensen M H 2003 *Phys. Rev.* E **67** 021805
[25] Shi Z, Krantz B A, Kallenbach N and Sosnick T R 2002 *Biochemistry-US* **41** 2120
[26] Liwo A, Kazmierkiewicz R, Czaplewski C, Groth M, Oldziej S, Rackovski R J, Pincus M R and Scheraga H A 1998 *J. Comput. Chem.* **19** 259
[27] Fain B and Levitt M 2003 *Proc. Natl Acad. Sci. USA* **100** 10700
[28] Sokal A D 1996 *Nucl. Phys.* B (Suppl. 47) 172
[29] Go N 1983 *Annu. Rev. Biophys. Biochem.* **12** 183