Geometry and symmetry presculpt the free-energy landscape of proteins

Trinh Xuan Hoang⁎, Antonio Trovato†, Flavio Seno‡, Jayanth R. Banavar⁎, and Amos Maritan§

⁎Institute of Physics, National Centre for Natural Science and Technology, 46 Nguyen Van Ngoc, Hanoi, Vietnam; †Italian National Institute for Materials Physics and Dipartimento di Fisica "G. Galilei," Università di Padova, Via Marzolo 8, 35131 Padua, Italy; ‡Department of Physics, 104 Davey Laboratory, Pennsylvania State University, University Park, PA 16802; and §Abdus Salam International Center for Theoretical Physics, Strada Costiera 11, 34014 Trieste, Italy

Communicated by Morrel H. Cohen, Rutgers, The State University of New Jersey, Piscataway, NJ, April 8, 2004 (received for review October 9, 2003)

We present a simple physical model that demonstrates that the native-state folds of proteins can emerge on the basis of considerations of geometry and symmetry. We show that the inherent anisotropy of a chain molecule, the geometrical and energetic constraints placed by the hydrogen bonds and sterics, and hydrophobicity are sufficient to yield a free-energy landscape with broad minima even for a homopolymer. These minima correspond to marginally compact native-state structures of proteins, which are assemblies of helices, hairpins, and planar sheets. A superior fit of a given non-zero tube thickness (21, 22). Here, we carefully incorporate these key geometrical features by means of an extensive statistical analysis of experimentally determined native-state structures in the Protein Data Bank (PDB).

A tube description places constraints on the radii of circles drawn through both local and nonlocal triplets of Cα positions of a protein native structure (22, 23). Furthermore, when one deals with a chain molecule, the tube picture underscores the crucial importance of knowing the context that an amino acid is in within the chain. The standard coarse-grained approach considers the locations of interacting amino acid pairs. Here, instead, we incorporate the strongly directional hydrogen bonding between a pair of amino acids, through an analysis of the PDB to determine the constraints on the mutual orientation of the local coordinate systems defined from a knowledge of the locations of the Cα atoms (see Methods and Fig. 1). The geometrical constraints associated with the tube and hydrogen bonds that we consider here are representative of the typical specific behavior of the interacting amino acids.

There are two other ingredients in the model: a local bending penalty, which is related to the steric hindrance of the amino acid side chains, and a pair-wise interaction of the standard type aspecific behavior of the interacting amino acids.

Methods

Tube Geometry. The protein backbone is modeled as a chain of Cα atoms (Fig. 2a) with a fixed distance of 3.8 Å between successive atoms along the chain, an excellent assumption for all but non-cis proline amino acids (24). The geometry imposed by chemistry dictates that the bond angle associated with three consecutive Cα atoms is between 82° and 148°.

Self-avoiding conformations of the tube whose axis is the protein backbone are identified by considering all triplets of Cα atoms and drawing circles through them and ensuring that none of their radii is smaller than the tube radius (25) (Fig. 2a). At the local level, the three-body constraint ensures that a flexible tube cannot have a radius of curvature any smaller than the tube thickness, to prevent sharp corners, whereas, at the nonlocal level, it does not permit any self-intersections. There is an inherent local anisotropy due to the special direction singled out by consecutive atoms along the chain, which enforces a preference for parallel alignment of neighboring tube segments in a compact conformation.

The backbone of Cα atoms is treated as a flexible tube of radius 2.5 Å, a constraint imposed on all (local and nonlocal)
three-body radii, an assumption validated for protein native structures (23). It is interesting to note that recent observations of residual dipolar couplings in short peptides (26) in the denatured state have demonstrated their stiffness and their anisotropic deformability; the building blocks of proteins are relatively stiff segments with strong directional preferences.

Sterics. Steric constraints require that no two nonadjacent C\textsubscript{α} atoms are allowed to be at a distance closer than 4 Å. Ramachandran and Sasisekaran (11) showed that steric considerations based on a hard sphere model lead to clustering of the backbone dihedral angles in two distinct α and β regions for non-glycyl and non-prolyl residues. The two backbone geometries that allow for systematic and extensive hydrogen bonding (14–16) are the α-helix and the β-sheet, obtained by a repetition of the backbone dihedral angles from the two regions, respectively (13). Short chains rich in alanine residues, which are a good approximation to a stretch of the backbone, can adopt a helical conformation in water (see refs. 27–32 for a detailed discussion of experimental conditions that would lead to a helical conformation). However, when one has more heterogeneous side chains, the helix backbone could sterically clash with some side chain conformers, resulting in a loss of conformational entropy (33). When the price in side chain entropy is too large, an extended backbone conformation results, pushing the segment toward a β-strand structure (13). These steric constraints are approximately imposed through an energy penalty (denoted by \( e_R \)) when the local radius of curvature is between 2.5 Å and 3.2 Å. (The magnitude of the penalty does not depend on the specific value of the radius of curvature, provided it is between these values.) There is no cost when the local radius exceeds 3.2 Å. Note that the tube constraint does not permit any local radius of curvature to take on a value less than the tube radius, 2.5 Å.

Hydrogen Bonds. We do not allow more than two hydrogen bonds to form at a given C\textsubscript{α} location. In our representation of the protein backbone, local hydrogen bonds form between C\textsubscript{α} atoms separated by two residues along the sequence with an energy defined to be \(-1\) unit whereas nonlocal hydrogen bonds are those that form between C\textsubscript{α} atoms separated by more than three residues 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 nonlocal hydrogen bonds (34). Cooperativity effects (35, 36) are taken into account by adding an energy of \(-0.3\) units when consecutive hydrogen bonds along the sequence are formed. There is some latitude in the choice of the values of these

---

**Fig. 1.** Sketch of the local coordinate system. For each C\textsubscript{α} atom \( i \) (except the first and the last one), the axes of a right-handed local coordinate system are defined as follows. The tangent vector \( \hat{t}_i \) is parallel to the segment joining \( i-1 \) with \( i+1 \). The normal vector \( \hat{n}_i \) joins \( i \) to the center of the circle passing through \( i-1, i \), and \( i+1 \), and it is perpendicular to \( \hat{t}_i \) and \( \hat{n}_i \) along with the three contiguous C\textsubscript{α} atoms lie in a plane shown in the figure. The binormal vector \( \hat{b}_i \) is perpendicular to this plane. The vectors \( \hat{t}_i, \hat{n}_i, \) and \( \hat{b}_i \) are normalized to unit length.

**Fig. 2.** Sketch of a portion of a protein chain. (a) The black spheres represent the C\textsubscript{α} atoms of the amino acids. The local radius of curvature \( r \) is defined as the radius of the circle passing through three consecutive atoms and is constrained to lie between 2.5 Å and 7.9 Å \( (r_{max}) \). A penalty \( e_R \) is imposed when \( 2.5 \leq r \leq 3.2 \) Å. The hydrophobic interaction, \( e_W \), is operative when two atoms separated by more than two along the sequence are within 7.5 Å of each other (see c). Note that two nonadjacent atoms cannot be closer than 4 Å. A flexible tube is characterized by the constraint that none of the three-body radii is less than the tube thickness, chosen here to be 2.5 Å (see b and d).
associated with each local hydrogen bond is positive. The interactions between 70° and 79.8°, viz. $i < j < k$

The binormal vectors are not defined. In order for such residues to form a hydrogen bond, the associated ending peptide link and the connecting vector $r_{ij}$, the distance between these atoms ought to be between 4.7 Å and 5.6 Å (4.1 Å and 5.3 Å) for the local (nonlocal) case. A study of protein native state structures reveals an overall nearly parallel alignment of the axes defined by three vectors: the binormal vectors at $i$ and $j$ and the vector $r_{ij}$ joining the $i$ and $j$ atoms. A hydrogen bond is allowed to form only when the binormal axes are constrained to be within 37° of each other whereas the angle between the binormal axes and that defined by $r_{ij}$ ought to be <20°. Additionally, for the cooperative formation of nonlocal hydrogen bonds, one requires that the corresponding binormal vectors of successive $C\alpha$ atoms make an angle >90°. The first and the last residues of the chain are special cases because their binormal vectors are not defined. In order for such residues to form a hydrogen bond (with each other or with other internal residues in the chain), it is required that the angle between the associated ending peptide link and the connecting vector to the other residue participating in the hydrogen bond is between 70° and 110°. As in real protein structures, when helices are formed, they are constrained to be right-handed. This chirality is defined as the sign of $(\mathbf{t}_{i+1} \times \mathbf{t}_{i+2}) \mathbf{r}_{i+2,i+3}$.

Our approach for the derivation of the geometrical constraints imposed by hydrogen bonds is similar to that carried out at the level of an all-atom description of the protein chain (37). For the simpler $C\alpha$ atom-based description, hydrogen bond energy functions have been introduced previously (38, 39) but without any input from a statistical analysis of protein structures.

Hydrophobic Interactions. The hydrophobic (hydrophilic) effects mediated by the water are captured through a relatively weak interaction $e_{\text{water}}$ (either attractive or repulsive) between $C\alpha$ atoms that are within 7.5 Å of each other (Fig. 2c). Note that hydrogen bonds can easily be formed between the amino acid residues in an extended conformation and the water molecules. Within our model, the intra-chain hydrogen bond interaction introduces an effective attraction, because water molecules are not explicitly present. The hydrophobicity scale is thus renormalized (e.g., even when $e_{\text{water}}$ is weakly positive, there could be an effective attraction resulting in structured conformations such as a single helix or a planar sheet). A negative $e_{\text{water}}$ is, in any case, crucial for promoting the assembly of secondary motifs in native tertiary arrangements. The properties of the model are summarized in Table 1.

Results and Discussion

Fig. 3 shows the ground state phase diagram obtained from Monte-Carlo computer simulations using the simulated annealing technique (40). [The solvent-mediated energy, $e_{\text{water}}$, and the local radius of curvature energy penalty, $e_{\text{R}}$ (see Methods for a description of the energy parameters) are measured in units of the local hydrogen bond energy.] When $e_{\text{water}}$ is sufficiently repulsive (hydrophilic) and $e_{\text{R}} > 0.3$ in the phase diagram, one obtains a swollen phase with very few contacts between the $C\alpha$ atoms. When $e_{\text{water}}$ is sufficiently attractive, one finds a very compact, globular phase with featureless ground states with a high number of contacts.

Between these two phases (and in the vicinity of the swollen phase), a marginally compact phase emerges (the interactions barely stabilize the ordered phase) with distinct structures including a single helix, a bundle of two helices, a helix formed by $\beta$-strands, a $\beta$-hairpin, three-stranded $\beta$-sheets with two distinct topologies, and a $\beta$-barrel-like conformation. Strikingly, these structures are the stable ground states in different parts of the phase diagram. Furthermore, other conformations, closely resembling distinct supersecondary arrangements observed in proteins (6), such as the $\beta$-$\alpha$-$\beta$ motif, are found to be competitive local minima whose stability can be enhanced by sequence design (for example, nonuniform values of curvature energy penalties for single amino acids and hydrophobic interactions for amino acid pairs) (Fig. 4 shows a compendium of various structures
obtained in our studies, including for comparison a generic compact conformation of a conventional polymer chain (with no tube geometry or hydrogen bonds), which neither is made up of helices or sheets nor possesses the significant advantages of protein structures. Although there is a remarkable similarity between the structures that we obtain and protein folds, our simplified coarse-grained model is not as accurate as an all-atom representation of the polypeptide chain in capturing features such as the packing of amino acid side chains.

The fact that different putative native structures are found to be competing minima for the same homopolymeric chain clearly establishes that the free-energy landscape of proteins is presculpted by means of the few ingredients used in our model. At the same time, relatively small changes in the parameters $e_W$ and $e_R$ lead to significant differences in the emergent ground state structure, underscoring the sensitive role played by chemical heterogeneity in selecting from the menu of native state folds.

Fig. 5a is a contour plot of the free energy at a temperature higher than the folding transition temperature (identified by the specific heat peak) for the parameter values $e_W = -0.08$ and $e_R = 0.3$ for which the ground state is an $\alpha$-helix (Fig. 3). The free energy landscape has just one minimum corresponding to the denatured phase whose typical conformations are somewhat compact but featureless. The contour plot at the folding transition temperature (Fig. 5b) has three local minima corresponding to an $\alpha$-helix, a three-stranded $\beta$-sheet, and the denatured state. At lower temperatures, the $\alpha$-helix is increasingly favored and the $\beta$-sheet is never the global free-energy minimum. Many protein-folding experiments show that, for small globular proteins at the transition temperature, only two states (folded and unfolded) are populated. The appearance in the present framework of multiple states for a homopolymer chain suggests that two-state folders might have been evolutionarily selected by sequence design favoring the native-state conformation over competing folds in the presculpted landscape.

Such a design is indeed straightforward within our model. For example, the $\alpha-\beta-\alpha$ motif shown in Fig. 4d (which is a local energy minimum for a homopolymer) can be stabilized into a global energy minimum for the sequence HPHHHHPHPHHP-PHPPPPPPhHPPPP, with $e_W = -0.4$ for HH contacts and $e_W = 0$ for other contacts, and $e_R = 0.3$ for all residues. It is interesting to note that lattice models of compact homopolymers yield large amounts of secondary structure (41); local radius of curvature constraints are built into lattice models. However, an all-atom study of polyalanine has shown that compactness alone is insufficient to obtain secondary structures (42). Even a simple tube subject to an attractive self-interaction favoring compaction has a tendency to form helices, hairpins and sheets when the ratio of the tube thickness to the range of attractive interaction is tuned properly (22). Our work here underscores the importance of hydrogen bonds in stabilizing both helices and sheets simultaneously (without any need for adjustment of the tube thickness), allowing the formation of tertiary arrangements of secondary motifs. Indeed, the fine tuning of the hydrogen bond and the hydrophobic interaction is of paramount importance in the selection of the marginally

---

**Fig. 3.** Phase diagram of ground state conformations. The ground state conformations were obtained by using Monte-Carlo simulations of chains of $24\ C$ atoms. $e_\alpha$ and $e_\beta$ denote the local radius of curvature energy penalty and the solvent-mediated interaction energy, respectively. Over 600 distinct local minima were obtained in different parts of parameter space, starting from a random conformation and successively distorting the chain with pivot and crankshaft moves commonly used in stochastic chain dynamics (43). A Metropolis Monte-Carlo procedure is used with a thermal weight $exp(-E/T)$, where $E$ is the energy of the conformation and the temperature $T$ is set initially at a high value and then decreased gradually to zero. In the orange phase, the ground state is a two-stranded $\beta$-hairpin. Two distinct topologies of a three-stranded $\beta$-sheet (dark and light blue phases) are found corresponding to conformations shown in conformations $i$ and $j$ in Fig. 4, respectively. The helix bundle shown in conformation $b$ in Fig. 4 is the ground state in the green phase whereas the ground state conformation in the magenta phase has a slightly different arrangement of helices. The white region in the left of the phase diagram has large attractive values of $e_W$, and the ground state conformations are compact globular structures with a crystalline order induced by hard sphere packing considerations (44) and not by hydrogen bonding (conformation $i$ in Fig. 4).

---

**Fig. 4.** MOLSCRIPT representation of the most common structures obtained in our simulations. Helices and strands are assigned when local or nonlocal hydrogen bonds are formed according to the described rules. Conformations $a$, $b$, $c$, $d$, $e$, $f$, and $g$ are competitive local minima. Conformation $i$ is that of a generic compact polymer chain, obtained by switching off hydrogen bonds, the tube constraint, and curvature energy penalty, and is obtained on maximizing the total number of hydrophobic contacts.

---

Hoang et al.
energy, dehydrophobic contacts
conformations from each of the minima are shown in the
Fig. 5. Contour plots of the effective free energy at high temperature (T = 0.22) and at the folding transition temperature T = 0.2. The effective free energy, defined as $\ln(P_{\text{N} + N_0, N}) = -\ln(P_{\text{N} + N_0, N})$ collected in equilibrium Monte-Carlo simulations at constant temperature. The spacing between consecutive levels in each contour plot is 1 and corresponds to a free energy difference of $kT$, where $T$ is the temperature in physical units. The darker the color, the lower the free-energy value. There is just one free-energy minimum corresponding to the denatured state at a temperature higher than the folding transition temperature (a) whereas one can discern the existence of three distinct minima at the folding transition temperature (b). Typical conformations from each of the minima are shown in the figure.

In summary, within a simple, yet realistic, framework, we have shown that protein native-state 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. The limited number of folds arises from the geometrical constraints imposed by steric and hydrogen bonds. In the marginally compact phase, not only does one have a space-filling conformation (the nearby backbone segments have to be placed near each other to avail of the attractive potential), which is effective in expelling water from the hydrophobic core, but also these segments need to have the right orientation with respect to each other to respect the geometrical constraints imposed by the hydrogen bonds.

We thank Buzz Baldwin, Huc-Sun Chan, Morrel Cohen, Russ Doolittle, Davide Marenduzzo, George Rose, and Harold Scheraga for their invaluable comments. This work was supported by Progetto Di Rilevante Interesse Nazionale 2003, Fondo Integrativo Speciale Ricerca 2001, the National Aeronautics and Space Administration, National Science Foundation (NSF) Integrative Graduate Education and Research Traineeship Grant DGE-9987589, NSF Materials Research Science and Engineering Centers, and the award of a postdoctoral fellowship at the Abdus Salam International Center for Theoretical Physics (to T.X.H.).