PERSPECTIVE

What Determines the Spectrum of Protein Native State Structures?

Timothy R. Lezon,[1] Jayanth R. Banavar,[1] Arthur M. Lesk,[2]* and Amos Maritan[3]
[1] *Department of Physics, The Pennsylvania State University, University Park, Pennsylvania*
[2] *Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania*
[3] *Dipartimento di Fisica "G. Galilei" and INFN, Università di Padova, Padova, Italy*

ABSTRACT **We present a brief summary of the key factors underlying protein structure, as developed in the investigations of Pauling, Ramachandran, and Rose. We then outline a simplified physical model of proteins that focusses on geometry and symmetry. Although this model superficially appears unrelated to the detailed chemical descriptions commonly applied to proteins, we show that it captures the essential elements of the chemistry and provides a unified framework for understanding the common characteristics of folded proteins. We suggest that the spectrum of protein native state structures is determined by geometry and symmetry and the role of the sequence is to choose its native state structure from this predetermined menu. Proteins 2006;63:273–277.**

Key words: folded proteins; native states; structural motifs

INTRODUCTION

Our understanding of the behavior of proteins has consistently made progress. However, although many general features of folding—such as burial of hydrophobic groups and structural motifs of native states—have been known for decades, many details remain unclear. Until recently, the problem of protein structure has been addressed through fully detailed models, involving all atoms and explicit charges. However, simplified models based on geometry and symmetry have shown themselves capable of rationalizing the nature of observed structures. The purpose of this essay is to illuminate the similarities and complementarity of these approaches.

INFERENCES ABOUT PROTEIN STRUCTURES DERIVED FROM STRUCTURAL CHEMISTRY

More than 50 years ago, Linus Pauling and his collaborators made a seminal advance in understanding protein structure based on an application of quantum chemistry and crystallography.[1,2] The key prediction, accurately confirmed by experiment, was that α-helices and β-sheets were the repeatable structures of choice for which backbone hydrogen bonds would provide the scaffolding. This ingenious result arose on considering the rules of quantum mechanics and details of the lengths and nature of covalent and hydrogen bonds. This result led to the idea that the "limited parts list"[3] of protein native state structures had helices and almost planar sheets as their principal components.

More than a decade later, Ramachandran and coworkers[4] studied the entirely different phenomena of sterics: the large energetic cost of the overlap of non-bonded atoms. This study was applicable not only to native state structures studded with backbone hydrogen bonds but also equally well to the denatured state of proteins, which is thought to have fewer hydrogen bonds between backbone atoms. The key discovery was that the allowed phase space, as described by the Ramachandran dihedral angles ϕ and ψ, was restricted by steric interactions. In other words, values of (ϕ, ψ) in only a few limited regions of phase space did not lead to steric overlaps. On incorporating the details of the structure of the peptide bond, the significantly populated allowed regions of the phase space correspond to the now familiar α-helix and β-strand. Indeed, the two backbone geometries that allow for systematic and extensive hydrogen bonding[1,2] are the α-helix and the β-sheet obtained by a repetition of the backbone dihedral angles from the two regions respectively.[5]

The sequence of amino acids plays an all-important role in the choice of the local structure. Poly-L-alanine, which is a good approximation to the backbone, readily forms a helix in water,[6] but for heterogeneous side-chains the helix backbone sterically clashes with some side-chain conform-

Grant sponsor: National Aeronautical and Space Administration; Grant sponsor: Penn State MRSEC; Grant sponsor: INFM; Grant sponsor: National Science Foundation IGERT grant; Grant number: DGE-9987589.

*Correspondence to: Arthur M. Lesk, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 108 Althouse Laboratory, University Park, PA 16802. E-mail: aml@mrc-lmb.cam.ac.uk

Received 26 July 2005; Revised 22 September 2005; Accepted 31 October 2005

Published online 9 February 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20869

ers, resulting in a loss of conformational entropy.[7] When the price in side-chain entropy is too large, an extended backbone conformation results pushing the segment toward a β-strand structure.[5]

Rose and his collaborators have built on the work of Pauling and Ramachandran in innovative and important ways. One key advance is the recognition of the hierarchical organization of proteins[8] and the idea of hierarchical folding initiated locally along the sequence and proceeding with the subsequent interaction of these marginally stable units to yield more complex folded regions and so on until the native state structure is obtained.[5,9] Another is the vivid demonstration[10,11] based on sterics that the Flory isolated-pair hypothesis,[12] which posits that the Ramachandran angles at successive positions along the sequence are independent of each other, breaks down in a nontrivial way so as to eliminate hybrid conformations of α-helices and β-strands. Fitzkee and Rose have observed[3] that this result is also consistent with the fact that the combination of Pauling's hydrogen bonds and Ramachandran's steric effects limit local backbone structures to a small spectrum of possible conformations.

The problem of determining the native state structure of a protein can be thought of as one of assembly from the list of available parts to create a harmonious whole in which the hydrogen bonds are in place, steric clashes are avoided, a hydrophobic core is created, and charged amino acids are preferentially in contact with the solvent. Of equal importance, even the structures in the denatured state are *not* featureless. The typical local structures in the denatured state are those that do not suffer from steric clashes and include α-helices, β-strands and polyproline II helices.[13–15]

The picture that emerges from the work of Pauling, Ramachandran, Rose and others encapsulates the essential structural chemistry underlying proteins and explains why protein native states are assemblages of α-helices and β-sheets and why there are a limited number of folds.

THE TUBE MODEL

The success of structural chemistry in rationalizing the role of α-helices and β-sheets[1,2] as the building blocks of protein structures has spurred investigations of proteins based on detailed models. In contrast, recent work[16,17] has suggested a unification of the various aspects of all proteins: symmetry and geometry determine the limited spectrum of folded conformations that a protein can choose from for its native state structure; these structures are in a marginally compact phase in the vicinity of a phase transition and therefore respond sensitively to certain perturbations; proteins are well-designed sequences of amino acids that fit well into one of these predetermined folds; and proteins are prone to misfolding and aggregation leading to the formation of amyloids, which are implicated in debilitating human diseases such as Alzheimers, type II diabetes, and spongiform encephalopathies.

We summarize the key ideas underlying this approach:

1. The symmetry of a system often plays a crucial role in determining its ordering. For example, a system of hard spheres exhibits either an isotropic fluid phase or an ordered crystalline phase depending on the packing fraction or sphere density. On replacing the isotropic spheres with anisotropic objects shaped like pencils, one can additionally obtain liquid crystal phases with translational order in fewer than three dimensions along with orientational order. It is therefore useful to consider the symmetry properties of protein chains. Like other linear polymers, polypeptide chains of proteins possess an inherent anisotropy—at each monomer (or amino acid) there is a special local direction. From the point of view of symmetry, therefore, a chain of *isotropic* spheres is not an appropriate simple model for a chain molecule. Instead, the simplest geometrically accurate model, which respects the correct symmetry, is that of a chain of coins or discs with the direction perpendicular to the face of the coin defining the local tangent. In the continuum limit, such a chain is akin to a flexible tube or the familiar garden hose. Physically, one requires that tube conformations be self-avoiding and this can be ensured by means of a suitable three-body interaction.[18]
2. One may next consider the nature of the ground states (at zero temperature) of a flexible tube subject to a self-attraction mimicking the effects of hydrophobicity. In other words, what are the classes of self-avoiding tube conformations that are best able to avail themselves of the attractive interactions? The answer to this question depends crucially on the value of a key dimensionless parameter:[19] the ratio of the tube thickness and the range of attractive interaction. When this ratio is close to unity, one obtains a marginally compact phase that is in the vicinity of a phase transition between the swollen phase (obtained for values of the ratio large compared to 1) and a generic compact phase (obtained for values of the ratio small compared to 1). Quite remarkably, the ratio is close to unity for proteins and so they are naturally poised in the marginally compact phase. This fine-tuning of the ratio is accomplished automatically in proteins because the side chains of amino acids determine both the effective tube radius (the side chains reside in the space within the tube) and the range of attractive interactions (the outer atoms of nearby side-chains have a short-range attraction screened by the surrounding water). For short tubes, there are very few marginally compact tube structures, most notably a space-filling helix[20,21] with a pitch to radius ratio within a few percent of that observed in real proteins and zig-zag strands assembled into almost planar sheets.[22] The proximity of the marginally compact phase to a phase transition provides exquisite sensitivity to these structures to the right types of perturbations and the fact that there are few structures lends itself to a simple energy landscape with few minima.
3. When one deals with unconstrained objects, it is sufficient to specify where the objects are located and using this information, one can construct mutual distances between pairs of objects. In contrast, when objects are

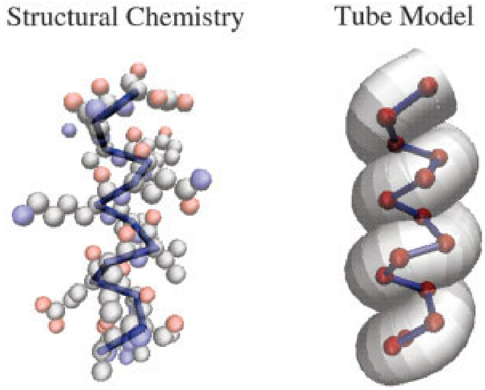

Fig. 1. Key differences between the structural chemistry and tube model approaches to understanding protein structure. At left is an all-atom representation of a helix, with the backbone darkened for clarity. On the right is the same helix represented as a flexible tube of radius 2.7 Å. The conventional chemistry-based picture typically models all atoms in the protein to accurately reproduce structure. Atomic coordinates for such a model are determined from known values of bond lengths and angles, which in turn result from the quantum mechanical properties of the atoms. The Hamiltonian for such a model often consists of additive pairwise interactions that are derived from known physical properties, such as hydrogen bonds, van der Waals, and electrostatic interactions. In the all-atom model backbone conformations are restricted by the ranges of the dihedral angles ϕ and ψ, and the particulars of the amino acid sequence determine protein structure. In contrast, the tube model is coarse-grained, such that only the locations of the C_α atoms are given attention. Each C_α is taken to reside on the axis of a flexible tube of nonzero thickness. The Hamiltonian here consists of an attractive potential between amino acids that mimics the effects of hydrophobicity, plus a three-body repulsive term that ensures self-avoidance of the tube. This three-body term provides the only limitation on the nature of the local backbone conformation and, as shown in Figure 2, excludes regions of the ϕ-ψ space that are sterically inaccessible in all-atom models. The only conformational constraint in this model is the fixed length of the virtual bonds between successive C_α atoms, and it is the form of the Hamiltonian that encourages the formation of protein-like structures. In the tube model, the details of the amino acid sequence play the secondary role of selecting the best fit geometrically allowed native state from a menu of possibilities determined by geometry and symmetry.

tethered along a chain, it is insufficient to describe a constituent object merely by providing its location, as if it were an isolated entity. Rather, associated with each object is a local coordinate system—for example, a Cartesian system whose three axes are the tangent to the chain, the normal to the chain and the binormal at that location. Thus for a chain molecule such as a protein, a more complete description is provided by the location as well as the specification of the local coordinate system at each amino acid location. Interestingly, a study of experimentally determined protein structures shows that sterics and hydrogen bonds conspire to place constraints on the relative orientation of the local coordinate systems of amino acids connected by covalent or hydrogen bonds. There is a great simplification associated with the fact that a vast majority of these geometrical constraints are independent of the identity of the amino acids involved in the bonding.[16,17]

4. A refined tube model of a homopolymer with these constraints along with a local bending energy penalty term, an overall hydrophobicity term and simple energy scores for local and nonlocal (along the sequence) hydrogen bonds as well as a reward for formation of cooperative hydrogen bonds leads to a surprisingly simple result. One finds, *even for a homopolymer*, that the ground states in the marginally compact phase (i.e., in the vicinity of a transition to a swollen phase) are assembled tertiary structures resembling protein native state structures. One obtains distinct assembled folds as the ground state on varying the overall hydrophobicity and/or the bending energy parameters.
5. In summary, this model suggests that the phase of matter employed by Nature to house protein structures is the marginally compact tube phase because of its advantages—a simple funnel-like energy landscape even at the homopolymer level with few minima and the sensitivity of structures because of their being marginally compact. The role of the sequence is not to *fashion* its native state structure but rather to *choose* it from

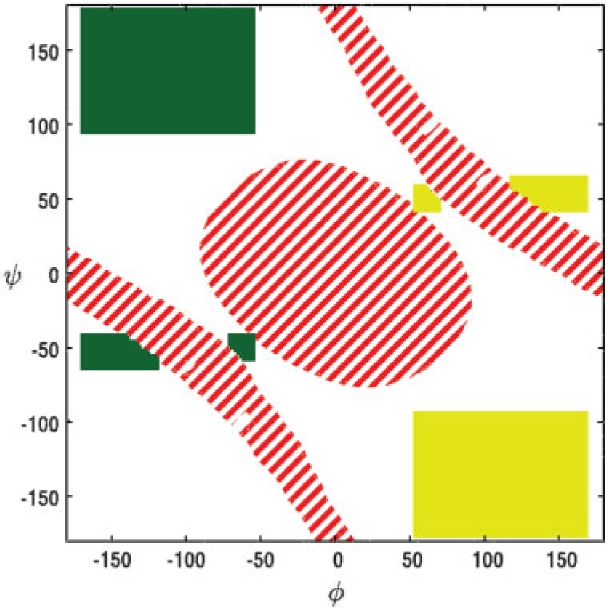

Fig. 2. Ramachandran plot showing the accessible and forbidden regions based on sterics and the tube constraint, respectively. Structurally repetitive peptides (i.e., those for which ϕ and ψ angles are identical for all amino acid residues) that are 12 residues in length are modeled for values of ϕ and ψ spanning the phase space. The tube radius—the radius of the smallest circle that passes through any three C_α atoms—is calculated for each peptide. The (red) hatching indicates regions that are *forbidden* by the three-body constraint: The radius of one or more circles passing through triplets of C_α atoms becomes less than the tube radius taken to be 2.7 Å. The forbidden region in the center of the plot is excluded by local effects alone, such that any residue with a (ϕ, ψ) value in this region will have a local radius of curvature below 2.7 Å. The excluded strips in the upper right and lower left result from nonlocal effects; repetitive structures with (ϕ, ψ) values from these regions have acceptably large local radii of curvature and only violate the three-body constraint if they are extended beyond the tri-peptide. The regions that are sterically *accessible* to 12 residue polyalanine peptides in an all-atom representation are plotted as solid: Allowed regions for poly-L-alanine have values $\phi < 0$ and are plotted in green (dark), and regions accessible to poly-D-alanine have values $\phi > 0$ and are plotted in yellow (light). Note that the regions near the α and α_L helices that are forbidden by the tube model are also prohibited by steric interactions in the L and D isomers, respectively. It is also interesting that the ideal α-helix, situated around (−63°, −41°), resides precariously between two regions that are forbidden in the tube model.

the menu of folds predetermined by geometry and symmetry. In other words, the mechanism by which the sequence determines the native state structure involves the sequence directing a *choice* from the menu of folds predetermined by geometry and symmetry. Protein native state structures conform to the constraints imposed by the spectrum of possible folding patterns.[23] The evolution of sequences and functionalities is subject to the same constraints. Finally, the same ingredients that lead to the menu of predetermined folds lead also to the existence of ordered aggregates of many proteins,[16] which resemble amyloid fibrils.

RECONCILING THE APPROACHES

The approaches described in the previous sections are quite distinct from each other, yet have certain features in common (see Fig. 1). The simple garden hose captures the sterics approximately in at least three ways: the self-avoidance of the tube; the constraint that the local radius of curvature can be no smaller than the tube radius; and the notion that the space within a tube can be thought of as "wiggle room" for the amino acid side chains. The directionality of the hydrogen bonds is encapsulated by the inherent anisotropy of a tube—in a close-packed arrangement, nearby tube segments are preferentially parallel and track each other rather than being perpendicular and progressively separating from each other. This is found to be the case in both helices and sheets. Furthermore, the requirement that nearby tube segments in the folded state of a protein in the marginally compact phase be placed alongside and parallel to each other promotes an all-or-nothing folding transition characteristic of a two-state protein.

Although the effects of steric constraints lead to the Ramachandran thinning of phase space at the single amino acid level and the Rose grammar at a slightly extended local level, the tube is different in several respects (see Fig. 2). First, enantiomorphic conformations give the same radius of curvature, so that the rejection of conformations inaccessible to L-amino acids but accessible to R-amino acids is a separate constraint. The preference for optimally compact conformations is common to the all-atom models, in which the preference arises from a combination of hydrophobicity and optimization of Van der Waals interactions, and the tube model. The full thinning out of phase space in the tube model, however, occurs only in the presence of an attraction-promoting compaction (in the absence of such an attraction, the dominant conformations are not compact). Furthermore, *nonlocal* effects imposed by the tube geometry and anisotropy play a key role in determining the optimal structures (a model in which one simply imposes a local radius of curvature constraint does not capture the physics of a flexible tube completely). The refined tube model is even more direct in incorporating the *amino acid aspecific* geometrical constraints arising from the effects of hydrogen bonds and sterics (aspecific except for proline residues); yet, it is remarkable that it yields assembled protein-like structures even for homopolymers.

Consider the sodium chloride structure adopted by ionic crystals such as NaCl, LiCl, KBr, and AgCl. The NaCl structure is a face-centered-cubic (fcc) arrangement for the Cl ions with the sodium ions occupying the octahedral holes. One can do a very careful quantum mechanical calculation and show that this fcc structure arises from considerations of electrovalent bonding. Alternatively, following the pioneering conjecture of Kepler[24] recently proved by Hales,[25] or the everyday experience of grocers, one may argue that a collection of spherical cannonballs or apples are best packed in a fcc lattice. One may then be emboldened to suggest that it is considerations of packing, periodicity, and the correct symmetry (note that a packing of cubes instead of spheres would not lead to a fcc lattice) that are the essential ingredients that determine the menu of possible crystal structures. In other words, the essential elements underlying the fcc structure are not the details of the interatomic interactions or even the quantum mechanics that describes the interactions of all matter but rather the Platonic considerations of geometry and symmetry. It is of course remarkable that Nature has found such a perfect fit between the quantum interactions in NaCl and the fcc structure. The key point is that the structure transcends the chemical housed in it and is determined by the overarching constraints of geometry and symmetry. That many protein sequences adopt the same fold and that the menu of possible folds is limited strongly suggest that similar considerations may be at play here as well even though proteins are neither infinite in extent nor periodic. The close packing of a flexible tube *in the marginally compact phase* is then the analog of the grocer's packing of apples for this problem.

ACKNOWLEDGMENTS

We are grateful to our collaborators, especially Trinh Hoang, Flavio Seno, and Antonio Trovato, for being our partners. George Rose has been our friend and teacher. We are deeply indebted to him and send him and Judy our warmest wishes.

REFERENCES

1. Pauling L, Corey RB, Branson HR. The structure of proteins: two hydrogen-bonded helical conformations of the polypeptide chain. Proc Natl Acad Sci USA 1951;37:205–211.
2. Pauling L, Corey RB. Conformations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. Proc Natl Acad Sci USA 1951;37:729–740.
3. Fitzkee NC, Fleming PJ, Gong H, Panasik N Jr, Street TO, Rose GD. Are proteins made from a limited parts list? Trends Biochem Sci 2005;30:73–80.
4. Ramachandran GN, Sasisekharan V. Conformation of polypeptides and proteins. Adv Prot Chem 1968;23:283–438.
5. Baldwin RL, Rose GD. Is protein folding hierarchic? I. Local structure and peptide folding. Trends Biochem Sci 1999;24:26–33.
6. Marqusee S, Robbins VH, Baldwin RL. Unusually stable helix formation in short alanine-based peptides. Proc Natl Acad Sci USA 1989;86:5286–5290.
7. Creamer TP, Rose GD. Side-chain entropy opposes α-helix formation but rationalizes experimentally determined helix-forming propensities. Proc Natl Acad Sci USA 1992;89:5937–5941.
8. Rose GD. Hierarchic organization of domains in globular proteins. J Mol Biol 1979;134:447–470.
9. Baldwin RL, Rose GD. Is protein folding hierarchic? II. Folding intermediates and transition states. Trends Biochem Sci 1999;24:77–83.

10. Fitzkee NC, Rose GD. Steric restrictions in protein folding: an α-helix cannot be followed by a contiguous β-strand. Protein Sci 2004;13:633–639.
11. Pappu RV, Srinivasan R, Rose GD. The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding. Proc Natl Acad Sci USA 2000;23:12565–12570.
12. Flory PJ. Statistical Mechanics of Chain Molecules; New York: Wiley; 1969.
13. Shi ZS, Olson CA, Rose GD, Baldwin RL, Kallenbach NR. Polyproline II structure in a sequence of seven alanine residues. Proc Natl Acad Sci USA 2002;99:9190–9195.
14. Pappu RV, Rose GD. A simple model for polyproline II structure in unfolded states of alanine-based peptides. Protein Sci 2002;11: 2437–2581.
15. Mezei M, Fleming PJ, Srinivasan R, Rose GD. Polyproline II helix is the preferred conformation for unfolded polyalanine in water. Proteins 2004;55:502–507.
16. Banavar JR, Hoang TX, Maritan A, Seno F, Trovato A. A unified perspective on proteins—a physics approach. Phys Rev E 2004;70: Art No 041905.
17. Hoang TX, Trovato A, Seno F, Banavar JR, Maritan A. Geometry and symmetry presculpt the free-energy landscape of proteins. Proc Natl Acad Sci USA 2004;101:7960–7964.
18. Banavar JR, Gonzalez O, Maddocks JH, Maritan A. Self-interactions of strands and sheets. J Stat Phys 2003;110:35–50.
19. Banavar JR, Maritan A, Micheletti C, Trovato A. Geometry and physics of proteins. Proteins 2002;47:315–322.
20. Maritan A, Micheletti C, Trovato A, Banavar JR. Optimal shapes of compact strings. Nature 2000;406:287–290.
21. Snir Y, Kamien RD. Entropically driven helix formation. Science 2005;307:1067.
22. Pappu RV, Hart RK, Ponder JW. Analysis and application of potential energy smoothing and search methods for global optimization. J Phys Chem B 1998;102:9725.
23. Denton M, Marshall C. Laws of form revisited. Nature 2001;410: 417.
24. Szpiro GG. Kepler's Conjecture; New York: John Wiley; 2003.
25. Sloane NJA. Kepler's conjecture confirmed. Nature 1998;395:435–436.