Secondary structure determines protein topology

PATRICK J. FLEMING, HAIPENG GONG, AND GEORGE D. ROSE

T.C. Jenkins Department of Biophysics, Johns Hopkins University, Baltimore, Maryland 21218, USA

(RECEIVED April 24, 2006; FINAL REVISION April 24, 2006; ACCEPTED May 18, 2006)

Abstract

Using a test set of 13 small, compact proteins, we demonstrate that a remarkably simple protocol can capture native topology from secondary structure information alone, in the absence of long-range interactions. It has been a long-standing open question whether such information is sufficient to determine a protein's fold. Indeed, even the far simpler problem of reconstructing the three-dimensional structure of a protein from its exact backbone torsion angles has remained a difficult challenge owing to the small, but cumulative, deviations from ideality in backbone planarity, which, if ignored, cause large errors in structure. As a familiar example, a small change in an elbow angle causes a large displacement at the end of your arm; the longer the arm, the larger the displacement. Here, correct secondary structure assignments (α -helix, β -strand, β -turn, polyproline II, coil) were used to constrain polypeptide backbone chains devoid of side chains, and the most stable folded conformations were determined, using Monte Carlo simulation. Just three terms were used to assess stability: molecular compaction, steric exclusion, and hydrogen bonding. For nine of the 13 proteins, this protocol restricts the main chain to a surprisingly small number of energetically favorable topologies, with the native one prominent among them.

Keywords: protein topology; protein folding; secondary structure; hydrogen bonding; confinement

Supplemental material: see www.proteinscience.org

We have been investigating the degree to which secondary structure alone determines a protein's overall topology under folding conditions (Przytycka et al. 1999; Gong and Rose 2005; Gong et al. 2005). Interest in this topic is motivated by both theoretical and practical concerns, the former stemming from inquiry into the mechanism of protein folding (Fitzkee et al. 2005a), and the latter from on-going success in secondary structure prediction. Protein topology is defined as the three-dimensional path described by the main chain. The exquisite packing that is a hallmark of folded proteins (Richards 1977) surely involves long-range interactions among side chains (Frieden 2003; Bradley et al. 2005). However, the focus here is on local interactions. Specif-

ically, we seek to determine the degree to which gross backbone topology and detailed packing are independent of one another. Simply put, is secondary structure alone sufficient to determine protein conformation at the molecular cartoon level (i.e., 5–6 Å level)?

The proposition that secondary structure determines tertiary structure often provokes an argumentative response: Some think it true but trivial; others are sure it is demonstrably false. The *true-but-trivial* group believes that little conformational latitude remains once α -helices, β -strands, and β -turns are fixed. Accordingly, it is important to realize that rebuilding a protein in three-dimensions is actually a daunting task, even when starting from exact knowledge of backbone torsion angles (Holmes and Tsai 2004; Gong et al. 2005). Our level of prior structural knowledge is less restricted yet. Backbone torsion angles are not assumed, only the four broad secondary structure categories: α -helix, β -strand, β -turns, polyproline II (P_{II}), or coil. The nontrivial nature of our approach is assessed via several controls, each of which

---

Reprint requests to: George D. Rose, T.C. Jenkins Department of Biophysics, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218; e-mail: grose@jhu.edu; fax: (410) 516-4118.

Article published online ahead of print. Article and publication date are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.062305106>.

omits one or more crucial aspects of the method. In contrast to the *true-but-trivial* group, the *demonstrably-false* group asserts that a given set of secondary structure elements can be interconnected in multiple ways, so we hasten to emphasize that peptide chain turns (β -turns) and P_{II} were included in our secondary structure definitions.

Our approach is computational: using Monte Carlo simulation, an extended polypeptide backbone chain was allowed to fold spontaneously under the influence of three simple energy terms: molecular compaction, steric exclusion, and hydrogen bonding. The first two terms are generic. Under folding conditions, water exerts a solvent-squeezing force on protein molecules (Liu and Bolen 1995), favoring compaction but limited by steric exclusion (Ramachandran et al. 1963). Only the third term—hydrogen-bonding between backbone donors and acceptors (Fleming and Rose 2005)—differs from one protein to another. Prior to simulation, secondary structure assignments— α -helix, β -strand, β -turns, and P_{II} —were extracted from 13 proteins of known structure. During the simulation, these assignments were used to constrain the main chain to corresponding broad regions of conformational space. All residues not included in these four secondary structure categories were classified as coil and allowed to sample all sterically allowed regions.

Anfinsen's thermodynamic hypothesis, that the native conformation is determined by the amino acid sequence (Anfinsen 1973), does not imply a folding mechanism—deliberately so. Nevertheless, the hypothesis has often been interpreted to mean that tertiary structure arises as a consequence of detailed side chain interactions in the folded state. Instead, the results presented here suggest a hierarchic folding mechanism (Rose 1979; Baldwin and Rose 1999a,b) in which secondary structure biases are established locally, leading iteratively to further, mutually stabilizing interactions and resulting ultimately in native topology. However, this process can be obscured by the overall coil-globule collapse of the protein (Sosnick et al. 1996) under suitable experimental conditions.

Materials and methods

The data set included 13 X-ray-elucidated protein domains having <100 residues, small axial ratios, and diverse secondary structure compositions (Table 1). Each protein was modeled as a polyalanine chain except at sequence positions with glycine or proline residues, which were retained. Backbone hydrogen atoms were included; other hydrogens were omitted.

Table 1. Structural clusters from each protein simulation ensemble

PDB code	Class ^a	Length (residues)	% Coil ^b	Cluster ID	Cluster size	% <6.0 Å RMSD ^c	Contains correct topology ^d
1vii	α	36	42	I	288	72.9	+
				II	44	48.8	+
1pgb	α/β	56	27	I	307	54.6	+
				II	50	0	—
1r69	α	63	27	I	100	0	+
				II	84	0	+
1ueb	β	63	24	I	186	68.6	+
				II	74	0	—
1cuk	β	64	22	I	366	57.5	+
1c9o	β	66	33	I	324	2.8	+
1aye	α	71	22	I	312	40.2	+
				II	56	0	—
1ubq	α/β	76	12	I	258	3.1	+
				II	104	0	—
1h75	α/β	76	18	I	274	3.7	+
1tig	α/β	88	16	I	124	21.1	+
				II	54	0	—
1gl4	β	89	13	I	212	0	+
1epf	β	92	21	I	128	0	+
				II	60	0	—
1htt	α/β	99	19	I	106	3.8	+

For each protein, 400 independent simulations were done and the minimum energy conformer from each simulation was saved. The saved conformers were clustered by structure as described in the text. Only clusters with at least 60% structural similarities and containing >40 conformers are included in the table. Data for the largest cluster in each case are in bold. Proteins are listed in order of increasing size.

^aStructural class as defined in the ASTRAL database (Chandonia et al. 2004).

^bPercentage of residues designated as coil during the simulations.

^cPercentage of conformers in cluster <6 Å RMSD compared to the native structure.

^dAs determined by visual inspection of the chain tracing in ribbon diagrams (see Figs. 2, 3).

Monte Carlo simulations (Metropolis et al. 1953) were performed using the move set and Metropolis criteria described next. Backbone torsion angles, ϕ and ψ , were allowed to sample values randomly within a range imposed by each residue's secondary structure assignment (see below). The peptide bond torsion angle, ω , was varied randomly within the range $180^\circ \pm 5^\circ$. For each trial of ϕ , ψ , and ω angles, a residue was chosen at random and moved individually, except in the case of β -turns (see below), where backbone torsion angles for the two central residues ($i + 1$, $i + 2$) (Rose et al. 1985) were varied concomitantly.

Within this protocol, hydrogen bonds are not “wired-in.” Rather, they are allowed to form or break spontaneously between any backbone donor and acceptor, subject to the Metropolis criterion. The hydrogen bond scoring function employed both distance and orientation criteria previously described (Fleming et al. 2005), with a maximum favorable score at a heavy atom donor-acceptor distance of ≤ 3.5 Å, decaying linearly to zero at a distance of 5 Å. A confinement score was devised to disfavor conformations with a geometric radius of gyration (R_g) larger than that predicted for a globular protein with the same number of residues ($R_{g\text{-glob}}$),

$$\text{Confinement score} = \begin{cases} 0, & R_g \leq R_{g\text{-glob}} \\ (R_g - R_{g\text{-glob}})^2, & R_g > R_{g\text{-glob}} \end{cases}$$

where $R_{g\text{-glob}} = 2.5 \times N^{0.34}$, and the prefactor was decreased slightly from that previously reported (Gong et al. 2005) to favor proteins with smaller than average R_g .

Secondary structure assignments for each protein were based solely on backbone torsion angles as previously described (Srinivasan and Rose 1999). Five standard categories were defined: extended (E), helix (H), turn (T), polyproline II (P), and coil (C), with turns subdivided into the six specific β -turn types: I, I', II, II', III, and III' (Rose et al. 1985). Every protein residue was assigned to one of these broad categories, and during Monte Carlo simulation, that residue's trial backbone torsion angles were chosen at random from the corresponding region in ϕ, ψ -space. Residue-specific coil constraints were taken from observed distributions in the protein coil library (<http://roslab.jhu.edu/coil/>) (Fitzkee et al. 2005b). ϕ, ψ -Maps of all backbone torsion angle categories can be found at <http://roslab.jhu.edu/movesets/> and in the Supplemental Material.

An initial simulation of 10,000 cycles (a cycle is $N - 2$ Monte Carlo moves, where N = the number of residues) was performed with only steric exclusion and a hydrogen bond score as the Metropolis criteria. The best scoring

structure from this initial simulation was further simulated for an additional 50,000 cycles, with the confinement score added to the Metropolis criteria. Control experiments were performed using steric exclusion alone or steric exclusion plus either confinement or hydrogen bonding, but not both; all simulations were constrained by their secondary structure assignments.

For each protein, 400 independent simulations were performed starting from an extended chain, and the conformer with the best combined hydrogen bond and confinement score was saved from each simulation. These 400 saved conformers were then clustered by structural similarity, with structure characterized by a “structure-vector” that included both the backbone torsion angles and all $C\alpha(i)$ to $C\alpha(i + 6, \dots, N)$ distances (i.e., an abbreviated $C\alpha$ -distance matrix). The $C\alpha$ -distance matrix component of this structure-vector was normalized by a function of N to balance its contribution against that of the torsion angle component. Hierarchic centroid linkage clustering of these vectors was performed with a modified version of Pycluster (de Hoon et al. 2004) after correcting for the periodicity of angular data.

All root mean squared difference (RMSD) values refer to $C\alpha$ backbone comparisons. Ribbon diagrams were made with PyMOL (DeLano 2003), and Ramachandran plots were made using Grace (<http://plasma-gate.weizmann.ac.il/Grace/>).

Results and Discussion

Controls

Three simulations that omit one or more energy terms were performed as controls: (1) steric exclusion only, (2) steric exclusion with confinement but no hydrogen bonding, and (3) steric exclusion with hydrogen bonding but no confinement. Ensembles obtained for these various combinations of incomplete scoring functions were generated for protein G (1pgb).

Control 1

Conformers satisfying secondary structure constraints and steric exclusion only are illustrated in Figure 1A. The ensemble resembles a random coil population (Fitzkee and Rose 2004), with a mean geometric radius of gyration, $\langle R_g \rangle$, of 24.7 ± 3.7 Å, similar to the value expected (22.4 Å) for a random coil in good solvent (de Gennes 1979). Helical segments are recognizable, but in the absence of hydrogen bonding, no β -hairpins or β -sheets were formed, and no conformers with native-like topologies were obtained, which is consistent with similar results reported earlier (Alexandrescu 2004). *This control is a vivid demonstration that secondary structure constraints alone are insufficient to capture the fold.*

Figure 1. (A) The protein G (1pgb) backbone simulated using only secondary structure constraints and steric exclusion. The native structure (*left*) and the five conformers with lowest RMSD to the native structure are shown; none of the five capture native topology. The overall population has a mean radius of gyration, $\langle R_g \rangle$, resembling that of unfolded proteins. Monte Carlo simulations were performed as described in the text with steric exclusion as the sole criterion for acceptance. (B) Conformations of the protein G backbone simulated as in A plus a confinement score. The native structure (*left*) and the five conformers with lowest RMSD to the native structure (RMSD of 5.0 Å–5.5 Å) are shown. Although compact, the population still lacks native topology. The confinement score is described in Materials and Methods. (C) Conformations of the protein G backbone simulated as in B plus a hydrogen bond score. The native structure (*left*) and five conformers with lowest normalized RMSD to the native structure (RMSD of 1.9 Å–3.2 Å) are shown. Conformers having native topology are now abundant. The hydrogen bond score is described in Materials and Methods.

Controls 2 and 3

The addition of a confinement score without hydrogen bonding selects for a compact population, illustrated in Figure 1B. A single, structurally disperse cluster was identified from this simulation, with $\langle R_g \rangle = 10.2 \pm 0.9$ Å. However, it is apparent that secondary structure together with compaction and steric exclusion is still insufficient to generate structures with correct topology. Conversely, addition of a hydrogen bonding score without confinement promotes formation of both α -helices and β -hairpins, but no β -sheet is formed, $\langle R_g \rangle$ resembles a random coil (21.6 ± 2.6 Å), and again the native-like topology is not observed (data not shown).

The combination of confinement and hydrogen bonding gives results that are dramatic. As seen in Figure 1C, the largest cluster includes many conformers with native topologies (Table 1). It should be emphasized that this result was achieved in the absence of side chain interactions; only secondary structure constraints and the three scoring functions were included.

Results for 13 proteins

Protein G is not unique. Simulations of eight additional proteins resulted in conformational ensembles that included native topologies, as seen in Figure 2. Table 1 summarizes results for all 13 proteins studied here. In

each case, only one or two major structural clusters were identified in the simulated conformational ensemble, with the native topology well represented in the larger cluster for at least nine proteins.

It is common practice to assess similarity between two structures by their RMSD, an imperfect measure that reduces a complex comparison involving hundreds of vectors to a single scalar number. We conform to this practice here by citing RMSD values, but with the following reservation. A small RMSD between two proteins indicates structural similarity, but the converse need not be true; two proteins can be topologically similar yet have a large RMSD between them.

In most proteins in Table 1, many conformers in the larger cluster have both native topology and RMSD < 6 Å (columns 6–7). Among the 13 proteins studied here, 10 have a population with RMSD ≤ 6 Å, but three fall in a different category, with RMSD > 6 Å for the most accurate conformer. These results demonstrate that secondary structure alone appears to be sufficient to define topology for small, single-domain proteins under folding conditions, although larger molecules with more complicated topologies may require inclusion of longer-range interactions. These results are consistent with our own previous fragment-assembly simulations (Gong et al. 2005) and with the fragment-assembly method of Chikenji et al. (2006) that included both local and long-range side chain interactions.

Figure 2. Proteins with correct topology and low RMSD scores. In each case, the native structure is on the *left*, and the structure with lowest RMSD from the largest cluster is on the *right*, labeled with its RMSD.

Four of the 13 test proteins resulted in less accurate topologies (Fig. 3). Although the helices in the all- α proteins, 1vii and 1r69, do describe the native chain path correctly, their rotational orientations would not engender a hydrophobic core if side chains were added (data not shown). This is to be expected because an α -helix can sample multiple rotational angles while maintaining the same overall orientation, and consequently, long-range interactions are probably required for further improve-

ment in these cases as well as in larger proteins with more complicated topologies.

Summary

Long-range interactions play an important role in determining the details of protein structure (Frieden 2003; Kihara 2005). But evidence is accumulating that local interactions are the primary determinant of secondary structure (Baldwin and Rose 1999a) and that secondary structure, in turn, delimits overall topology (Przytycka et al. 1999; Hoang et al. 2004; Gong et al. 2005). Even so, it is clear that approximate ϕ, ψ torsion angles alone are insufficient to determine tertiary folds (Alexandrescu 2004). Here, we use Monte Carlo simulations to show that correct secondary structure assignments (α -helix, β -strand, β -turn, polyproline II, and coil), together with global confinement and maximal hydrogen bonding, can capture the topology of a test set of small globular proteins in the absence of long-range interactions. Guided by these factors, the conformational search is narrowed significantly and the correct topology is sampled frequently, providing opportunities for additional stabilizing interactions, including those involving side chains. Very recently, Zhang et al. (2006) also concluded that a combination of compaction and hydrogen bonded elements of secondary structure limits the number of feasible folds.

Acknowledgments

We thank Lauren Perskie, Buzz Baldwin, Nicholas Fitzkee, and Timothy Street for much fruitful discussion; an anonymous referee for useful comments; and The Mathers Foundation for support.

Figure 3. Proteins with similar topology but poor accuracy. In each case, the native structure is on the *left*, and the structure with lowest RMSD from the largest cluster is on the *right*, labeled with its RMSD. For the two helical proteins, hydrogen bonds alone are insufficient to select the correct helix orientation; long-range hydrophobic interactions are probably required. The two larger proteins have more complicated topologies and may also suffer from omission of long-range interactions.

References

- Alexandrescu, A.T. 2004. Strategy for supplementing structure calculations using limited data with hydrophobic distance restraints. *Proteins* **56**: 117–129.
- Anfinsen, C.B. 1973. Principles that govern the folding of protein chains. *Science* **181**: 223–230.
- Baldwin, R.L. and Rose, G.D. 1999a. Is protein folding hierarchic? I: Local structure and peptide folding. *Trends Biochem. Sci.* **24**: 26–33.
- . 1999b. Is protein folding hierarchic? II: Folding intermediates and transition states. *Trends Biochem. Sci.* **24**: 77–83.
- Bradley, P., Misura, M.S., and Baker, D. 2005. Toward high-resolution de novo structure prediction for small proteins. *Science* **309**: 1868–1871.
- Chandonia, J.M., Hon, G., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M., and Brenner, S.E. 2004. The ASTRAL compendium in 2004. *Nucleic Acids Res.* **32**: D189–D192.
- Chikenji, G., Fujitsuka, Y., and Takada, S. 2006. Shaping up the protein folding funnel by local interaction: Lesson from a structure prediction study. *Proc. Natl. Acad. Sci.* **103**: 3141–3146.
- de Gennes, P.-G. 1979. *Scaling concepts in polymer physics*. Wiley, New York.
- de Hoon, M.J.L., Imoto, S., Nolan, J., and Miyano, S. 2004. Open source clustering software. *Bioinformatics* **20**: 1453–1454.
- DeLano, W. 2003. *The PyMOL molecular graphics system*. DeLano Scientific LLC, San Carlos, CA.
- Fitzkee, N.C. and Rose, G.D. 2004. Reassessing random-coil statistics in unfolded proteins. *Proc. Natl. Acad. Sci.* **101**: 12497–12502.
- Fitzkee, N.C., Fleming, P.J., Gong, H., Panasik Jr., N., Street, T.O., and Rose, G.D. 2005a. Are proteins made from a limited parts list? *Trends Biochem. Sci.* **30**: 73–80.
- Fitzkee, N.C., Fleming, P.J., and Rose, G.D. 2005b. The Protein Coil Library: A structural database of nonhelix, nonstrand fragments derived from the PDB. *Proteins* **58**: 852–854.
- Fleming, P.J. and Rose, G.D. 2005. Do all backbone polar groups in proteins form hydrogen bonds? *Protein Sci.* **14**: 1911–1917.
- Fleming, P.J., Fitzkee, N.C., Mezei, M., Srinivasan, R., and Rose, G.D. 2005. A novel method reveals that solvent water favors polyproline II over β -strand conformation in peptides and unfolded proteins: Conditional hydrophobic accessible surface area (CHASA). *Protein Sci.* **14**: 111–118.
- Frieden, C. 2003. The kinetics of side chain stabilization during protein folding. *Biochemistry* **42**: 12439–12446.
- Gong, H. and Rose, G.D. 2005. Does secondary structure determine tertiary structure in proteins? *Proteins* **61**: 338–343.
- Gong, H., Fleming, P.J., and Rose, G.D. 2005. Building native protein conformation from highly approximate backbone torsion angles. *Proc. Natl. Acad. Sci.* **102**: 16227–16232.
- Hoang, T.X., Trovato, A., Seno, F., Banavar, J.R., and Maritan, A. 2004. Geometry and symmetry prescript the free-energy landscape of proteins. *Proc. Natl. Acad. Sci.* **101**: 7960–7964.
- Holmes, J.B. and Tsai, J. 2004. Some fundamental aspects of building protein structures from fragment libraries. *Protein Sci.* **13**: 1636–1650.
- Kihara, D. 2005. The effect of long-range interactions on the secondary structure formation of proteins. *Protein Sci.* **14**: 1955–1963.
- Liu, Y. and Bolen, D.W. 1995. The peptide backbone plays a dominant role in protein stabilization by naturally occurring osmolytes. *Biochemistry* **34**: 12884–12891.
- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. 1953. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**: 1087–1092.
- Przytycka, T., Aurora, R., and Rose, G.D. 1999. A protein taxonomy based on secondary structure. *Nat. Struct. Biol.* **6**: 672–682.
- Ramachandran, G.N., Ramakrishnan, C., and Sasisekharan, V. 1963. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**: 95–99.
- Richards, F.M. 1977. Areas, volumes, packing, and protein structure. *Annu. Rev. Biophys. Bioeng.* **6**: 151–176.
- Rose, G.D. 1979. Hierarchic organization of domains in globular proteins. *J. Mol. Biol.* **134**: 447–470.
- Rose, G.D., Gierasch, L., and Smith, J.A. 1985. Turns in peptides and proteins. In *Advances in protein chemistry* pp. 1–109. Academic Press, New York.
- Sosnick, T.R., Mayne, L., and Englander, S.W. 1996. Molecular collapse: The rate-limiting step in two-state cytochrome c folding. *Proteins* **24**: 413–426.
- Srinivasan, R. and Rose, G.D. 1999. A physical basis for protein secondary structure. *Proc. Natl. Acad. Sci.* **96**: 14258–14263.
- Zhang, Y., Hubner, I.A., Arakaki, A.K., Shakhnovich, E., and Skolnick, J. 2006. On the origin and highly likely completeness of single-domain protein structures. *Proc. Natl. Acad. Sci.* **103**: 2605–2610.