Coevolution of Function and the Folding Landscape: Correlation with Density of Native Contacts

Ronald D. Hills Jr.* and Charles L. Brooks III*†

*Department of Molecular Biology and Kellogg School of Science and Technology, The Scripps Research Institute, La Jolla, California; and †Department of Chemistry and Biophysics Program, University of Michigan, Ann Arbor, Michigan

ABSTRACT The relationship between the folding landscape and function of evolved proteins is explored by comparison of the folding mechanisms for members of the flavodoxin fold. CheY, Spo0F, and NtrC have unrelated functions and low sequence homology but share an identical topology. Recent coarse-grained simulations show that their folding landscapes are uniquely tuned to properly suit their respective biological functions. Enhanced packing in Spo0F and its limited conformational dynamics compared to CheY or NtrC lead to frustration in its folding landscape. Simulation as well as experimental results correlate with the local density of native contacts for these and a sample of other proteins. In particular, protein regions of low contact density are observed to become structured late in folding; concomitantly, these dynamic regions are often involved in binding or conformational rearrangements of functional importance. These observations help to explain the widespread success of Gō-like coarse-grained models in reproducing protein dynamics.

Received for publication 29 July 2008 and in final form 11 August 2008.

Address reprint requests and inquiries to Charles L. Brooks III, Tel.: 734-647-6682; E-mail: brookscl@umich.edu.

Evolved proteins have a funneled folding landscape heavily biased toward the native structure. Hence, main-chain topology and interactions present in the native state dictate the mechanism by which a protein adopts its unique native fold (1). In addition to determining the folding mechanism, the three-dimensional structure of a protein determines its biological function, whether it entails ligand binding, catalysis, or conformational allostery. If structure can explain both folding and function, what then is the relationship between the folding landscape and function?

To address the relation of folding to function, we consider the success of simple coarse-grained models, known as Gō models, in describing folding. In the most common form, a Gō model represents a protein as a string of C_α-beads that interact via a Lennard-Jones-like potential in which only residue pairs that are in contact in the native state experience a pairwise attractive force. The result is a smooth and highly funnel-like energy landscape on which protein folding is a computationally tractable problem for even large proteins, unlike the case for all-atom simulations. Molecular dynamics trajectories can be followed from the unfolded state, which rapidly folds to the native structure, and the relative sequence of structure formation events can be monitored to glean insight into the mechanism of folding. Gō models have been successful in reproducing folding transition states and intermediates for a wide range of proteins (2–4). This success has been interpreted as resulting from the fact that folding landscapes are determined by native state topology and are evolved to be dominated by native interactions (1). Nonnative interactions, which cause energetic frustration in the landscape, play a minor role in the structure of transition states (3). Gō-like landscapes ignore nonnative interactions and lack energetic frustration but may contain topological frustration, which arises when native interactions form in the incorrect order.

Gō-like simulations were recently applied to compare topological frustration present in three members of the common flavodoxin fold (5,6). CheY, NtrC, and Spo0F are bacterial response regulators with unrelated functions and low sequence homology (~30% pairwise identity) but share an identical $\beta\alpha$-repeat topology (Fig. 1 *A*). To explore the role of sequence in folding the three proteins were studied using a ''flavored'' Gō model in which the interaction energies of side-chain native contacts were scaled according to their abundance in the Protein Data Bank. Phi-value experiments with CheY identified two folding subdomains: an N-terminal subdomain highly structured in the folding transition state, and a C-terminal subdomain unstructured in the transition state. The Gō simulations supported an N-terminal nucleation mechanism for the folding of CheY, NtrC, and Spo0F.

It has been observed that van der Waals contacts in CheY are weaker in the C-subdomain than the N-subdomain, resulting in flexibility in helix 4 of the C-subdomain that is important for function. Similarly, the Gō model assigned an average of 1.6/1.3 (CheY), 1.2/1.1 (NtrC), and 1.6/1.3 (Spo0F) native contacts per residue in the N-/C-subdomains, respectively. The folding mechanism that has emerged is that

Editor: Kathleen B. Hall.

doi: 10.1529/biophysj.108.143388

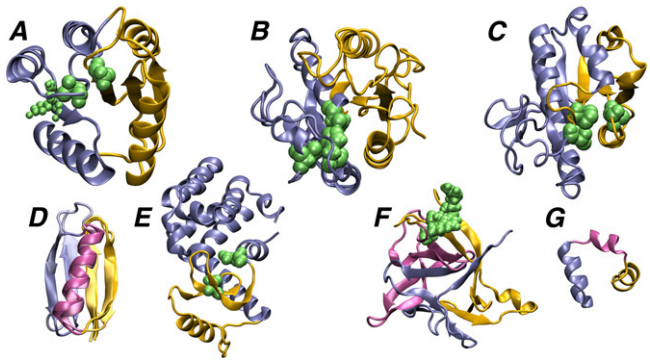

FIGURE 1 N-terminal (*yellow*), C-terminal (*blue*) and, where applicable, central (*magenta*) subdomains for CheY (*A*), cutinase (*B*), flavodoxin (*C*), protein L/G (*D*), T4L (*E*), IL-1β (*F*), and HP35 (*G*). Active site residues are shown in green.

the formation of the stable N-terminus is rate-limiting and serves to nucleate the weaker C-terminus. It is worth noting that while the model assigned different interaction strengths to each native contact to take into account sequence effects, contact energies were uniformly distributed throughout the proteins such that the number of contacts was a good predictor of folding events. In particular, the relative order of secondary structure formation observed within each protein was also well explained by the local contact density.

We express the contact density, which can be regarded as a measure of packing efficiency in the atomic protein structure, as the ratio of native contacts within some subset of the protein versus the number of residues comprising the subset. The thermodynamic basis for this ratio is as follows. The enthalpy of forming some element of structure is (neglecting small sequence-dependent variations) proportional to the number of native contacts assigned in the region, while the entropic cost of forming that structure from the denatured state is approximately proportional to the length of the peptide chain in the region expressed in number of residues. (Entropy is proportional to the number of rotatable torsion angles (7), and for large sequences, the average number of torsions per side chain approaches some constant.) An increased ratio of native contacts to residues is therefore predicted to have a more favorable free energy change associated with folding. Contact density has previously been used to explain downhill folding (8), association mechanisms in homodimers (9), and B-factors (10).

CheY, NtrC, and Spo0F contain 0.83, 0.88, and 1.06 contacts/residue, respectively, between helix 4 and strands 4 and 5. The enhanced contacts in Spo0F can be attributed to the introduction of bulkier residues into what is an alanine-rich cavity in CheY and NtrC. Stabilizing contacts in Spo0F's C-terminus resulted in a competition for van der Waals contacts between the N- and C-subdomains, causing topological frustration to be observed in the Gō landscape. Denatured states prematurely structured in the C-terminus were observed to unfold, or backtrack, before N-terminal folding could occur and proceed to the native state. Off-pathway frustration was minimized within the C-termini of CheY and NtrC. It turns out that frustration is correlated with the conformation dynamics of the three proteins. CheY and NtrC obey an allosteric population shift mechanism whereby a dynamic helix 4 and strand 5 are rigidified upon phosphorylation to bind their downstream targets. Spo0F, however, undergoes limited conformational rearrangement in its more rigid C-terminus, which does not participate in Spo0F's phosphorylation function. Evidently, an evolutionary relationship exists among protein stability, folding efficiency, and functional competence.

This correspondence between folding and function encouraged us to examine other proteins (Fig. 1). Phi-value experiments show flavodoxin folds via nucleation of its C-subdomain. Its C-subdomain has 1.5 contacts/residue compared to the N-subdomain's 1.3. Moreover, the flexible hydrophobic surface residues involved in nucleotide binding reside in the N-subdomain. The flavodoxin homolog cutinase contains flexible hydrophobic surface residues in the C-subdomain, which are responsible for its lipase activity. Concomitantly, its N- and C-subdomains have 1.6 and 1.4 contacts/residue, respectively. Contact density also correctly identifies the nucleating subdomains in the predominantly α-helical T4 lysozyme (T4L) and the β-trefoil protein interleukin-1β (IL-1β). Topological frustration due to subdomain competition has also been observed in T4L and IL-1β. For small proteins such as proteins L and G and HP35, the notion of subdomain is ill defined, and the local contact density is rather uniform throughout such that contact density is a poor predictor of folding order. References to experimental and theoretical characterization of the folding and functional dynamics of all proteins compared are given in Data S1 in the Supplementary Material.

Subdomain contact densities were compared to residue statistics. For the proteins compared, no correlation is observed between the contact density of a subdomain and the average volume (11) of its residues ($r = 0.02$). A weak correlation is seen between contact density and subdomain hydrophobicity (12) ($r = 0.43$). Contact density is well correlated ($r = -0.89$) with the compaction of the chain topology, expressed as the square of the subdomain radius of gyration, R_g, divided by the number of residues. Fig. 2 shows that chain compaction can be a predictor of folding order along with contact density when we define the compaction ratio as the square of the later folding subdomain's R_g normalized by its number of residues divided by the square of the nucleating subdomain's R_g normalized by its number of residues. The contact ratio is defined as the number of native contacts per residue in the nucleating subdomain divided by the native contacts per residue in the later folding subdomain. Given a protein of unknown function, one should be able to predict folding order and structurally dynamic regions if subdomains differ moderately in both contact density and chain compaction.

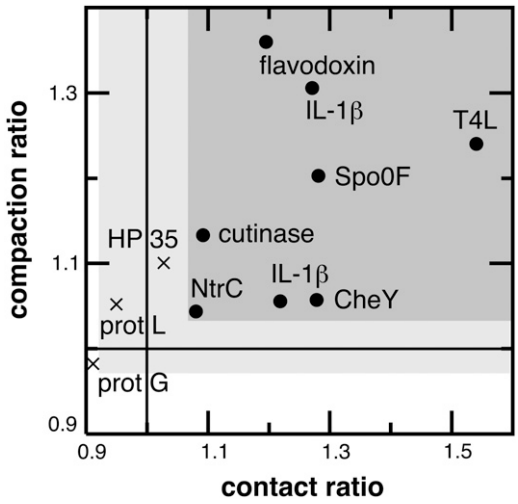

FIGURE 2 Subdomain folding order can be predicted from large (*dark shaded*) ratios of contact density and chain compaction (see text for definitions). Ratios <1.0 denote incorrect predictions. For small proteins (×), one or both predictors approach unity.

Evolution must arrive at a compromise between folding stability and the flexibility requisite for function (13,14). Folding frustration is observed in Gō simulations of the C-terminally stabilized CheY homolog Spo0F. Topological frustration offers a mechanistic basis for the experimental observation that stabilizing mutations decrease folding rates (15,16). Localized sites of frustration have been proposed as a means for predicting binding sites (17); conversely, functional sites have been implicated in frustration (18). The correlation of folding with local contact density provides an explanation for the success of even the simplest of folding models, such as analytical Ising models that solely consider contact enthalpy and chain entropy at the expense of a three-dimensional representation (19). Furthermore, the relation between contact density and protein dynamics helps to explain why Gō-like coarse-grained models (20,21) and elastic network normal mode analysis (22,23) are capable of unraveling the functional mechanisms behind evolution's molecular machines.

SUPPLEMENTARY MATERIAL

To view all of the supplemental files associated with this article, visit www.biophysj.org.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grant No. GM48807 and a La Jolla Interfaces in Science predoctoral fellowship to R.D.H.

REFERENCES and FOOTNOTES

1. Onuchic, J. N., and P. G. Wolynes. 2004. Theory of protein folding. *Curr. Opin. Struct. Biol.* 14:70–75.
2. Clementi, C., H. Nymeyer, and J. N. Onuchic. 2000. Topological and energetic factors: what determines the structural details of the transition state ensemble and ''en-route'' intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* 298:937–953.
3. Karanicolas, J., and C. L. Brooks III. 2003. Improved Gō-like models demonstrate the robustness of protein folding mechanisms towards non-native interactions. *J. Mol. Biol.* 334:309–325.
4. Lee, S. Y., Y. Fujitsuka, D. H. Kim, and S. Takada. 2004. Roles of physical interactions in determining protein folding mechanisms: molecular simulation of protein G and α-spectrin SH3. *Proteins.* 55:128–138.
5. Hills, R. D. Jr., and C. L. Brooks III. 2008. Subdomain competition, cooperativity and topological frustration in the folding of CheY. *J. Mol. Biol.* 382:485–495.
6. Hills, R. D. Jr., S. V. Kathuria, C. R. Matthews, and C. L. Brooks III. 2008. Sequence effects on folding in CheY-related proteins. *J. Mol. Biol.* Pending publication.
7. Ohkubo, Y. Z., and C. L. Brooks III. 2003. Exploring Flory's isolated-pair hypothesis: statistical mechanics of helix-coil transitions in polyalanine and the C-peptide from RNase A. *Proc. Natl. Acad. Sci. USA.* 100:13916–13921.
8. Zuo, G. H., J. Wang, and W. Wang. 2006. Folding with downhill behavior and low cooperativity of proteins. *Proteins.* 63:165–173.
9. Levy, Y., P. G. Wolynes, and J. N. Onuchic. 2004. Protein topology determines binding mechanism. *Proc. Natl. Acad. Sci. USA.* 101: 511–516.
10. Halle, B. 2002. Flexibility and packing in proteins. *Proc. Natl. Acad. Sci. USA.* 99:1274–1279.
11. Harpaz, Y., M. Gerstein, and C. Chothia. 1994. Volume changes on protein-folding. *Structure.* 2:641–649.
12. Pacios, L. F. 2001. Distinct molecular surfaces and hydrophobicity of amino acid residues in proteins. *J. Chem. Inf. Comput. Sci.* 41: 1427–1435.
13. Hubner, I. A., M. Oliveberg, and E. I. Shakhnovich. 2004. Simulation, experiment, and evolution: understanding nucleation in protein S6 folding. *Proc. Natl. Acad. Sci. USA.* 101:8354–8359.
14. Jager, M., Y. Zhang, J. Bieschke, H. Nguyen, M. Dendle, M. E. Bowman, J. P. Noel, M. Gruebele, and J. W. Kelly. 2006. Structure-function-folding relationship in a WW domain. *Proc. Natl. Acad. Sci. USA.* 103:10648–10653.
15. Kim, D. E., H. D. Gu, and D. Baker. 1998. The sequences of small proteins are not extensively optimized for rapid folding by natural selection. *Proc. Natl. Acad. Sci. USA.* 95:4982–4986.
16. Lopez-Hernandez, E., P. Cronet, L. Serrano, and V. Munoz. 1997. Folding kinetics of Che Y mutants with enhanced native α-helix propensities. *J. Mol. Biol.* 266:610–620.
17. Ferreiro, D. U., J. A. Hegler, E. A. Komives, and P. G. Wolynes. 2007. Localizing frustration in native proteins and protein assemblies. *Proc. Natl. Acad. Sci. USA.* 104:19819–19824.
18. Gosavi, S., P. C. Whitford, P. A. Jennings, and J. N. Onuchic. 2008. Extracting function from a β-trefoil folding motif. *Proc. Natl. Acad. Sci. USA.* 105:10384–10389.
19. Takada, S. 1999. Gō-ing for the prediction of protein folding mechanisms. *Proc. Natl. Acad. Sci. USA.* 96:11698–11700.
20. Koga, N., and S. Takada. 2006. Folding-based molecular simulations reveal mechanisms of the rotary motor F-1-ATPase. *Proc. Natl. Acad. Sci. USA.* 103:5367–5372.
21. Takagi, F., and M. Kikuchi. 2007. Structural change and nucleotide dissociation of myosin motor domain: dual Gō model simulation. *Biophys. J.* 93:3820–3827.
22. Lu, M. Y., and J. P. Ma. 2005. The role of shape in determining molecular motions. *Biophys. J.* 89:2395–2401.
23. Tama, F., and C. L. Brooks III. 2006. Symmetry, form, and shape: guiding principles for robustness in macromolecular machines. *Annu. Rev. Biophys. Biomol. Struct.* 35:115–133.