son and Beese, 2004). The same mispair, on the other hand, is poorly extended by Dpo4: it was shown to form a reverse wobble, which misaligns the 3′-end of the primer and therefore precludes nucleophilic attack (Trincao et al., 2004). This example clearly illustrates that molecular mechanisms, which lead to either polymerase stalling or bypass, should be studied in the context of the DNA sequence and polymerase alike. Let the (mis)match games begin.

Sylvie Doublié
Department of Microbiology and
Molecular Genetics
The University of Vermont
Burlington, Vermont 05405

Selected Reading

Brautigam, C.A., and Steitz, T.A. (1998). Curr. Opin. Struct. Biol. *8*, 54–63.

Doublié, S., Sawaya, M.R., and Ellenberger, T. (1999). Struct. Fold. Des. *7*, R31–R35.

Echols, H., and Goodman, M.F. (1991). Annu. Rev. Biochem. *60*, 477–511.

Hatahet, Z., Zhou, M., Reha-Krantz, L.J., Morrical, S.W., and Wallace, S.S. (1998). Proc. Natl. Acad. Sci. USA *95*, 8556–8561.

Hogg, M., Wallace, S.S., and Doublié, S. (2004). EMBO J. *23*, 1483–1493.

Johnson, S.J., and Beese, L.S. (2004). Cell *116*, 803–816.

Kool, E.T. (2002). Annu. Rev. Biochem. *71*, 191–219.

Krahn, J.M., Beard, W.A., and Wilson, S.H. (2004). Structure, this issue, 1823–1832.

Kunkel, T.A. (2004). J. Biol. Chem. *279*, 16895–16898.

Sawaya, M.R., Prasad, R., Wilson, S.H., Kraut, J., and Pelletier, H. (1997). Biochemistry *36*, 11205–11215.

Steitz, T.A. (1999). J. Biol. Chem. *274*, 17395–17398.

Trincao, J., Johnson, R.E., Wolfle, W.T., Escalante, C.R., Prakash, S., Prakash, L., and Aggarwal, A.K. (2004). Nat Struct Mol Biol *11*, 457–462.

Structure, Vol. 12, October, 2004, DOI 10.1016/j.str.2004.09.001

Protein Folding: Simple Models for a Complex Process

Twenty-eight years after its original publication, the diffusion-collision model has successfully been applied to describe the folding kinetics of two proteins with the same native structure but different sequences (Islam et al., 2004, this issue of *Structure*). The calculations show the relative importance of the primary and tertiary structure on the sequence of events and folding. For both proteins, the model suggests parallel folding pathways, a finding which has wide implications for the interpretations of experiments.

To be functional, proteins must fold into a particular three-dimensional structure (native state). Protein folding is a complex process involving noncovalent interactions throughout the entire molecule, many degrees of freedom, and a fine balance between enthalpic and entropic contributions to the free energy (Dill and Chan, 1997; Karplus, 2000). For single-domain proteins of less than 100 residues, it is known that those with predominantly α-helical structure in the native state fold faster than β-sheet proteins. Yet, a better understanding of the relative role of native state structure and amino acid sequence in the kinetics of protein folding is still needed. The improved understanding of the folding process will be of practical relevance to researchers in biotechnology and medicine because a large volume of predicted protein sequences obtained from the human (and other) genome projects require folding kinetic analysis before they can be put to use. Moreover, insights in the folding process will help elucidate events leading to misfolding and resulting pathologies such as prion diseases.

The paper by Islam and collaborators (2004) in this issue shows that the diffusion-collision model, a simple phenomenological model with a coarse-grained approximation of protein structure, correctly describes the sequence of events and folding kinetics of two 60 residue α/β protein domains of proteins G and L. Specifically, the domains under observation are the B1 segments of the IgG binding domain of each protein. The diffusion-collision model approximates regular elements of secondary structure (α helices and β-hairpins) as hard spheres connected by flexible featureless strings (loops). These secondary structure elements are in fast equilibrium between the native and denatured state and the folding process involves diffusion of these elements, collision, and occasional coalescence (Karplus and Weaver, 1976). Folding rates, and the probability of isolating any particular folded species over time, are calculated using a diffusion equation. Remarkably, the validity of the diffusion-collision model was recently confirmed for α helical proteins by combined experimental (ϕ-value) analysis and an explicit water molecular dynamics simulation study for a 61-residue three-helix bundle (Engrailed homeodomain from *Drosophila melanogaster*) (Mayor et al., 2003).

The two proteins chosen by Islam et al. (2004) not only extend the range of applicability of the diffusion-collision model to proteins with significant β sheet content but also allow also one to isolate effects due to kinetics and the primary or tertiary structure. This is possible because, despite a sequence identity of only 15%, protein G and protein L have the same native structure. Their folded state is symmetric and consists of a central α-helix packed on a four-stranded mixed β-sheet formed by N- and C-terminal β-hairpins. The

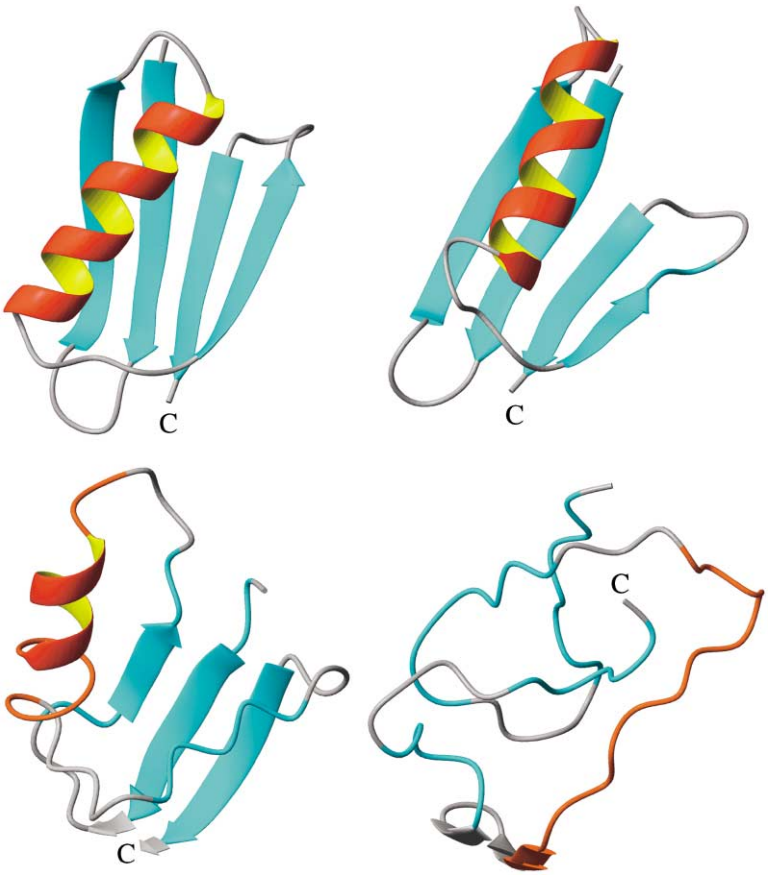

Figure 1. Representative Snapshots (from Left to Right and Top to Bottom) of an Implicit Solvent Molecular Dynamics Simulation of Protein G Unfolding at 385 K (U. Haberthuer and A.C., unpublished results)

In this simulation, unfolding starts at the C-terminal hairpin, but in other runs it begins at the N-terminal hairpin (not shown). In all frames, segments of the backbone that correspond to α-helix and β-sheet in the native fold are colored in red and blue, respectively.

ϕ-value analysis of proteins L and G indicates that for proteins with symmetric native structure more than one folding pathway may exist and the route selected depends on the sequence, i.e., most favorable interactions (McCallister et al., 2000). This is exactly what Islam and colleagues found using the diffusion-collision model, which also shows that both proteins are two-state folders (Islam et al., 2004). The calculated sequence of events for folding is in agreement with both experimental data (McCallister et al., 2000), and previous molecular dynamics simulations that used a minimalist model where each residue was described by one bead located at the α-carbon position (Karanicolas and Brooks, 2002). On the other hand, the observed agreement in the folding rates obtained by diffusion-collision and the experimental means might, in part, originate from the researchers choice of parameters to approximate the stability of individual elements of regular secondary structure.

Although not explicitly discussed by the authors, the application of the diffusion-collision model to protein G and protein L reveals the presence of multiple parallel pathways. For protein G, multiple pathways were previously observed in all atom Monte Carlo simulations using a Go potential (which preferentially stabilizes interactions present in the native structure) (Shimada and Shakhnovich, 2002) as well as in implicit solvent molecular dynamics simulations of unfolding at 385 K (U. Haberthür and A.C., unpublished results, Figure 1). Conversely, a single pathway for protein G was proposed on the basis of experimental data (McCallister et al., 2000). However, given the symmetric native state conformations of protein G and protein L, it is reasonable to expect a statistical distribution of parallel folding pathways. This is supported by the experimentally observed switch in pathways upon weakening of the C-terminal β-hairpin (Nauli et al., 2001). In this context, Jane Clarke and coworkers have recently reported changes in the flux between different transition states on the basis of upward curvature in the guanidinium chloride-dependent unfolding kinetics of a β-sandwich protein (Wright et al., 2003). The final sentence of their paper states: "It remains to be established whether evolution has selected sequences that fold via a single pathway, rather than designing proteins capable of folding via multiple routes, or whether what is unusual is not the existence of parallel pathways, but the fact that they can be experimentally detected and resolved."

Multiple pathways were also detected in implicit solvent molecular dynamics simulations of two designed 20-residue peptides, which have a sequence identity of 15% but the same folded state (a three-stranded antiparallel β-sheet with tight turns at residues 6–7 and 14–15) (Ferrara and Caflisch, 2001). Two folding pathways were observed for each of the two structured peptides in the simulations; they involved the formation of either of the two β-hairpins followed by consolidation of the unstructured strand. For one peptide, about 1/3 and 2/3 of the folding pathways started by formation of the N-terminal and C-terminal hairpin, respectively. For the other peptide, the statistical predominance was the

opposite. These simulation results on structured peptides demonstrated that the possible pathways are defined primarily by the native state structure while the amino acid sequence determines the statistically predominant order of events (Ferrara and Caflisch, 2001).

Finally, it is remarkable that despite the ever-increasing speed of computers, many useful insights in protein folding have been gained from simple models with a coarse-grained description of the protein or the solvent. These approaches include the diffusion-collision model, as well as lattice models which have played a key role in understanding fast folding on a funnel-like energy landscape (Leopold et al., 1992; reviewed in Dill and Chan, 1997), off-lattice coarse-grained models (reviewed in Mirny and Shakhnovich, 2001), and implicit solvent molecular dynamics simulations (Lazaridis and Karplus, 1997; Ferrara and Caflisch, 2001). As is often the case in science, simple, approximate treatments can provide precious hints for solving a complex problem. More detailed, fully atomistic methods are more useful for a direct comparison with experiments (Mayor et al., 2003) but have so far played a smaller role than the simplified models to shed light on the complexity of protein folding.

Acknowledgments

I thank Dr. Urs Haberthür for preparing Figure 1.

Amedeo Caflisch
Department of Biochemistry
University of Zurich
Winterthurerstrasse 190
CH-8057 Zurich
Switzerland

Selected Reading

Dill, K., and Chan, H.S. (1997). Nat. Str. Biol. *4*, 10–19.

Ferrara, F., and Caflisch, A. (2001). J. Mol. Biol. *306*, 837–850.

Islam, S.A., Karplus, M., and Weaver, D.L. (2004). Structure *12*, this issue, 1833–1845.

Karanicolas, J., and Brooks, C.L., III. (2002). Protein Sci. *11*, 2351–2361.

Karplus, M. (2000). J. Chem. Phys. B *104*, 11–27.

Karplus, M., and Weaver, D.L. (1976). Nature *260*, 404–406.

Lazaridis, T., and Karplus, M. (1997). Science *278*, 1928–1931.

Leopold, P.E., Montal, M., and Onuchic, J.N. (1992). Proc. Natl. Acad. Sci. USA *89*, 8721–8725.

Mayor, U., et al. (2003). Nature *421*, 863–867.

Mirny, L., and Shakhnovich, E. (2001). Annu. Rev. Biophys. Biomol. Struct. *30*, 361–396.

McCallister, E.L., Alm, E., and Baker, D. (2000). Nat. Struct. Biol. *7*, 669–673.

Nauli, S., Kuhlman, B., and Baker, D. (2001). Nat. Struct. Biol. *8*, 602–605.

Shimada, J., and Shakhnovich, E. (2002). Proc. Natl. Acad. Sci. USA *99*, 11175–11180.

Wright, C.F., Lindorff-Larsen, K., Randles, L.G., and Clarke, J. (2003). Nat. Struct. Biol. *10*, 658–662.