provides information on motifs and domains within each protein sequence, drawn from many sources. Functional units within proteins have been identified and grouped into families and superfamilies using various criteria, appearing in a variety of classifications such as Pfam (Finn et al., 2006), Worldwide PDB (Berman et al., 2006), CCD (Marchler-Bauer et al., 2006), SCOP (Wilson et al., 2006), and many others (see UniProt). The addition of information on the consequences of interactions between SCOP superfamily domains adds yet deeper and more complex understanding of domain function, which no doubt will be incorporated into the practice of the prediction of protein function from sequence.

In order to improve application of knowledge about domains to genome annotation work, a useful step would be to expand databases to provide explicit information on domain function. In bridging biological fields, there is a problem in communicating information in a form that can be used by the nonexpert genome annotator. Some domains are well known and well described; for example, the various NAD(P)H binding site domains. However, in the SCOP superfamily listings, domains are systematized by alphanumeric coding and are also given brief names: sometimes mnemonics, sometimes understandable abbreviations, and sometimes seemingly opaque labels. PFAM also assigns names, but these are often specific to the function in the protein(s) studied first and might not be literally transferable. Consequently, the nonspecialist may not understand the biological activity of most domains by their labels. Over time and for the sake of the scientific community, descriptions will need to be expanded in an effort to inform the nonspecialist about attributes expected of a domain in an unknown protein, making better use of one biological field's knowledge for another.

Thus, the elegant paper by Bashton and Chothia provides highly specific information about domains and their interactions, particularly for multidomain proteins. In some cases, domains did not change their actions when combined in multidomain proteins, but in other cases, a variety of effects on function resulted from interaction. The detailed information provided in these many examples contributes not only to the field of structural chemistry of proteins, but it also presages the kind of careful and detailed information that will accelerate our understanding of evolutionary mechanisms and will aid the practice of predicting functions of unknown proteins from their sequences.

REFERENCES

Bashton, M., and Chothia, C. (2007). Structure *15*, this issue, 85–99.

Berman, H., Henrick, K., Nakamura, H., and Markley, J.L. (2006). Nucleic Acids Res., in press. Published online November 16, 2006. 10.1093/nar/gkl971.

Finn, R.D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., et al. (2006). Nucleic Acids Res. *34*, D247–D251.

Galperin, M.Y., and Koonin, E.V. (1998). In Silico Biol. *1*, 55–67.

Marchler-Bauer, A., Anderson, J.B., Derbyshire, M.K., Deweese-Scott, C., Gonzales, N.R., Gwadz, M., Hao, L., He, S., Hurwitz, D.I., Jackson, J.D., et al. (2006). Nucleic Acids Res., in press. Published online November 29, 2006. 10.1093/nar/gkl951.

Wilson, D., Madera, M., Vogel, C., Chothia, C., and Gough, J. (2006). Nucleic Acids Res., in press. Published online November 10, 2006. 10.1093/nar/gkl910.

Tying the Knot That Binds

Lisa M. Gloss[1,*]
[1] School of Molecular Biosciences, Washington State University, Pullman, WA 99164, USA
*Correspondence: lmgloss@wsu.edu
DOI 10.1016/j.str.2006.12.001

The recent determination of protein structures with knots in their backbone topology has defied previous conventional wisdom. How proteins can fold with a knot is an intriguing question that has been explored for YibK from *Haemophilus influenzae* in this issue of *Structure* (Mallam and Jackson, 2007a).

It has been over 40 years since Anfinsen and colleagues demonstrated that a protein's sequence contains all the necessary information to determine its structure, stability, and folding mechanism. Deciphering how this information is encoded by the sequence is a holy grail of structural biology. Since the mid-90s, many efforts have focused on studying small, single domain, monomeric proteins. These simple structures often fold by two-state kinetic mechanisms, with no transiently populated intermediates (Jackson, 1998). These experimental systems are also amenable to detailed computational studies, and this synergy has provided new insights and

some atomic level understanding of protein stability and folding (Daggett and Fersht, 2003; Onuchic and Wolynes, 2004).

Fueled by the insights gained from these simple two-state systems, investigators are now turning toward model systems of greater structural complexity, including oligomeric proteins. Unlike monomeric proteins, the folding instructions for oligomers are written in more than one polypeptide. The formation of secondary structures and tertiary interactions must be coordinated with appropriate association of the protein chains, requiring a protein concentration-dependent description of the folding landscape. The challenge is that the folding monomeric species, which may have exposed hydrophobic surfaces, must oligomerize correctly, while avoiding inappropriate intermolecular interactions that could lead to aggregation.

Most studies of oligomeric protein folding have focused on dimers, which exhibit a variety of folding mechanisms (reviewed in Jaenicke and Lilie, 2000). Small dimers, like the P22 Arc repressor, appear to fold by two-state kinetics. As seen for larger monomeric proteins, larger dimers often populate kinetic intermediates, many of which are on-pathway and productive in the formation of the native dimeric species N_2. SecA and Trp repressor from *E. coli*, *Vibrio harveyi* bacterial luciferase and human A1-1 glutathione transferase are examples where partial folding in the monomeric species precedes association. These proteins, as well as *E. coli* factor for inversion stimulation (FIS), bacterial ketosteroid isomerases, and the eukaryotic histone heterodimers, also populate dimeric intermediates before folding to the final N_2 species (Placek and Gloss, 2005 and references therein).

Mallam and Jackson (Mallam and Jackson, 2005, 2006, 2007a) have added an intriguing twist to dimer folding—a trefoil knot within the backbone topology. In a bona fide knot, a significant segment of the protein, twenty residues or more, must be threaded *through* a loop. The prevailing wisdom was that naturally occurring proteins couldn't be knotted, but recently several structures have been reported for proteins with different types of topological knots (Mallam and Jackson, 2005). The dimeric α/β methyltransferases are a growing family of proteins with trefoil knots within each monomer. These proteins utilize S-adenosyl-methionine (SAM), and the knotted region is close to or part of the cofactor binding site in each monomer. YibK from *Haemophilus influenzae* is the first family member for which extensive equilibrium and kinetic urea-induced unfolding and refolding studies have been reported. While it is conceivable that the knot remains in the chemically denatured state, this seems unlikely. Given this caveat, these studies demonstrate that despite the topological complexity of a knot, the protein sequence alone is sufficient to encode the instructions for the efficient folding to the native dimer (Mallam and Jackson, 2005). However, the folding pathway is not a simple one (Mallam and Jackson, 2006).

YibK initially folds to two partially folded monomers through two parallel channels that then converge to form a dimerization-competent monomer, which subsequently folds to the native dimer (Mallam and Jackson, 2006). The parallel folding channels result from unfolded isomers that apparently differ in their proline conformations. YibK has 10 Pro residues with one native *cis* Pro bond. Surprisingly, faster folding occurs through a nonnative intermediate in which all prolines seem be in the *trans* conformation. The most recent study (Mallam and Jackson, 2007a) describes the construction of monomeric versions of YibK that recapitulate the folding features of the dimerization-competent monomeric kinetic intermediate. The secondary structure content of the monomeric YibK mutants is similar to the native dimer, but the tertiary structure, monitored by Trp fluorescence, is only ~40% of the folded dimer. The SAM binding site and the trefoil knot are adjacent to, but not part of, the dimer interface; however, the ability of the monomeric variants to bind SAM is severely impacted. Thus, while the YibK monomer contains significant structure, dimerization is necessary for complete folding, cofactor binding, and by inference, catalysis.

The studies of Mallam and Jackson are a necessary step in the elucidation of how a protein can fold with a knot, but two major questions are still unanswered: (1) when and how is the knot formed, and (2) what is the functional significance of the knot?

Does knot formation occur in a monomeric or dimeric state? What interactions guide the threading of the chain through itself? The similar knotted dimer, YbeA from *E. coli*, folds by a comparable mechanism to YibK, with a partially folded dimerization-competent monomeric intermediate (Mallam and Jackson, 2007b), although YbeA doesn't populate the initial monomeric intermediate(s). The subsequent dimerization of both proteins is slow, 2–4 × 10^2 $M^{-1}s^{-1}$, and rate limited by a first order conformational change. Many association-coupled folding reactions exceed 10^4 $M^{-1}s^{-1}$ and several approach the diffusion limit (Arc repressor mutants, Trp repressor, SecA, FIS, and eukaryotic histones; see Placek and Gloss, 2005 and reference therein). Is knot formation responsible for the slow association kinetics of YibK and YbeA?

While dimerization is clearly essential for complete folding and function (Mallam and Jackson, 2007a), the requirement for a knotted structure is less clear. The classical α/β SAM-dependent methyltransferase fold does not contain knots (Lim et al., 2003). Analysis of the three-dimensional structure of YibK shows that minor shuffling of the sequence connectivity could remove the trefoil knot with no significant repacking of the protein core or secondary structural elements (Lim et al., 2003). Would an unknotted YibK dimer retain activity? Would the folding kinetics, particularly dimerization, be accelerated by decreased topological complexity?

While significant progress has been made on elucidating how to fold and knot a protein, further studies are needed to provide a complete description of this exciting new twist in protein folding.

REFERENCES

Daggett, V., and Fersht, A. (2003). Nat. Rev. Mol. Cell Biol. *4*, 497–502.

Jackson, S.E. (1998). Fold. Des. *3*, R81–R91.

Jaenicke, R., and Lilie, H. (2000). Adv. Protein Chem. *53*, 329–401.

Lim, K., Tempczyk, A., Parsons, J.F., Bonander, N., Toedt, J., Kelman, Z., Howard, A., Eisenstein, E., and Herzberg, O. (2003). Proteins *51*, 56–67.

Mallam, A.L., and Jackson, S.E. (2005). J. Mol. Biol. *346*, 1409–1421.

Mallam, A.L., and Jackson, S.E. (2006). J. Mol. Biol. *359*, 1420–1436.

Mallam, A.L., and Jackson, S.E. (2007a). Structure *15*, this issue, 111–122.

Mallam, A.L., and Jackson, S.E. (2007b). J. Mol. Biol., in press. Published online November 10, 2006. 10.1016/j.jmb.2006.11.014.

Onuchic, J.N., and Wolynes, P.G. (2004). Curr. Opin. Struct. Biol. *14*, 70–75.

Placek, B.J., and Gloss, L.M. (2005). J. Mol. Biol. *345*, 827–836.

Bypassing Translation Initiation

Christopher U.T. Hellen[1,*]
[1] Department of Microbiology and Immunology, State University of New York Downstate Medical Center, Brooklyn, NY 11203, USA
*Correspondence: christopher.hellen@downstate.edu
DOI 10.1016/j.str.2006.12.002

A high-resolution cryo-EM reconstruction of a ribosome-bound dicistrovirus IRES (Schüler et al., 2006) and the crystal structure of its ribosome binding domain (Pfingsten et al., 2006) provide new insights into an exceptional eukaryotic translation mechanism.

Internal ribosomal entry sites (IRESs) are elements present in a subset of eukaryotic mRNAs that mediate translation initiation by noncanonical, end-independent mechanisms known collectively as internal ribosomal entry. Over the last decade, the outlines of three such mechanisms have been elucidated that have different requirements for eukaryotic initiation factors (eIFs). The simplest mechanism is used by the ~200 nt long intergenic region (IGR) IRESs that separate the two large coding regions in the RNA genomes of dicistroviruses such as Cricket paralysis virus (CrPV) and *Plautia stali* intestine virus (PSIV). Initiation on IGR IRESs occurs at GCU/GCA/GCC (alanine) or CAA (glutamine) codons rather than at AUG initiation codons and involves neither initiator tRNA (Met-tRNA$^{Met}_i$) nor eIFs (Sasaki and Nakashima, 2000; Wilson et al., 2000; Pestova and Hellen, 2003; Jan et al., 2003). The recent complementary structural advances reported by Schüler et al. (2006) and Pfingsten et al. (2006) provide new insights into how IGR IRESs promote a process that usually requires at least 11 eIFs.

The canonical initiation mechanism ("scanning initiation") comprises a coordinated series of events that include binding of the eIF2•GTP•Met-tRNA$^{Met}_i$ ternary complex to the 40S ribosomal subunit, attachment of the resulting 43S complex to the 5′ end of an mRNA, scanning to the initiation codon to form a 48S complex, and joining with a 60S ribosomal subunit to form an 80S ribosome in which the Met-tRNA$^{Met}_i$ anticodon is base paired to the AUG codon in the ribosomal peptidyl (P) site.

IGR IRESs enable ribosomes to bypass this process and begin elongation directly. They bind to 40S subunits and to 80S ribosomes independently of eIFs such that the IRES's 3′-terminal CCU triplet occupies the P site (Wilson et al., 2000). This interaction accounts for the competition between IGR IRESs and Met-tRNA$^{Met}_i$ for the P site (Wilson et al., 2000; Pestova et al., 2004). Translation begins following delivery of cognate aminoacyl-tRNA to the ribosomal aminoacyl (A) site by eukaryotic elongation factor (eEF) 1 and its translocation by eEF2 to the P site, which exceptionally occurs without prior peptide bond formation or a deacylated tRNA in the P site (Wilson et al., 2000; Jan et al., 2003; Pestova and Hellen, 2003). In addition to binding to the 40S subunit and mimicking the initiation codon/Met-tRNA$^{Met}_i$ anticodon in the P site, IGR IRESs establish the correct reading frame for translation and might facilitate their own translocation out of the P site. Pfingsten et al. (2006) and Schüler et al. (2006) have established a structural framework for understanding these different steps.

IGR IRESs have closely related structures (Kanamori and Nakashima, 2001): the three domains each contain an essential pseudoknot (Figure 1A). The base-paired CCU triplet that occupies the P site is in domains 3′s pseudoknot (PK I). Domain 3 is connected to PKII (part of domain 1), and domain 1 folds with domain 2, the ribosome binding element (which contains PKIII), to form a stable double-nested pseudoknot. In their 3.1 Å crystal structure (Figure 2), Pfingsten et al. (2006) report that the constituent elements of PSIV IRES, domains 1 and 2 pack together tightly as a result of multiple stabilizing A-minor interactions involving both strands of the large L1.2 loop and the minor groove of helix P2.2. Mutagenesis and footprinting experiments established that the conserved SL-IV and SL-V stem-loops make direct, functionally important interactions with the 40S subunit (Nishiyama et al., 2003). They emerge from the same side of this highly structured core, almost at right angles to

