Opinion TRENDS in Biochemical Sciences Vol.30 No.1 January 2005

Protein folding and the organization of the protein topology universe

Kresten Lindorff-Larsen[1], Peter Røgen[2], Emanuele Paci[3], Michele Vendruscolo[1] and Christopher M. Dobson[1]

[1]University of Cambridge, Department of Chemistry, Lensfield Road, Cambridge, UK, CB2 1EW
[2]Department of Mathematics, Technical University of Denmark, Building 303, DK-2800 Kongens Lyngby, Denmark
[3]University of Zürich, Department of Biochemistry, Winterthurerstrasse 190, 8057 Zürich, Switzerland

The mechanism by which proteins fold to their native states has been the focus of intense research in recent years. The rate-limiting event in the folding reaction is the formation of a conformation in a set known as the transition-state ensemble. The structural features present within such ensembles have now been analysed for a series of proteins using data from a combination of biochemical and biophysical experiments together with computer-simulation methods. These studies show that the topology of the transition state is determined by a set of interactions involving a small number of key residues and, in addition, that the topology of the transition state is closer to that of the native state than to that of any other fold in the protein universe. Here, we review the evidence for these conclusions and suggest a molecular mechanism that rationalizes these findings by presenting a view of protein folds that is based on the topological features of the polypeptide backbone, rather than the conventional view that depends on the arrangement of different types of secondary-structure elements. By linking the folding process to the organization of the protein structure universe, we propose an explanation for the overwhelming importance of topology in the transition states for protein folding.

The widespread application of the methods of structural biology is beginning to provide a comprehensive picture of the variety of possible native folds available to proteins [1–4]. Folding to these structures is, in most cases, the final and crucial step in the transformation of genetic information into a specific biological function. A full understanding of the mechanisms by which folding occurs therefore represents the solution to a central problem in molecular biology [5–7].

Procedures have recently been developed to provide a molecular description of the conformations that are rate-limiting in the folding of a given protein, the transition-state ensemble (TSE; see Glossary), by incorporating the results of a mutational analysis of folding kinetics into computer simulations [8]. Using this approach, ensembles of conformations representing the TSE have been determined for a series of proteins [8–15]. Examination of these ensembles has shown that establishing the correct overall topology of the polypeptide chain is a crucial aspect of protein folding. This observation is in accord with a series of studies that have shown that the folding rate of a protein, to a first approximation, can be related to the entropic cost of forming the native-like topology [16–22].

The structural changes occurring during protein folding have also been analysed in detail for a series of proteins and we discuss some of these studies here. The results enable the topological view of folding to be reconciled with the well-established concept of nucleation [23] by showing that – despite the many different ways in which a given topology could, in principle, be generated – individual proteins use interactions between a specific and limited set of residues to define the fold [8,15,23,24]. Together with methods that aim to describe protein structures in terms of general topological quantities [25,26], these results indicate how the underlying principles that determine the native-state structure are closely related to those that guide the protein-folding reaction.

Glossary

Generalized Gauss integrals: A family of geometric measures constructed to classify protein conformations in terms of general conformational properties [26]. An example of one of these measures is the average number of times a chain segment crosses over and under any other segment when averaged over all directions from which the chain is seen.

Molecular dynamics simulations: A computational method to calculate the time-dependent behaviour of a molecular system. In classical molecular dynamics simulations, a force field associated with the potential energy of a protein is used and Newton's equations of motion are integrated to sample the relevant conformations of all the atoms in a protein molecule. In restrained molecular dynamics simulations, the force field is modified to take experimental data into account, to bias the simulations towards those regions of conformation space that are consistent with the experiment. This procedure enables protein conformations that are in agreement with available experimental data (e.g. Φ-values) to be obtained even if they have free energies that are far from the minimum (e.g. at the TSE) in the unrestrained simulations [10].

Transition-state ensemble: An ensemble of conformations, the formation of which is rate-limiting for folding. If folding from the denatured state is modelled as occurring on a free-energy landscape, the transition state is associated with the largest barrier that needs to be crossed to reach the native state [5,7].

Φ-Value analysis: A kinetic method for obtaining structural information about the transition state for protein folding. Individual amino acid mutations are made throughout the protein sequence, and the effects of the mutations on the folding and unfolding kinetics, in addition to the thermodynamics of folding, are measured. The Φ-value for a specific mutation is the ratio of the stability change in the transition state and in the native state accompanying that mutation and, thus, reports the extent to which the interactions found in the native state can also be found in the transition state [7].

Corresponding authors: Vendruscolo, M. (mv245@cam.ac.uk), Dobson, C.M. (cmd44@cam.ac.uk).
Available online 7 December 2004

 0968-0004/$ - see front matter doi:10.1016/j.tibs.2004.11.008

The nature of folding transition states

The TSEs for three mainly β-sheet-containing proteins, muscle acyl-phosphatase (AcP) [27], the α-spectrin Src homology 3 (SH3) domain [28] and the third fibronectin type III domain of human tenascin (TNfn3) [29], have been studied in particular detail, including comprehensive mutational Φ-value analysis [7] of the interactions that are present within the TSEs. Structural models of the TSEs of these proteins have been determined by incorporating the experimentally determined Φ-values into molecular dynamics simulations. The results of such a procedure is that the simulations are able to identify the conformations that are fully compatible with the experimental data [10,11,15] and, hence, to determine the TSE (Figure 1). Analysis of structures determined in this way has revealed that many features of the native states, including well-defined secondary-structure elements and the burial from solvent of the amino acid residues in the hydrophobic core, are only partially formed in the conformations that make up the TSEs. Nevertheless, despite the fact that these conformations have average root mean square deviations from the native state typically of ~7 Å [10,11,15], and that some conformations have values even larger than 10 Å, inspection of the structures within the TSEs suggests that they have overall topologies that are similar to those of their corresponding native states.

To examine this topological similarity more quantitatively, we have explored ways of analysing the relationships between conformations in a TSE and the known structures of native proteins [15]. The method that has proved most valuable is based on structural alignments between the TSE conformations and a large database of native protein structures. Representative members of the conformations in the TSEs of AcP, TNfn3 and the α-spectrin SH3 domain were aligned with the native-state structures of 921 protein domains of 40–110 residues, extracted from the major structural classes (α, β, α/β, α+β) of the SCOP (structural classification of proteins) database [1,3] (http://scop.mrc-lmb.cam.ac.uk/scop). The structural comparisons were carried out using a distance matrix alignment procedure implemented in the DALI program [2,30,31], in which comparisons of matrices of pairwise C_α distances are used to quantify structural and topological similarities. An alignment from DALI can be judged by its Z-score, a length-normalized measure of the structural similarity; the higher the Z-score the higher the level of structural similarity [31].

The results of this analysis have revealed that the majority of conformations in the TSEs of these proteins can be characterized as being structurally more similar to their native folds than to any other fold within the SCOP database [15] (Table 1). Thus, for AcP, 27 out of 29 (93%) representative TSE structures, selected from a much

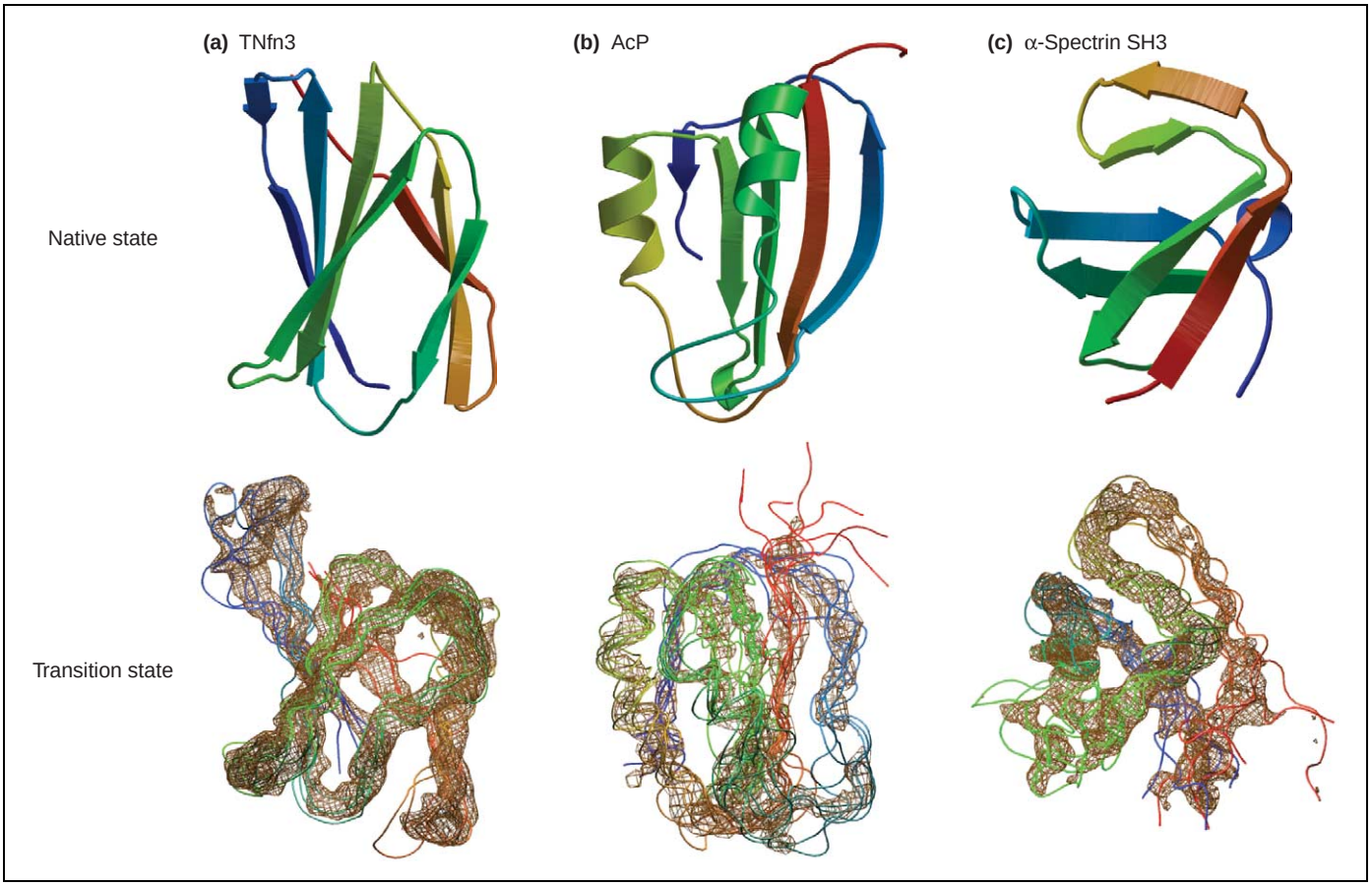

Figure 1. Native structures and transition-state ensembles (TSEs) for **(a)** the third fibronectin type III domain of human tenascin (TNfn3), **(b)** acyl-phosphatase (AcP) and **(c)** α-spectrin Src homology 3 (SH3) domain. Native-state structures were produced using Bobscript [43]. The transition-state structures show a subset of the more native-like conformations within the TSEs traced within atomic-density maps [15,44]. The envelope shown here corresponds to the 30% amplitude isosurface of the density of atoms in the polypeptide main chain. The structures are coloured from their N terminus (red) to their C terminus (blue).

Table 1. Closest matching SCOP domains in alignment with TSE[a]

Protein	Number of TSE conformations	Number of hits in SCOP family[b]	Number of hits in SCOP fold[b]
AcP	29	16 (d.58.10.1)	27 (d.58)
TNfn3	90	63 (b.1.2.1)	78 (b.1)
α-Spectrin SH3	31	28 (b.34.2.1)	29 (b.34)

[a]Abbreviations: AcP, acyl-phosphatase; SCOP, structural classifications of proteins database; SH3, Src homology 3; TNfn3, third fibronectin type III domain of human tenascin; TSE, transition-state ensemble.
[b]Entries within parentheses correspond to SCOP classifications, for example, the AcP-like family, d.58.10.1, which is a sub-class of the ferredoxin-like fold d.58.

larger ensemble by a clustering procedure [10], have a domain from the SCOP ferredoxin-like fold (SCOP classification d.58 – to which AcP belongs) as their closest matching native structure. (In the SCOP nomenclature, families of structurally similar domains are labelled in a hierarchical manner, for example, 'd.58.10.1' for the AcP family, where 'd' refers to the fact that the protein belongs to the α+β class, 'd.58' is the ferredoxin-like fold and 'd.58.10' is the AcP superfamily.) For the α-spectrin SH3 domain, 29 out of 31 (94%) representative TSE structures have a protein from the SCOP SH3 fold (SCOP classification b.34) as their closest matching native structure. For TNfn3, 78 out of 90 (87%) TSE representatives have a member of the immunoglobulin-like β-sandwich fold (SCOP classification b.1), and a further nine transition-state structures have other β-sandwich domains as their closest native structures. Thus, a total of 97% of the representative TSE structures have their closest match with a domain that has a β-sandwich fold. Nearly all the TSE structures for each of the three proteins can, therefore, be characterized as having the overall fold characteristic of the native state.

Strikingly, however, the absolute similarities between the conformations making up the TSEs and the closest matching SCOP structures, as judged by the DALI Z-scores, have been found to be quite low. For the α-spectrin SH3 domain, the average Z-score between the TSE structures and their closest matching SCOP domain is 1.6. For AcP and TNfn3, the equivalent average Z-scores are 1.7 and 2.9, respectively. These are well below the Z-scores that are characteristic of pairs of native-state structural homologues, which typically have $Z > 5$ [32]. This result indicates that, although the TSE structures have the key features of the protein architecture that define the overall native fold, the degree of structural similarity to the native state of typical members of the TSE is quite low.

The universe of protein folds

A global analysis of the 'universe' of protein folds has previously been used to provide an overall view of the structural similarities and differences between native-state structures [2,4,26,30]. Particularly useful for this purpose are low-dimensional projections that aim to capture the overall features of the high-dimensional space of protein conformations [2,4,26]. Therefore, to visualize the results of the structural comparisons described here, Figure 2a shows a 3D projection of the combined universe of SCOP folds and an average representation of the AcP TSE. In this projection, each fold is represented as a point and these points have been distributed in the plot so that their pair-wise distances correspond as closely as possible to the structural similarities determined by DALI [4,33]. As the dimensionality of conformational space is significantly larger than three [4], any such projection provides only an approximate view of the relationships between different folds, that is, the distances between the points in the plot correspond only approximately to the structural similarities. Nevertheless, the figure clearly shows the expected clustering, based on the SCOP classes of α, β, α/β and α+β folds, indicating that a 3D projection is sufficient to capture the overall features of protein-fold space [4]. Furthermore, the TSE, represented on the plot by its centre-of-mass in the projection, can be seen to be located at the edge of the region of fold-space that corresponds to native proteins. This is in line with the fact that the structural similarities, even to the closest matching SCOP domains, are quite low. However, as we have discussed, the majority of these structures are closer to a domain with the correct native fold than to any other native fold; this conclusion is evident in the figure from the clustering between the TSE structures and the corresponding native SH3 domains present in SCOP. Similar conclusions can be drawn from examination of the analogous projections for the α-spectrin SH3 domain and TNfn3 (Figure 2b,c).

As a means of resolving the apparent paradox that a specific native-like fold can be established despite a low structural similarity, we have recently developed a description of protein conformations that is independent of secondary structure and, instead, is focused on topological properties of the polypeptide chain [26]. By first applying a smoothing procedure [25,34] to the C_α traces of the domains in the SCOP database (Figure 3), a series of space curves is obtained that we define as the underlying chain topologies of native proteins. The features of these smoothed structures have been analysed by calculating a series of generalized gauss integrals [26] that enable the topological similarity between protein structures to be quantified. Figure 3 is a 3D projection of topology space that uses the smoothed-chain representation to illustrate the relationship between native proteins on this basis. Remarkably, a clear preference is observed for domains to cluster according to their SCOP secondary-structure classes. This observation reveals that there is a strong link between secondary-structure composition and chain topology in native proteins by showing that, when the smoothed-chain representations of protein structures are compared using the Gauss integrals, they still cluster according to their secondary-structure classification. This result suggests that there is an almost exact correspondence between a native-state fold (i.e. the non-smoothed structures) and the more general native-state topology defined by the smoothed-chain representations.

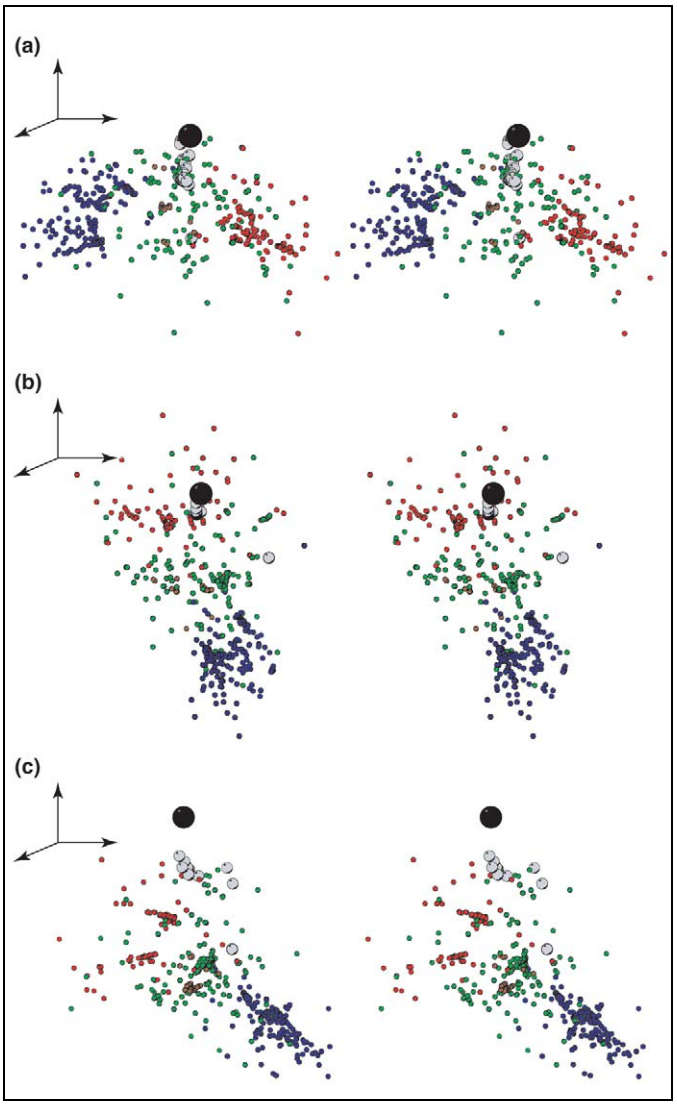

Figure 2. The universe of native-protein folds and transition-state structures. 3D projections (stereo-view) of the fold-space for 921 domains from the SCOP database (http://scop.mrc-lmb.cam.ac.uk/scop). **(a)** Acyl-phosphatase (AcP) transition-state ensemble (TSE) structures, **(b)**) the third fibronectin type III domain of human tenascin (TNfn3) TSE structures, and **(c)** α-spectrin Src homology 3 (SH3) domain TSE structures. The SCOP domains are coloured according to their overall classification: blue, α; red, β; brown, α/β; green, $\alpha+\beta$. Larger grey spheres correspond to SCOP domains from (a) the ferredoxin-like fold, (b) the immunoglobulin-like β-sandwich fold and (c) the SH3-like barrel fold. The black spheres represent the centre of mass in the projection of the representative members of the TSE; all the TSE conformations were used to generate the projections, but their positions in the projections were subsequently averaged so as to represent the TSE only as a single point. Grey outlying spheres represent domains that are evolutionarily related, but are classified by the DALI program as structurally dissimilar to other members of the fold; an example being the structure of a domain swapped dimer of an SH3 domain [45]. The projections were prepared using the Isomap procedure [33], which uses local distance information only – here, over the 20 nearest neighbours – to reconstruct geodesic distances across the multi-dimensional protein-structure universe. In these plots, the high dimensional conformation space of proteins is visualized in a low-dimensional projection by distributing individual points representing each protein domain so that the pair-wise distances in the plot reflect their structural similarities as closely as possible. Z (similarity)-scores were converted into distances (dissimilarity scores) using the empirically chosen function $d(Z) = \gamma/(1+\exp[(Z-Z_0)/\beta])$ with $Z_0=2$, $\beta=1$ and γ chosen to ensure $d(0)=1$. For $i=j$, we set the distance of a protein to itself to zero. The low-dimensional projections were obtained using the pair-wise structural similarities but no other information about the protein structures. A consequence of this procedure is that the axes in the projections carry no specific structural information, but are chosen simply to obtain optimal agreement between the distances in the projections and the structural similarities.

The observation that there is a simple relationship between native folds and the general topological properties of the polypeptide chain is further supported by the results of a structural classification of 1674 SCOP domains with $<95\%$ pair-wise sequence identities and of lengths between 40 and 110 residues. In particular, the question was asked whether or not the nearest structural neighbour of any given domain, defined as having the highest structural similarity based on the Gauss integrals, belongs to the same SCOP fold. Using the non-smoothed (fold) representation of the native-state domains 81% of the 1674 SCOP domains have another domain with the same fold as its closest neighbour; this number drops only to 76% when the classification is based on the smoothed (topology) representation of the native state. This result indicates that the overall chain topology, rather than the explicit secondary structure, does indeed specify the native-state fold. As the smoothing procedure removes local information about secondary structure, this observation is likely to be caused by the inherent differences in overall packing patterns between protein folds of different types of secondary structure [35].

The determinants of protein folding

To rationalize the experimentally observed relationships between protein structure and folding rates, a model for folding has been suggested in which the formation of the overall topology of the native state is a rate-limiting step in protein folding [22]. Such models do not, however, easily explain the experimental observation that certain amino acid residues have a more important role than others in the folding of individual proteins [23]. Recently, it has been shown that the formation of a network of contacts involving a limited set of key residues is sufficient to determine the overall architecture of the TSE for the folding of a given protein [8,10,15,24]. Furthermore, it has been shown for several proteins, including those discussed here, that the overall fold formed in the TSE of a protein is, in fact, predominantly that of the native state [15,36]. These observations enable the descriptions of protein folding that are based on the concepts of nucleation and of topological constraints to be reconciled. In particular, it has been shown that the formation of a specific network of interactions between amino acid residues that can be described as an extended nucleus [20,23], most often found within the hydrophobic core, restricts the region of conformational space available to the protein chain in the TSE in such a way that the topology is defined [15]. Thus, although the formation of the overall topology is the key event in the folding of many small proteins, this topology results from the interactions between sets of specific residues; an example of a mechanism by which this occurs in the SH3 domain of the Src tyrosine kinase is shown in Figure 4 [15].

Such observations by themselves do not, however, explain how the fully native structure can be formed from the loosely assembled structures characteristic of a TSE. In this respect, the close link between a protein fold and its smoothed-chain representation presented here suggests that once the overall chain topology has been established, any local secondary structure not present in

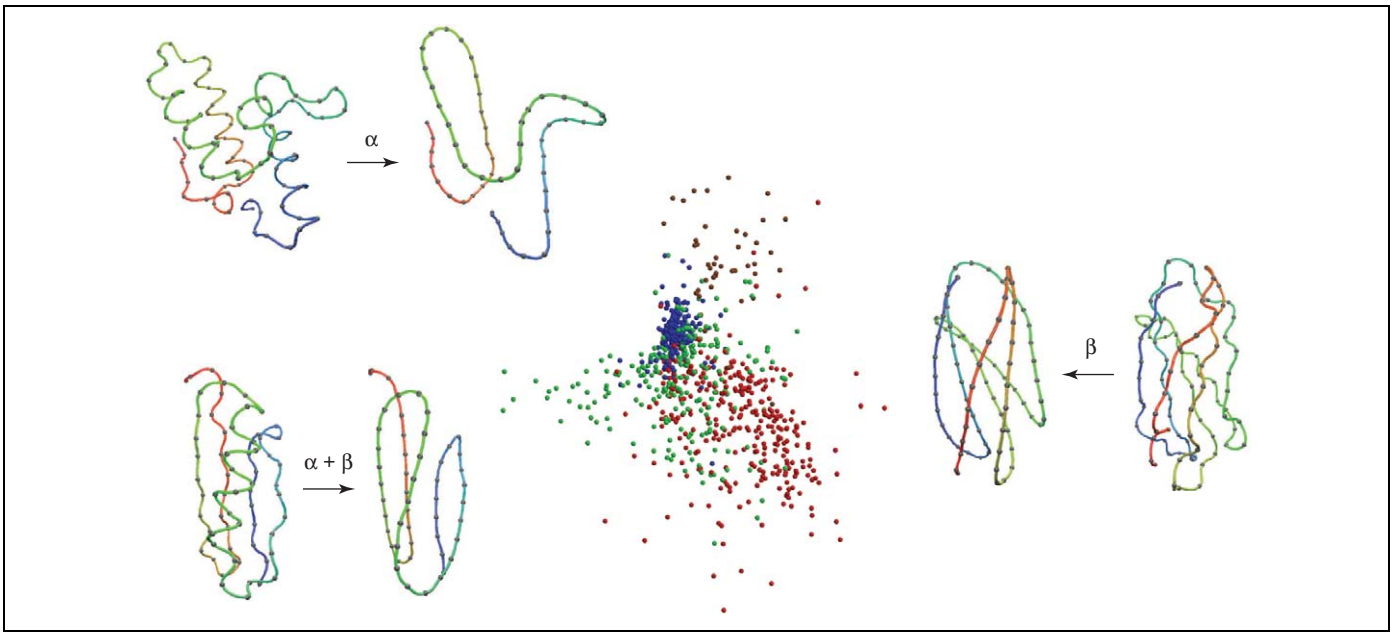

Figure 3. A topological view of the protein structure universe. The structures show the effect of a smoothing procedure on three proteins representing α, β and α+β folds. The three proteins and their smoothed representations are coloured along their sequence from the N terminus (red), through the central region (green), to the C terminus (blue); black spheres correspond to C_α atoms. Protein-chain smoothing was performed by modifying a previously described procedure [34]: C_α atoms were moved to new positions calculated according to $r_i^{new} = (r_{i-2} + ar_{i-1} + br_i + ar_{i+1} + r_{i+2})/(2 + 2a + b)$, where r_i is the position of the ith C_α atom, and $a = 2.4$ and $b = 2.1$. The latter values were chosen to minimize the total chain curvature [46] over different proteins. The update from r_i to r_i^{new} was, however, only performed if this did not involve any chain-crossing events. It is clear from the figure that the smoothing procedure essentially removes the signature of secondary-structural elements such as α helices. However, the local density of atoms on such a smoothed chain is dependent on whether the chain corresponds to, for example, a helix or a strand [25]. To minimize bias caused by this effect, C_α pseudo-atoms were distributed at equidistant (2 Å) positions along the smoothed chain. The central plot in the figure provides a visualization of the overall features of the universe of these smoothed representations of proteins, by showing a 3D projection of fold space consisting of 921 SCOP (http://scop.mrc-lmb.cam.ac.uk/scop) domains after smoothing as described. The smoothed representations were grouped using generalized Gauss integrals [26,47] to determine their topological similarities. The domains are coloured according to their secondary-structure classification in the SCOP database: blue, α; red, β; brown, α/β;green, α+β.

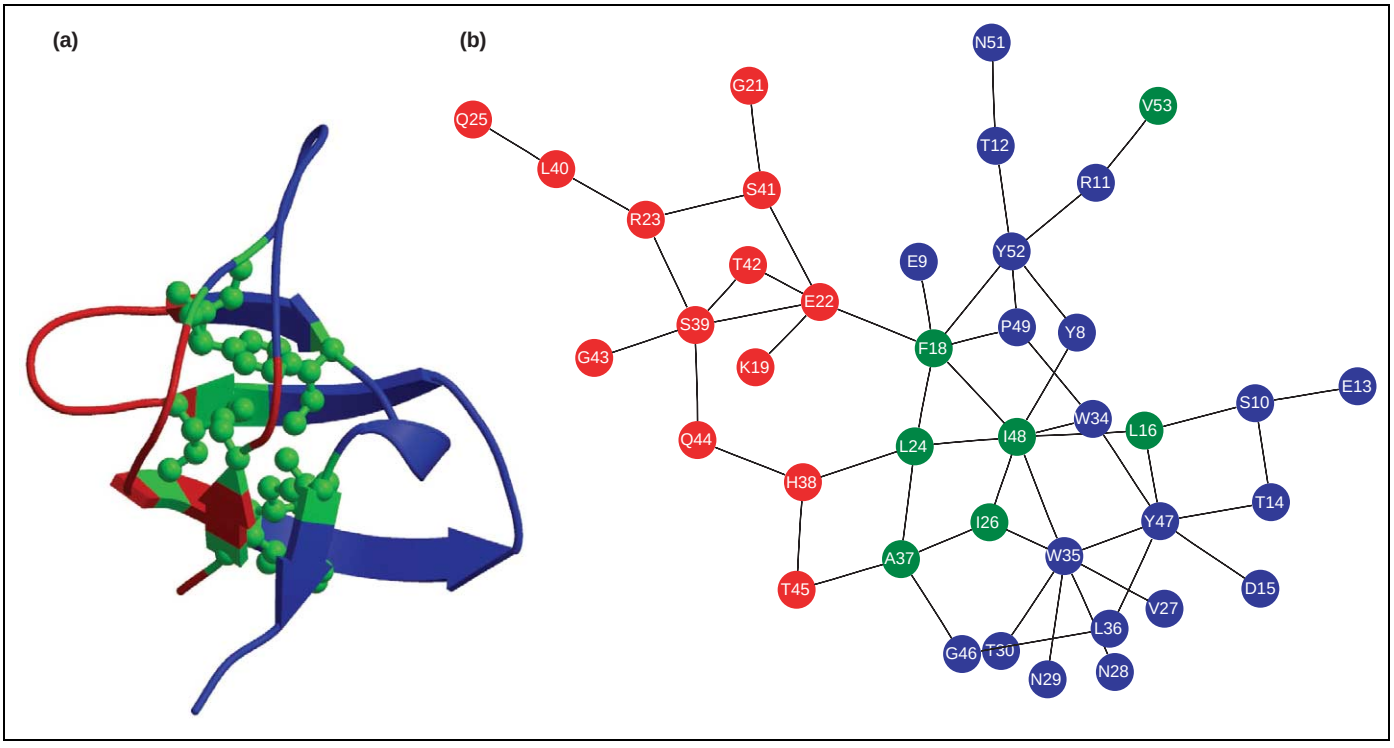

Figure 4. Definition of the topology in the transition state ensemble (TSE) of the Src tyrosine kinase Src homology 3 (SH3) domain by a network of interactions. **(a)** Native-state structure of the SH3 domain in which the residues in the hydrophobic core are shown in a ball-and-stick representation and are coloured green, and the left- and right-hand region of the protein structure in this view is coloured red and blue, respectively. **(b)** Network of interactions in the TSE of SH3. Each node corresponds to a single amino acid residue, each of which is colour-coded using the scheme in (a). The connecting lines indicate strong non-covalent interactions in the TSE between amino acids more than two residues apart [15]. A subset of residues in the hydrophobic core is required to define the overall network of interactions in the TSE. Furthermore, it has been shown that the formation of a network between the same six residues in the core is essential for defining the topology in the TSEs of three different SH3 domains [15].

the TSE can readily form during the last stages of the folding reaction in which the side chains lock together to generate the closely packed native state. This observation can explain the importance of obtaining a native topology in the TSE and provides a mechanism by which the observed robustness of transition-state structures, even towards large changes in sequence [37,38], can be generated.

Concluding remarks: encoding the native fold

The studies reviewed here provide a framework for understanding the mechanism of folding of small globular proteins and, in particular, how the interactions involving just a few residues can play a crucial part in encoding the native fold despite the fact that the TSE lacks many native-like properties such as the complete formation of secondary structure, the close packing of side chains and the full burial from solvent of hydrophobic-core residues [10,11,15]. Such a situation is unlikely to occur in any random amino acid sequence and the sequences of naturally occurring proteins seem to have been selected during evolution to have properties such as the ability to fold efficiently to the closely packed structures that are typical of the majority of cellular proteins [6].

With the use of the approaches described here, in combination with the increasing success of protein design procedures, it might become possible to test the extent to which the selection of amino acid sequences through evolution has influenced the folding and misfolding properties of proteins [39–41]. In addition, it seems that for larger proteins the topology can develop locally within domains of the structure that subsequently dock together, with a few key residues again enabling the correct overall architecture to be established [36,42]. Thus, the principles of folding determined from detailed studies of simple systems can be extended to describe, at least in outline, the manner in which even the most intricate protein structures can be attained.

In conclusion, the combination of the results from protein-folding experiments, computer simulations and the methods of structural bioinformatics is beginning to explain how the sequence of a protein can encode the information needed both to generate the correct native fold and to guide the protein-folding reaction so that it can efficiently adopt its functional native state.

Acknowledgements

K.L-L. is supported by the Danish Research Agency. M.V. is a Royal Society University Research Fellow. The research of M.V. and C.M.D. is supported, in part, by Programme Grants from the Wellcome and Leverhulme Trusts.

References

1 Murzin, A.G. *et al*. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536–540
2 Holm, L. and Sander, C. (1996) Mapping the protein universe. *Science* 273, 595–602
3 Brenner, S.E. *et al*. (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.* 28, 254–256
4 Hou, J. *et al*. (2003) A global representation of the protein fold space. *Proc. Natl. Acad. Sci. U. S. A.* 100, 2386–2390
5 Dinner, A.R. *et al*. (2000) Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem. Sci.* 25, 331–339
6 Dobson, C.M. (2003) Protein folding and misfolding. *Nature* 426, 884–890
7 Daggett, V. and Fersht, A.R. (2003) Is there a unifying mechanism for protein folding? *Trends Biochem. Sci.* 28, 18–25
8 Vendruscolo, M. *et al*. (2001) Three key residues form a critical contact network in a protein folding transition state. *Nature* 409, 641–645
9 Li, L. and Shakhnovich, E.I. (2001) Constructing, verifying, and dissecting the folding transition state of chymotrypsin inhibitor 2 with all-atom simulations. *Proc. Natl. Acad. Sci. U. S. A.* 98, 13014–13018
10 Paci, E. *et al*. (2002) Determination of a transition state at atomic resolution from protein engineering data. *J. Mol. Biol.* 324, 151–163
11 Paci, E. *et al*. (2003) Self-consistent determination of the transition state for protein folding: application to a fibronectin type III domain. *Proc. Natl. Acad. Sci. U. S. A.* 100, 394–399
12 Paci, E. *et al*. (2004) Comparison of the transition states ensembles for folding of Im7 and Im9 determined using all-atom molecular dynamics simulations with Φ value restraints. *Proteins* 54, 513–525
13 Hubner, I.A. *et al*. (2004) Commitment and nucleation in the protein G transition state. *J. Mol. Biol.* 336, 745–761
14 Hubner, I.A. *et al*. (2004) Simulation, experiment, and evolution: understanding nucleation in protein S6 folding. *Proc. Natl. Acad. Sci. U. S. A.* 101, 8354–8359
15 Lindorff-Larsen, K. *et al*. (2004) Transition states for protein folding have native topologies despite high structural variability. *Nat. Struct. Mol. Biol.* 11, 443–449
16 Plaxco, K.W. *et al*. (1998) Contact order, transition state placement and the refolding rate of single domain proteins. *J. Mol. Biol.* 277, 985–994
17 Debe, D.A. *et al*. (1999) The topomer-sampling model of protein folding. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2596–2601
18 Baker, D. (2000) A surprising simplicity to protein folding. *Nature* 405, 39–42
19 Plaxco, K.W. *et al*. (2000) Topology, stability, sequence and length: Defining the determinants of two-state protein folding kinetics. *Biochemistry* 39, 11177–11183
20 Fersht, A.R. (2000) Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism. *Proc. Natl. Acad. Sci. U. S. A.* 97, 1525–1529
21 Makarov, D.E. and Plaxco, K.W. (2003) The topomer search model: a simple, quantitative theory of two-state protein folding kinetics. *Protein Sci.* 12, 17–26
22 Gillespie, B. and Plaxco, K.W. (2004) Using protein folding rates to test protein folding theories. *Annu. Rev. Biochem.* 73, 837–859
23 Fersht, A.R. (1997) Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* 7, 3–9
24 Vendruscolo, M. *et al*. (2002) Small-world view of the amino acids that play a key role in protein folding. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 65, 061910
25 Taylor, W.R. *et al*. (2001) Protein structure: geometry, topology and classification. *Rep. Prog. Phys.* 64, 517–590
26 Røgen, P. and Fain, B. (2003) Automatic classification of protein structure by using Gauss integrals. *Proc. Natl. Acad. Sci. U. S. A.* 100, 119–124
27 Chiti, F. *et al*. (1999) Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nat. Struct. Biol.* 6, 1005–1009
28 Martinez, J.C. and Serrano, L. (1999) The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nat. Struct. Biol.* 6, 1010–1016
29 Hamill, S.J. *et al*. (2000) The folding of an immunoglobulin-like greek key protein is defined by a common-core nucleus and regions constrained by topology. *J. Mol. Biol.* 297, 165–178
30 Holm, L. and Sander, C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* 233, 123–138
31 Holm, L. and Park, J. (2000) DaliLite workbench for protein structure comparison. *Bioinformatics* 16, 566–567
32 Dietmann, S. *et al*. (2002) Automatic detection of remote homology. *Curr. Opin. Struct. Biol.* 12, 362–367

33 Tenenbaum, J.B. *et al*. (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 2319–2323
34 Taylor, W.R. (2000) A deeply knotted protein structure and how it might fold. *Nature* 406, 916–919
35 Chothia, C. *et al*. (1977) Structure of proteins: packing of α-helices and pleated sheets. *Proc. Natl. Acad. Sci. U. S. A.* 74, 4130–4134
36 Fersht, A.R. and Daggett, V. (2002) Protein folding and unfolding at atomic resolution. *Cell* 108, 573–582
37 Cobos, E.S. *et al*. (2003) A thermodynamic and kinetic analysis of the folding pathway of an SH3 domain entropically stabilised by a redesigned hydrophobic core. *J. Mol. Biol.* 328, 221–233
38 Yi, Q. *et al*. (2003) Structural and kinetic characterization of the simplified SH3 domain FP1. *Protein Sci.* 12, 776–783
39 Ventura, S. *et al*. (2002) Conformational strain in the hydrophobic core and its implications for protein folding and design. *Nat. Struct. Biol.* 9, 485–493
40 Kuhlman, B. *et al*. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science* 302, 1364–1368
41 Scalley, M. and Baker, D. (2004) Characterization of the folding energy landscapes of computer generated proteins suggests high folding free energy barriers and cooperativity may be consequences of natural selection. *J. Mol. Biol.* 338, 573–583
42 Vendruscolo, M. *et al*. (2003) Structures and relative free energies of partially folded states of proteins. *Proc. Natl. Acad. Sci. U. S. A.* 100, 14817–14821
43 Esnouf, R.M. (1997) An extensively modified version of MolScript that includes greatly enhanced coloring capabilities. *J. Mol. Graph. Model.* 15, 132–134
44 Schwieters, C.D. and Clore, G.M. (2002) Reweighted atomic densities to represent ensembles of NMR structures. *J. Biomol. NMR* 23, 221–225
45 Kishan, K.V. *et al*. (1997) The SH3 domain of Eps8 exists as a novel intertwined dimer. *Nat. Struct. Biol.* 4, 739–743
46 Rackovsky, S. and Scheraga, H.A. (1978) Differential geometry and polymer conformation 1. Comparison of protein conformations. *Macromolecules* 11, 1168–1174
47 Røgen, P. and Bohr, H. (2003) A new family of global protein shape descriptors. *Math. Biosci.* 182, 167–181