Available online at www.sciencedirect.com

SciVerse ScienceDirect

Current Opinion in Structural Biology

Take home lessons from studies of related proteins

Adrian A Nickson, Beth G Wensley* and Jane Clarke

The 'Fold Approach' involves a detailed analysis of the folding of several topologically, structurally and/or evolutionarily related proteins. Such studies can reveal determinants of the folding mechanism beyond the gross topology, and can dissect the residues required for folding from those required for stability or function. While this approach has not yet matured to the point where we can predict the native conformation of any polypeptide chain *in silico*, it has been able to highlight, amongst others, the specific residues that are responsible for nucleation, pathway malleability, kinetic intermediates, chain knotting, internal friction and Paracelsus switches. Some of the most interesting discoveries have resulted from the attempt to explain differences between homologues.

Address
Department of Chemistry, University of Cambridge, Lensfield Rd, Cambridge CB2 1EW, UK

Corresponding author: Nickson, Adrian A (aan23@cam.ac.uk)

* Current address: MedImmune, Granta Park, Cambridge CB21 6GH, UK.

Current Opinion in Structural Biology 2013, **23**:66–74

This review comes from a themed issue on **Folding and binding**

Edited by **Jayant Udgaonkar** and **Susan Marqusee**

For a complete overview see the Issue and the Editorial

Available online 20th December 2012

0959-440X/$ – see front matter,

http://dx.doi.org/10.1016/j.sbi.2012.11.009

Introduction

In the fifty years since the protein-folding field was first established, there have been thousands of papers detailing the thermodynamic or kinetic characterization of hundreds of different proteins. One particularly useful approach is 'The Fold Approach' [1], which involves a detailed analysis of the folding of several topologically, structurally and/or evolutionarily related proteins in order to discern patterns and trends in folding (stability, pathways and mechanisms).

In this manuscript, we describe a number of studies that highlight how comparisons within and between related protein families have affected our understanding of protein folding. This article builds on our recent review [2•] incorporating significant results from the last few years. Here, we focus on the folding of isolated domains and do not discuss multidomain proteins, misfolding or aggregation.

The malleability of protein folding pathways

A unifying folding mechanism

In the early days of the 'protein-folding problem', three competing mechanisms were proposed that described how a polypeptide chain might fold to the native state: nucleation [3], hydrophobic-collapse [4] and diffusion-collision (framework) [5]. However, an early Φ-value analysis of the small protein chymotrypsin inhibitor 2 (CI2) demonstrated that none of these mechanisms was appropriate, since secondary and tertiary structure formed concomitantly [6]. Thus the nucleation-condensation mechanism was introduced [7], in which long-range contacts set up the initial topology of the protein (incurring a substantial entropic loss with minimal enthalpic gain), followed by a rapid collapse to the native state (with minimal entropic loss but substantial enthalpic gain). Under these conditions, the transition state is usually an expanded form of the native state [8], which helps to explain the strong correlation between native topological complexity (Contact Order) and folding rates, as noted by Plaxco and Baker in the late 1990s [9].

Although the nucleation-condensation mechanism is observed to be widely applicable, several proteins have been shown to fold in a more hierarchical manner. In particular, the engrailed homeodomain (En-HD) was seen to fold via a classical framework mechanism [10]. To investigate whether this result was owing to the simple architecture of the protein, Fersht and co-workers studied four other members of the homeodomain-like superfamily: c-Myb, hRAP1, Pit1 and hTRF1. They observed a slide in mechanism a slide from hTRF1 (pure nucleation-condensation) to En-HD (pure framework) through c-Myb, hRAP1 and Pit1 (mixed mechanisms), which correlated with the innate secondary structural propensity of each domain [11,12•]. The authors used this result to conclude that nucleation-condensation and diffusion-collision are thus ''different manifestations of a common unifying mechanism'' for protein folding. This variation is not unique, and a continuum of mechanisms has also been seen for different members of the PSBD superfamily, where it is again linked to secondary structural propensity [13].

The foldon concept

Further reconciliations between apparently different folding pathways have also been proposed using the concept of 'foldons'. This term was initially used to describe the C-terminal domain of bacteriophage T4 fibritin [14], but was quickly adopted by Wolynes and co-workers to describe independently folding units of a protein chain [15]. Although originally referring solely to

contiguous regions of polypeptide sequence, Englander [16] and Oliveberg [17,18] redefined the term 'foldon' to describe any kinetically competent submotif within a protein (i.e. any subset of residues that can fold cooperatively to a defined structural state).

Perhaps the most successful application of the foldon hypothesis comes from studies of the ferredoxin-like family of proteins including U1A and the small ribosomal protein S6 from Thermus thermophilus ($S6_T$). Here, Oliveberg and co-workers observed that, while the wild-type $S6_T$ protein folded through a globally diffuse transition state that typified nucleation-condensation, a circular permutant (with conjoined wild-type termini and a different backbone cleavage site) exhibited an extremely polarized transition state [19]. Moreover, two alternate circular permutants demonstrated that entropy mutations could be used to shift the position of the nucleus within the topology of the $S6_T$ protein [20]. This finding was particularly interesting, since it reconciled the folding of $S6_T$ and U1A with that of $S6_A$ and ADA2h: two other homologous ferredoxin-like proteins that appeared to fold through a different pathway (although still by nucleation condensation). Oliveberg explained these results by suggesting that all ferredoxin-like proteins comprise two overlapping foldons, but that the specific folding pathway is determined by the primary sequence of each domain [18].

It is, perhaps, easiest to compare these foldons to tandem repeat proteins. In these proteins, each repeat is unstable in isolation – and yet each repeat has a defined native structure to which it will fold [21,22•]. Interactions between these repeats can provide sufficient stabilization to produce a globally stable native state, and a cooperatively folding protein [23]. In the same way, isolated foldons are unstable – but the combination of several foldons will lead to a stable, structured protein domain. In the ankyrin repeat protein myotrophin, it is the C-terminal repeat that is most stable (least unstable) in isolation, and hence folding begins in this region of the protein. However, when this repeat is destabilized by mutation, it is now the N-terminal repeat that is most stable, and the protein will fold from the opposite end over a different pathway [24], similar to that of Internalin B [25]. A similar rerouting of the folding pathway has also been achieved by mutations in the Notch ankyrin domain [26]. In an analogous manner, the folding of the ferredoxin-like proteins is controlled by which of the two component foldons is the most stable (least unstable), hence the differences in transition state structure between U1A/$S6_T$ and $S6_A$/ADA2h [18].

How do folding pathways respond to sequence changes?

Both experiment [27] and theory [28] suggest that the protein-folding nucleus can be subdivided into two distinct sections (Figure 1). The *obligate nucleus* comprises those few interactions that commit the polypeptide chain to fold to the correct native state topology. Such residues pack early, (with high Φ-values), and incur a substantial entropy cost with little enthalpic gain. They are surrounded by the *critical nucleus*, which is a shell of additional interactions that are necessary to turn the free-energy profile downhill (i.e. additional interactions that are accumulated up to the global transition state). These interactions are more plastic, and each folding event may use a different subset of residues within the critical nucleus to effect a barrier crossing. The foldon idea can be combined with that of the obligate and critical folding nucleus to explain the many types of pathway malleability: this is described in Figure 2, and exemplified by members of the immunoglobulin-like (Ig-like) fold.

When considering the folding of related proteins, perhaps the most thoroughly studied fold is that of the Ig-like domains. These all-β proteins have a complex Greek-key architecture, and are extremely common in eukaryotes with over 40 000 distinct domains identified to date [29]. They were chosen for study because, despite their complex topology, there is low sequence identity within each superfamily – and virtually no sequence identity between different superfamilies. Early studies on fibronectin type III (fnIII) domains (TNfn3 and FNfn10) revealed the presence of four key hydrophobic residues in the B, C, E and F strands that constituted the obligate nucleus: interactions of these residues was necessary, but sufficient, to set up the correct topology of the protein [30–32]. Interestingly, the size of the critical nucleus was very different in these two proteins – it is far more extensive in FNfn10 than in TNfn3 (Figure 2B). Moreover, in FNfn10, a few mutations resulted in a small change in the unfolding *m*-value that could indicate a shift in the critical nucleus (Figure 2C). Most importantly, the obligate nucleus of the evolutionarily unrelated Ig domain titin I27 comprised residues that were structurally equivalent to those in the fnIII domains [33]. Thus, these proteins share an obligate nucleus, which is required to set up the correct topology of these complex Greek-key domains and allow folding to proceed. Indeed, the hydrophobic residues of this obligate nucleus were so well conserved that a search of the Protein Data Bank (PDB) was undertaken to find an Ig-like domain that did not contain this nucleation motif. The resultant domain, CAfn2, was subject to a detailed Φ-value analysis that produced a gratifying result: the folding nucleus had simply 'slipped' down the core to use an adjacent pair of hydrophobic residues [34] – both the obligate and critical nuclei have moved in response to sequence changes (Figure 2D).

A final surprise in this analysis of pathway malleability in Ig-like domains came from a more detailed analysis of