Communications

Alternate Pathways for Folding in the Flavodoxin Fold Family Revealed by a Nucleation-growth Model

Erik D. Nelson* and Nick V. Grishin

Howard Hughes Medical Institute, University of Texas Southwestern Medical Center 6001 Forest Park Blvd., Room ND10.124, Dallas, TX 75235-9050, USA

A recent study of experimental results for flavodoxin-like folds suggests that proteins from this family may exhibit a similar, signature pattern of folding intermediates. We study the folding landscapes of three proteins from the flavodoxin family (CheY, apoflavodoxin, and cutinase) using a simple nucleation and growth model that accurately describes both experimental and simulation results for the transition state structure, and the structure of on-pathway and misfolded intermediates for CheY. Although the landscape features of these proteins agree in basic ways with the results of the study, the simulations exhibit a range of folding behaviours consistent with two alternate folding routes corresponding to nucleation and growth from either side of the central β-strand.

Keywords: fold families; equilibrium intermediates; non-native interactions

*Corresponding author

From a folding perspective, the topology of a protein is interpreted by the shape of its native backbone which loosely determines the pattern of atom-to-atom cross-links between its amino acid residues. Over the past several years, simple theoretical and computational models based essentially on topology and minimal entropy loss1–3 have demonstrated that native topology is a "first order" effect deciding the way a protein folds.4–12 While the data so far still provide a very incomplete picture, it suggests that if we could provide any consistent description of protein folding it would be that evolutionary changes which, roughly speaking, conserve topology13–15 and act as perturbations affecting mainly the depths of intermediates and the heights of free energy barriers on a protein’s folding landscape rather than the basic mechanism16–18 that allows it to fold.

However, among these results have now appeared a growing number of excursions away from axiomatic correspondence between folding and topology that must somehow find a place within this picture.19–24 For example, the small proteins L and G share an almost identical, symmetric topology, but both proteins nucleate one of their two β-sheets preferentially, breaking the symmetry of the native fold.20,21 The small, all-helical proteins Im7 and Im9 share essentially the same topology, but Im7 folds through an on-pathway intermediate in which a distorted arrangement of its helices is stabilised by non-native interactions.22,23 Perhaps, it is not so surprising that the folding mechanisms of these proteins are varied. Their native shapes are not frustrated mechanically so they should have greater freedom to respond to structural and energetic perturbations, and their responses (the modulation of intermediates and pathways by these perturbations) may even be somewhat continuous.

On the other hand, even small perturbations such as amino acid substitutions can sometimes cause discrete interconversions of protein structure within a fold family (for instance, changing β-strands to β-helices24,25). Moreover, the structural family of a protein (its fold type or fold classification) often allows large loop insertions, sometimes within secondary structure units, and the substitution of one secondary structure type for another, all of which can affect the entropy of its folding units, the pattern of native contacts between them, and the capacity of these units to evolve more favourable contacts. Accordingly, this more flexible interpretation of topology (fold type) should permit more substantial variations to occur among protein folding mechanisms.

The landscape features that define the folding pathways of larger proteins (~200 amino acid residues) are more discrete, and should have more capacity to accommodate perturbations. These
features still appear to be guided by native topology, however, given the larger and less predictable variations in structure that can be admitted into the fold families of larger proteins, a manifestly pathway-like protein could conceal, in an evolutionary sense, alternate folding routes due to multiple folding units that are responsive to preferential stabilization by a suitable accumulation of these perturbations. Therefore, as with proteins L and G, a purely structural classification of protein families can permit substantial variations among the folding routes of a given fold type, but for larger proteins this may start to define “discrete spectra” of mechanical differences, or “modes” for folding within a family.

If multiple routes do exist for a particular fold type, when does nature choose from among them, and when does it admit mixtures of the routes? These types of problems are just now beginning to be explored, and they are of interest not simply in terms of the physics of how proteins fold but because they may provide information about low lying conformational sub-states that decide how proteins function. Because of the complexities involved in obtaining this information experimentally, simple, computationally efficient folding models, such as those recently used to describe protein transition state structures, could be very useful to infer folding properties and thus direct the process of these measurements more effectively. Here, we use one of these models for a detailed exploration of CheY and two other large proteins from the flavodoxin fold family.

The model is one of an extremely simple type in which amino acid residues are allowed to exist in just two states, either folded (frozen) or unfolded (a discussion of the model is given in the Appendix). Its energetics are heterogeneous and Go-like, the interaction between any two amino acid residues being proportional to the number of atom-to-atom contacts that would exist between them in the native crystal structure of the protein. Each collective state of the amino acid residues is intended to represent a small micro-ensemble consisting of the conformational states of unfolded segments constrained by the frozen amino acid residues and the cross-links that form between them. The entropy of the micro-ensembles is described using simple estimates from polymer theory in which the unfolded segments are modeled as random flight (gaussian) chains and only the space occupied by frozen parts of the molecule is excluded.

In current applications of this model, the micro-ensembles are limited to very simple objects (for example, a nucleus or nuclei with two or fewer loops) for the sake of simplifying the computations. However, it is known that these approximations begin to break down around >100 amino acid residues, precisely where the fine scale features of folding start to matter less and where, due to its speed, the model could be of most use. In a recent paper, we developed an approach to sample more complex micro-ensemble topologies excluded in previous work in order to investigate larger proteins with multiple folding units. We found that including these topologies often led to qualitative improvements in the calculation of transition state structure, and that the dominantly occurring micro-ensembles turned out to have a simple scaling form (see the Appendix) for which an explicit calculation of excluded volume effects of the type noted above would not be too forbidding. Although we account for these effects in only an order of magnitude sort of way, this approximation seems to be enough to draw the kinds of conclusions we need for this work.

The CheY topology studied here seems particularly well suited to description by this model. The transition state structure of CheY (3chy.pdb) compares relatively well with available protein engineering data (correlation coefficient 0.62 or 0.94 if volume increasing mutations are excluded) and the model detects the misfolded and on-pathway intermediate states thought to reflect topological frustration between interior (β-sheet) and exterior (α-helix) layers of the fold that bridge two weakly interpenetrating domains on either side of the central β strand. The level of agreement is surprising since the misfolded intermediate is thought to result from the dynamical connection between these layers and lead to a non-native distortion of the helices, yet we observe the intermediate in a model without explicit dynamical constraints and native-only interactions (Figures 1 and 2). On-pathway the agreement is surprising as well. In crossing the transition state, CheY nucleates from its N-terminal domain and growth is thought to proceed by strands of the β-sheet which frustrates the accretion of α-helices onto the exterior. Again, this is exactly what we observe in our simulations. In rough agreement with the experimental results of Lopez-Hernandez & Serrano, the nuclear region includes β1–β2–β3 and part of α2 (we refer to regions on either side of the central β strand as domains A and B below). The minima in the free energy profile (Figure 2) register with the formation of β-strands and the helices start to form just before the maxima so that the conflict in stability between interior and exterior regions of the fold is periodically resolved in crossing the barriers. The unusual unfolding and refolding features of helices α1 and α2 in Figure 1(a) and the accentuation of the intermediate barrier after β4 in Figure 2 may signify non-native interactions in the actual folding path as we explain later below.

The flavodoxin study of Bollen & van Mierlo suggests that proteins from the same fold family (CheY, cutinase and anabaena apoflavodoxin in this instance) may exhibit a similar pattern of on and off-pathway intermediates. These proteins have lengths ranging from 128 to 197 amino acid residues and very low sequence identity, and protein engineering results exist only for the smallest member, CheY. Interestingly, both cutinase...
Alternate Pathways for Flavodoxin Folding

(1agy.pdb) and apoflavodoxin (1ftg.pdb) contain a number of flexible loop insertions (in cutinase these include \(\alpha\)-helical fragments) at points where \(\alpha\)-helices would connect to \(\beta\)-strands in the B (C-terminal) domain of CheY. These insertions could relax the interior–exterior frustration effect suggested by these authors and allow for greater stability of the B domain which could lead to variations among flavodoxin fold pathways.

Our results for cutinase and apoflavodoxin do share many of the same features described for CheY. Like CheY, the key folding event is growth of the nucleus up to and across the \(\beta_3\) strand dividing the A and B domains of the fold. Also, each protein exhibits, to varying degrees, the signal of a misfolded intermediate observed by Clemente and co-workers in which helices but not strands or loops (except in the nucleus) are folded, and minima (maxima) in the landscape register with the formation of \(\beta\)-strands (\(\alpha\)-helices) consistent with frustration between the interior and exterior regions of the protein. However, at least for apoflavodoxin, the structural mechanism for folding is quite different. The nucleus of apoflavodoxin is on the opposite side (C-terminal, or B-side) of the \(\beta_3\) strand, including most of the C-terminal helix \(\alpha_6\), strand \(\beta_5\), helix \(\alpha_5\) and connecting loops (see Figures 3 and 4) and growth proceeds toward the N-terminal end of the \(\beta\)-sheet. This result is at first difficult to accept given the simplicity of the model and the fact that part of the protein (the N-terminal strand \(\beta_1\)) is confined in its interior, and we will return to this subject later below.

However, we note here that the number of atom-to-atom contacts per residue in the native states of CheY and apoflavodoxin are also weighted in opposite directions (see Figure 4) and this effect, together with the structural differences in the folds seems to explain the results of the simulations.

The atom-to-atom contact profile for cutinase, similar to its folding landscape, could best be pictured as intermediate to CheY and apoflavodoxin. As in CheY, the formation of strands tends to line up with minima in the free energy profile but now the helices are included more within the minima. Although we do not present folding plots for cutinase, it is useful to summarize the results. First, the CheY helix \(\alpha_4\) is unstructured in cutinase

**Figure 1.** Projection of the folding landscape onto (a) \(\alpha\)-helices and (b) \(\beta\)-strands for CheY. \(P_n(q)\) is the probability that sub-structure \(n\) (helix or strand 1–5) is folded when there are \(q\) frozen amino acid residues. The folding process is stepwise, nucleated by domain A (\(\beta_1-\alpha_1-\beta_2\)) at \(q \approx 38\) and proceeding to accrete each section, \(\alpha_n-\beta_{n+1}\), in order along the interior \(\beta\)-sheet. After the protein is nucleated, the addition of each new helix (strand) leads to a maxima (minima) in the free energy profile (Figure 2). The misfolded “helical” intermediate observed by Clemente and co-workers is detected near \(q \approx 24\). Across this region, the strands and loops in domain B remain unfolded totally, the number of nuclei jumps (the probability of four nuclei reaching about 0.1 at \(q = 24\)) and the distribution of nuclear sizes changes abruptly from bimodal (distributed about 2 and \(q\) amino acid residues) to unimodal (distributed about two amino acid residues, the segment size used in the simulations) to bimodal before reaching the transition state. (c) Ribbon diagram of the CheY crystal structure. Light blue regions indicate amino acid residues with native contacts defined by Nelson & Grishin" and Shea et al."
so its C-terminal helix gets indexed as $\alpha_4$. In the unfolded wing of the cutinase free energy profile, part of its domain B is folded, including helices $\alpha_3$ and $\alpha_4$, and strands $\beta_4$ and $\beta_5$. Although $\alpha_3$ remains frozen into the folded wing of the profile, most of the segments unfold near $q=L/2$ ($L$ is the length of the protein) and are “simultaneously” replaced by domain A, $\alpha_2$, and $\beta_3$ before the reaction proceeds. The folding plots have an all or none character that suggests the exchange of B-like for A-like nuclei is part of the folding pathway \(^{27}\) even though the molecule begins this process from a partially misfolded state.

Aside from structural processes, the results above appear roughly consistent with the experimental data. The sizes of free energy barriers are comparable in scale to the results reported by Bollen & van Mierlo, and although it is difficult to establish the topography of the landscape near the misfolded intermediate, the profiles seem as if they could be classified in a similar way. For example, the CheY kinetics were analysed with both on and off-pathway models by the Serrano group to indicate that they lead to the same results. \(^{40}\) This is consistent with the fact that the main transition state can be reached by a partial exchange of helical structure in domain B for nuclear structure in domain A as is indicated by our own results. However, in apoflavodoxin and cutinase domain B folds first, so the exchange should be qualitatively different, and this may explain why an off-pathway kinetic model \(^{27}\) could describe these two experiments better.

Does this over-simplified model predict the basic signature of the folding landscapes?

The model appears to be operating as intended. (i) The transition state structure of the CheY topology agrees well with experiment. (ii) Complex diagrams (nested, inter-linked loop, etc.\(^{36}\)) are very infrequent in simulations for this fold type. (iii) There are very few contacts between domain A and domain B (after strand $\beta_3$) so the nuclei in these two regions are free to fold in parallel (see the Appendix).

---

Figure 2. Structural events along the free energy profile, $F(q)$, of CheY. The misfolded helical intermediate is centered at $q=24$ ($\omega$), and the nucleus, $\beta_2-\alpha_3-\beta_5$, is formed at $q=38$ ($\tau$). Each major basin in the profile corresponds to the completion of a helix $\alpha_n$ (left side of basin), formation of a strand $\beta_{n+1}$ (middle of basin), and partial formation of the following helix $\alpha_{n+1}$ (right side of basin). The structure of the misfolded intermediate, and the registry of helices with maxima in $F(q)$ indicates topological frustration between the $\beta$-interior and $\alpha$-exterior of the protein as suggested by Bollen & van Mierlo.\(^{27}\) The depth of minima (height of maxima) in this region reflect loop closure events that are sensitive to the entropy approximations used in these types of models. The basic structure of the profile is in agreement with that in Clementi et al.\(^4\) except for the placement of the transition state.

Figure 3. Projection of the folding landscape onto (a) $\alpha$-helices and (b) $\beta$-strands for apoflavodoxin. The helix indices follow the crystal structure data in which $\alpha_2-\alpha_3$ corresponds to the CheY helix $\alpha_2$. The strand indices are the same in all three of the flavodoxin proteins. The nucleus of apoflavodoxin includes part of the C-terminal helix $\alpha_6$, all of $\beta_3$ most of a large loop $\alpha_6$ preceding, or inserted into $\beta_6$, a small part of helix $\alpha_8$ and the loop preceding it (see Figure 4). As the transition state is crossed, the rest of $\alpha_6$ forms, and folding continues to alternate from $\alpha$ to $\beta$ moving from C to N-terminal ends until the protein is folded. Again, there are two minima (basins) in folded wing of the free energy profile, comparable in size to CheY, that register with the formation of $\beta_n-\alpha_{n-1}$ layers. The signature of an intermediate with helical structure is visible near $q=36$. 

---
The patterns of atom-to-atom contacts are consistent with the way each protein folds, and although the entropy cost to freeze unfolded segments of proteins depends on amino acid composition, it seems unlikely that including this dependence could lead to something concerted enough to reverse the effect. Finally, in mechanical unfolding of apoflavodoxin, both domain A (the CheY nuclear region) and the helix-strand combination $\alpha_6-\beta_5$ in domain B (the apo-flavodoxin nucleus) are dynamically confined by their local environments, moving essentially as fixed units while the protein unfolds and remaining so long after the core of the protein is exposed to solvent.

As we noted above, non-native interactions can have a substantial impact on, or even control the folding of certain proteins, and some of our results seem to suggest these effects. Although the model does not include non-native interactions directly, proteins do, and the results may reflect their absence in the model at certain points along the folding profiles. The effects of non-native interactions have never been looked at using this type of model and hence it is difficult to decide when they could be present, or what signature they would leave on the model kinetics. Consequently, we decided to look at the Im7 folding landscape where these effects have been mapped out.

Im7 folds through a single intermediate in which three of its four helices ($\alpha_1$, $\alpha_2$, and $\alpha_4$) are structured but distorted non-natively, maximizing the burial of hydrophobic side-chains that would be exposed had the helices adopted their native positions. In crossing the transition state into the native fold, the helices acquire their native orientations, and the binding site for helix $\alpha_3$ is exposed allowing it to fold and ultimately lock the whole protein into its native structure. Our results for Im7 are shown in Figure 5. Its sister protein, Im9, folds across a smooth free energy barrier but still shows some indications of an intermediate perhaps suggesting the results seen at low pH. Both proteins condense into relatively large, partially unfolded ensembles (Figure 5(b)) due to exposed side-chains in the turn regions of the folds. This situation can be improved a bit by extending the contact radii or by including the dependence of the entropy on amino acid type, however, the results here are still very instructive.

Again, in the intermediate parts of the protein are stabilized by non-native interactions. When the transition state is crossed, these stabilizing contacts are exchanged for native contacts and the energetics of the protein and the model converge. Any qualitative differences that exist between the model and the protein due to the missing non-native interactions should be evident before the transition region where these interactions are lost and the differences between the two pathways are reversed. Regions of the protein that are stabilized by non-native interactions in the intermediate should be less stable in the model and may tend to fold late, while regions that are not stabilized by these interactions would tend to fold early. Because this behaviour is reversed on crossing the transition state, it should be evident (if the effect is strong enough) as some type of “wrinkle” in the time order for folding the sub-structures involved in the intermediate, and this is exactly what we observe.
(ii) the free energy of freezing $\alpha_3$ with $\alpha_4$ unfolded. Apparently, forming the loop by native interactions only is unfavourable in both the protein and the model, but the protein can avoid this situation by escaping into the non-native dimension of the free energy landscape\cite{41} where it finds more favourable contacts. The model initially folds helix $\alpha_3$ first, but when the model and protein pathways begin to converge near $q^*$, the stabilizing energy of the protein transiently compensates the $\alpha_3$ loop.

The effects of non-native interactions therefore emerge here as a consequence of the lower dimensionality of the model.\cite{41} Whatever conformations are dynamically accessible to a protein and are somehow stabilized that are not available to the model protein will be subject to the effects described above. To some degree, it should be possible to infer the existence of non-native intermediates from the time reversal of domain stabilities, however, even for Im7 it is difficult to decode the actual sequence of folding events from $P_n(q)$ alone. Thus, although the exchange region in cutinase resembles the transition in Im7, the results could be explained equally well by some inherent frustration in the protein. On the other hand, the CheY on-pathway intermediate involves sub-structures that surround a small pocket, or cavity in the fold, and it seems possible that helix $\alpha_3$ could initially pack non-natively onto the protein with helix $\alpha_4$ partially unfolded (very similar to $\alpha_4$ and $\alpha_3$ in Im7) to better stabilize this part of the fold.

In summary, although these fluctuations, or time-order wrinkles do not provide an absolute test for non-native interactions, their absence suggests that native topology is the dominant effect that guides the folding process. Consequently, for all of the reasons cited above, we believe our results demonstrate that, at the very least, flavodoxin-like folds permit alternate folding pathways corresponding to nucleation from either side of the central $\beta_3$ strand. If this result turns out to be true, an interesting subject for experiment would be to find out whether these pathways can exist in parallel (so that both ends fold and join in the middle) or whether, as it now appears, there is some structural reason why one end or the other of flavodoxin-like topologies tends to fold preferentially.

Looking back on our results, it is remarkable that this system, which is essentially just an Ising model with non-local interactions, can distinguish among the kinetic attributes of these extremely complex objects. The fact that complicated features such as the off-pathway intermediate in CheY and the dynamical confinement of nuclear regions in apoflavodoxin can be detected by a theory that essentially consists of contact weighted native cross-links and loop closure entropy indicates a very basic connection between the structural attributes of local regions in a protein (their shape and connectivity to the rest of the protein) and how such regions are organized dynamically, which ultimately decides the different ways proteins can fold.

**Supplementary Data**

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2006.02.026

**Appendix**

The justification of using this type of model to study proteins as something even minimally reasonable extends from the work of Munoz & Eaton where it was directly applied to quantitatively describe the folding of a small $\beta$-hairpin molecule and later predict the folding rates of small two-state proteins.\cite{29-31,33,34} The approach we describe here is an extension of a separate approach\cite{33,34} that...
provided a starting point to account for the excluded volume effects described in the text. The free energy of a micro-ensemble $\gamma$ in Galzitskaya & Finkelstein\(^{33}\) is defined as:

$$F(\gamma) = \epsilon \sum_{i<j}^{\gamma} \delta(i,j) - T \left[(L-q)\sigma + \sum_{p=1}^{\gamma} s(p)\right]$$  \hspace{1cm} (A1)$$

where in the first (energetic) part of this expression, $\delta(i,j)$ is the number of heavy-atom contacts (including main-chain atoms) between residues $i$ and $j$ in the native crystal structure and the sum $\sum_{i<j}^{\gamma}$ includes all pairs of amino acids that are frozen in $\gamma$. In the entropic part of the expression, $L$ is the chain length, $q$ is the number of folded residues, $\sigma = 2.3R$ is the entropy cost to freeze an amino acid, and $s(p)$ is the entropy cost to link the ends of an unfolded segment into a loop $p$. Unfolded ends and open segments are described as free chain segments while $s(p)$ is approximated by a gaussian chain with ends attached to an impenetrable surface (in the work done by Galzitskaya & Finkelstein,\(^{33}\) the number of loops and open segments in a micro-state is limited to two). The energy scale $\epsilon$ is set by performing the experiments at equilibrium between native ($q=1$) and unfolded ($q=0$) states.

The approach we developed in Nelson & Grishin\(^{36}\) extended this model in effect to permit all orders, or complexities of the micro-ensemble topologies. The dominant contributions in proteins were found to originate from a simple class of scaling topologies (specifically, one or more folded nuclei, each potentially decorated by loops and ends, joined together by open unfolded segments) so that the unfolded part of the system could still be described as a non-interacting soup of open segments and loops just as in the original model. Although this approach neglects the shape (size) of the nuclei and excluded volume effects from nuclei attached to open segments and ends, it is (i) correct in order of magnitude, (ii) leads to qualitative level improvements in the calculation of transition state structures for most (but not all) of the small to moderate size (X-ray) proteins in Galzitskaya & Finkelstein\(^{33}\) and Garbuzynskiy et al.\(^{34}\) (about 30% on average)\(^{36}\) and (iii) identifies intermediates (for example, the CheY misfolded intermediate) that are unavailable to the lower complexity models.

It should be pointed out, however, that even this approach inhibits some kinetically important processes. An inherent constraint of the two-state model is that every pair of amino acids that are in contact in the crystal structure become cross-linked when they freeze into their folded states. Consequently, sub-domains that interpenetrate, or are otherwise strongly connected in the crystal structure are inhibited from folding independently (i.e. in parallel)\(^{41}\). This situation can only be addressed by including an additional state per amino acid which substantially complicates the problem, but it occurs in perhaps one out of 20 or so proteins studied so far (staphylococcal nuclease) and plays at most a very limited part in the systems investigated here.

### References

Alternate Pathways for Flavodoxin Folding


Edited by M. Levitt

(Received 12 October 2005; received in revised form 10 February 2006; accepted 10 February 2006)