Predicting Protein Folding Pathways

Mohammed J. Zaki, Vinay Nadimpally, Deb Bardhan, Chris Bystroff

Springer-Verlag

Berlin Heidelberg New York
London Paris Tokyo
Hong Kong Barcelona
Budapest

Chapter 1. Predicting Protein Folding Pathways

Summary.

A structured folding pathway, which is a time ordered sequence of folding events, plays an important role in the protein folding process and hence, in the conformational search. Pathway prediction, thus gives more insight into the folding process and is a valuable guiding tool to search the conformation space. In this paper, we propose a novel "unfolding" approach to predict the folding pathway. We apply graph based methods on a weighted secondary structure graph of a protein to predict the sequence of unfolding events. When viewed in reverse this yields the folding pathway. We demonstrate the success of our approach on several proteins whose pathway is partially known.

1.1 Introduction

Proteins fold spontaneously and reproducibly (on a time scale of milliseconds) into complex three-dimensional (3D) globules when placed in an aqueous solution, and, the sequence of amino acids making up a protein appears to completely determine its three dimensional structure [17, 4]. At least two distinct though inter-related tasks can be stated.

1. *Problem:* Given a protein amino acid sequence (i.e., linear structure), determine its three dimensional folded shape (i.e., tertiary structure).
2. *Problem:* Given a protein amino acid sequence and its three dimensional structure, determine the time ordered sequence of folding events, called the folding pathway, that leads from the linear structure to the tertiary structure.

The structure prediction problem is widely acknowledged as an open problem, and a lot of research in the past has focused on it. The pathway prediction problem, on the other hand, has received almost no attention. It is clear that the ability to predict folding pathways can greatly enhance

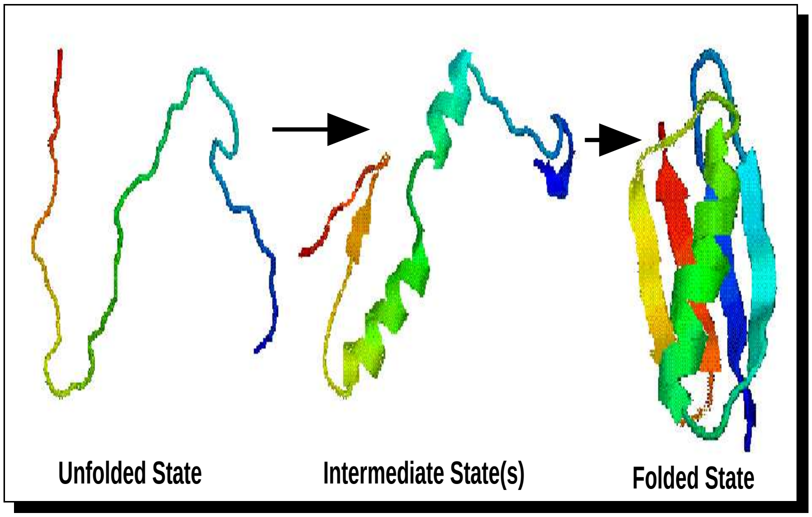

Fig. 1.1. Folding Pathway

structure prediction methods. Folding pathway prediction is also interesting in itself, since protein misfolding has been identified as the cause of several diseases such as Creutzfeldt-Jacob disease, cystic fibrosis, hereditary emphysema and some cancers. In this paper we focus on the pathway prediction problem. Note that while there has been considerable work to understand folding intermediates via molecular dynamics and experimental techniques, to the best of our knowledge ours is one of the first works to *predict* folding pathways.

Traditional approaches to protein structure prediction have focused on detection of evolutionary homology [3], fold recognition [6, 22], and where those fail, ab initio simulations [23] that generally perform a conformational search for the lowest energy state [21]. However, the conformational search space is huge, and, if nature approached the problem using a complete search, a protein would take millions of years to fold, whereas proteins are observed to fold in milliseconds. Thus, a structured folding pathway, i.e., a time ordered sequence of folding events, must play an important role in this conformational search [4]. The nature of these events, whether they are restricted to “native contacts,” i.e., contacts that are retained in the final structure, or whether they might include non-specific interactions, such as a general collapse in size at the very beginning, were left unanswered. Over time, the two main theories for how proteins fold became known as the “molten globule/hydrophobic collapse” (invoking non-specific interactions) and the “framework/nucleation-condensation” model (restricting pathways to native contacts only).

Strong experimental evidence for pathway-based models of protein folding has emerged over the years, for example, experiments revealing the structure of the “unfolded” state in water [18], burst-phase folding intermediates [10], and the kinetic effects of point mutations (“phi-values” [20]). These pathway models indicate that certain events always occur early in the folding process and certain others always occur later (see Figure 1.1).

Currently, there is no strong evidence that specific non-native contacts are required for the folding of any protein [7]. Many simplified models for folding, such as lattice simulations, tacitly assume that non-native contacts are “off pathway” and are not essential to the folding process [14]. Therefore, we choose to encode the assumption of a “native pathway” into our algorithmic approaches. This simplifying assumption allows us to define potential folding pathways based on a known three-dimensional structure. We may further assume that native contacts are formed only once in any given pathway.

Knowledge of pathways for proteins can give important insight into the structure of proteins. To make pathway based approaches to structure prediction a reality, plausible protein folding pathways need to be predicted. One approach to enumerate folding pathways is to start with an unfolded protein and consider the various possibilities for the protein to fold. This approach is infeasible due to the explosively large number of possibilities to consider for the pathways. Our novel approach is to start with a folded protein in its final state and learn how to “unfold” the protein in a time-ordered sequence of steps, to its unfolded state. The reversal of such a sequence could be a plausible protein folding pathway. Our contributions stem from this basic approach. In this paper, we explore the role of minimum cuts on weighted graphs in determining a plausible sequence of unfolding steps.

1.2 Preliminaries

1.2.1 Protein Contact Maps

The 3D conformation of a protein may be compactly represented in a symmetrical, square, boolean matrix of pairwise, inter-residue contacts called the . The contact map of a protein is a particularly useful representation of protein structure. Two amino acids in a protein that come into contact with each other form a non-covalent interaction (hydrogen-bonds, hydrophobic effect, etc.). More formally, we say that two amino acids (or residues) a_i and a_j in a protein are in *contact* if the 3D distance $\delta(a_i, a_j)$ is at most some threshold value t (a common value is $t = 7\mathring{A}$), where $\delta(a_i, a_j) = |\mathbf{r_i} - \mathbf{r_j}|$, and $\mathbf{r_i}$ and $\mathbf{r_j}$ are the coordinates of the α-Carbon atoms of amino acids a_i and a_j (an alternative convention uses β-carbons for all but the glycines). We define *sequence separation* as the distance between two amino acids a_i and a_j in the amino acid sequence, given as $|i - j|$. A contact map for a protein with N residues is an $N \times N$ binary matrix C whose element $C(i, j) = 1$ if residues i and j are in contact, and $C(i, j) = 0$ otherwise.

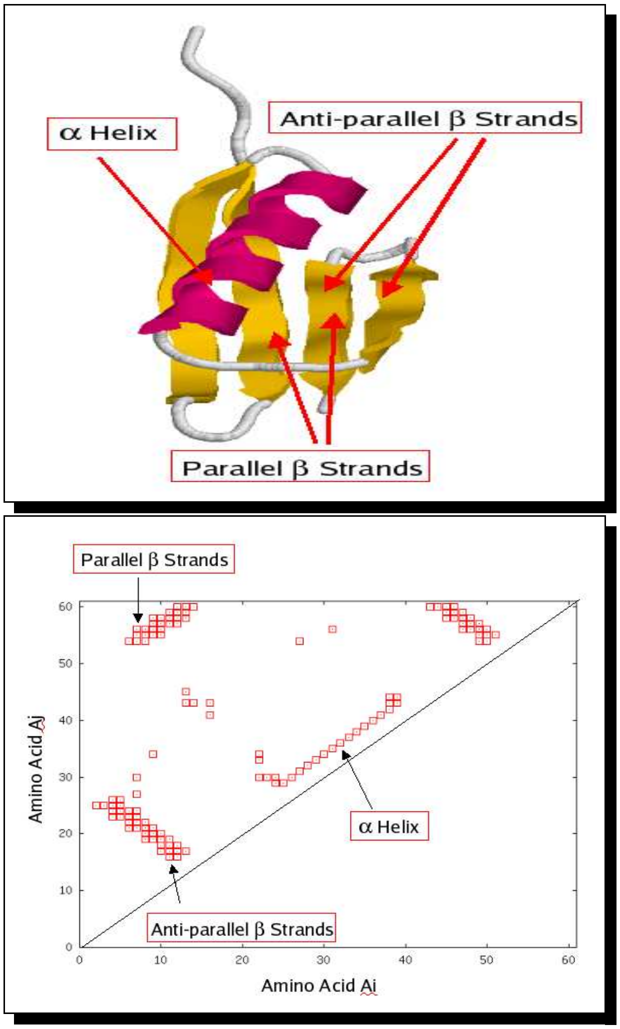

Fig. 1.2. 3D structure for protein G (PDB file 2IGD, Sequence Length 61), and its Contact Map. Clusters of contacts indicate secondary structure elements (SSE); the cluster along the main diagonal is an α-helix, and the clusters parallel and anti-parallel to the diagonal are parallel and anti-parallel β-sheets, respectively.

Figure 1.2 shows the contact map for IgG-binding protein from the (PDB), with PDB code 2IGD (61 residues). A contact map provides useful information about the protein's (SSEs; namely, α-helices and β-strands), and it also captures non-local interactions giving clues to its tertiary structure. For example, clusters of contacts represent certain secondary structures: α-Helices appear as bands along the main diagonal since they involve contacts between one amino acid and its four successors; β-Sheets are thick bands parallel or anti-parallel to the main diagonal. Moreover, a contact map is rotation and translation invariant, an important property for data mining. It is also possible to recover the 3D structure from contact maps [26].

1.2.2 Graphs and Minimum Cuts

An undirected graph $G(V,E)$ is a structure that consists of a set of vertices $V = \{v_1, v_2, \cdots, v_n\}$ and a set of edges $E = \{e_i = (s,t) | s,t \in V\}$, i.e., each edge e_i is an unordered pair of vertices. A is a graph with an associated weight function $W : E \rightarrow \Re^+$ for the edge set. For each edge $e_i \in E$, $W(e_i)$ is called the *weight* of the edge e_i.

A *path* between two vertices $s,t \in V$ is an ordered set of vertices $\{v_1, v_2, ..., v_k\}$ such that $v_1 = s$, $v_k = t$ and for every $1 \leq j < k$, $(v_j, v_{j+1}) \in E$. Two vertices $s,t \in V$ are said to be *connected* in G if there exists a path between s and t. A *connected component* K is a maximal set of vertices $K \subseteq V$, such that for every $s,t \in K$, s and t are connected in G. A graph is said to be a connected graph if $\forall s,t \in V$, s and t are connected.

Let $G = (V,E)$ be a simple undirected, connected, weighted graph. An *(edge) cut* C, is a set of edges $C \subseteq E$, which when removed from the graph, partitions the graph into two connected components V_1 and V_2 (with $V_1 \bigcap V_2 = \emptyset$, $V_1 \bigcup V_2 = V$, $V_1 \neq \emptyset$, $V_2 \neq \emptyset$). An edge *crosses* the if its endpoints are in different partitions of the cut. The *capacity* of the edge cut C is the sum of the weights of edges crossing the cut, given as $W(C) = \sum_{e \in C} W(e)$.

A cut C is a *s-t cut* if vertices s and t are in different partitions of the cut. A *minimum s-t cut* is a $s-t$ cut of minimum capacity. A *(global) (mincut)* is a minimum $s-t$ cut over all pairs of vertices s and t. Note that mincut need not be unique.

1.2.3 Weighted SSE Graph

A protein can be represented as a *(WSG)*, where the vertices are the SSEs comprising the protein and the edges denote proximity relationship between the secondary structures. Furthermore, the edges are weighted by the strength of the interaction between two SSEs. Following the convention used in protein topology or TOPS diagrams [24, 29], we use triangles to represent β-strands, and circles to represent α-helices.

To correctly model the secondary structure elements and their interaction, the edge construction and their weights are determined from the protein's contact map. The edge weights are determined as follows: we determine the list of SSEs and their sequence positions from the known 3D structure taken from the Protein Data Bank (PDB) [1]. Every SSE is a vertex in the WSG. Let $V = \{v_1, v_2, \cdots, v_n\}$ denote a protein with n SSEs. Each SSE v_i has starting

[1] http://www.rcsb.org/pdb/

($v_i.s$) and ending ($v_i.e$) sequence positions, where $1 \leq v_i.s < v_i.e \leq N$, and N is the length of the protein.

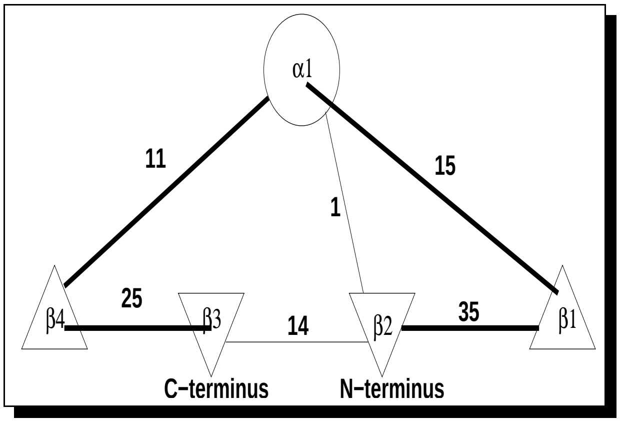

Fig. 1.3. WSG for Protein 2IGD

Let v_i and v_j be a pair of SSEs. Let the indicator variable $b(v_i, v_j) = 1$ if v_i and v_j are consecutive on the protein backbone chain, else $b(v_i, v_j) = 0$. The number of contacts between the two SSEs in the contact map is given as $\kappa(v_i, v_j) = \sum_{i=v_i.s}^{v_i.e} \sum_{j=v_j.s}^{v_j.e} C(i,j)$. An edge exists between two SSEs if there are a positive number of contacts between them, i.e., $\kappa > 0$, or if the two SSEs are on linked on the backbone chain. The weight assigned to the edge (v_i, v_j) is given as follows: $W(v_i, v_j) = \Delta \times b(v_i, v_j) + \kappa(v_i, v_j)$, where Δ is some constant. In out study we set Δ as the average number of (non-zero) contacts between SSEs, i.e., $\Delta = \frac{\mathcal{S}}{|\mathcal{S}|}$, where $\mathcal{S} = \{\kappa(v_i, v_j) > 0 \mid v_i, v_j \in V\}$. This weighting scheme gives higher weights to backbone edges and also to SSEs with greater bonding between them. The backbone edges are given higher weight since they represent strong covalent bonds, while the other contacts represent weaker non-covalent bonds. An example WSG for protein 2IGD is shown in Figure 1.3. The thick line denote backbone edges. SSEs are arranged from the N-terminus (start) to the C-terminus (end), and numbered as given in the PDB file. 2IGD has 5 SSEs, $\beta_2\beta_1\alpha_1\beta_4\beta_3$ arranged from the N- to C-terminus.

1.3 Predicting Folding Pathways

In this section we outline our approach to predict the folding pathway of a protein using the idea of "unfolding". We use a graph representation of a protein, where a vertex denotes a secondary structure and an edge denotes the interactions between two SSEs. The edges are weighted by the strength of the SSE interactions obtain from the *protein contact map.* The basic intuition behind our approach is to break as few contacts as possible, and to avoid splitting a SSE held at both ends. Among several choices, the best option is to pick to one that has the least impact on the remaining part of the protein. Through an series of *minimum cuts* on the weighted graph, we predict the most likely sequence of *unfolding* events. Reversing the unfolding steps yields plausible pathways for protein folding. A detailed description of our approach appears below.

1.3.1 via Mincuts

The basic intuition behind the unfolding process stems from the belief that unfolding occurs by *breaking as few contacts as possible.* Given an weighted SSE graph for a protein, a mincut represents the set of edges that partition the WSG into two components that have the smallest number of contacts (i.e., the bonds) between them. Hence, minimum capacity edge cuts on WSGs can help us determine the points in the protein where unfolding is likely to occur.

. The problem of determining the mincuts of weighted graphs is a well studied problem in graph theory (see [1] for a comprehensive review). We chose the Stoer-Wagner (SW) [25] deterministic polynomial-time mincut algorithm, since it is very simple, and yet is one of the fastest current methods, running in time $O(|V||E| + |V|^2 \log |V|)$. It relies on the following observation: either the global mincut is a $s - t$ mincut or it is not. In the former case if we find the $s - t$ mincut, we are done. In the latter case, it is sufficient to consider a mincut of $G - \{s, t\}$.

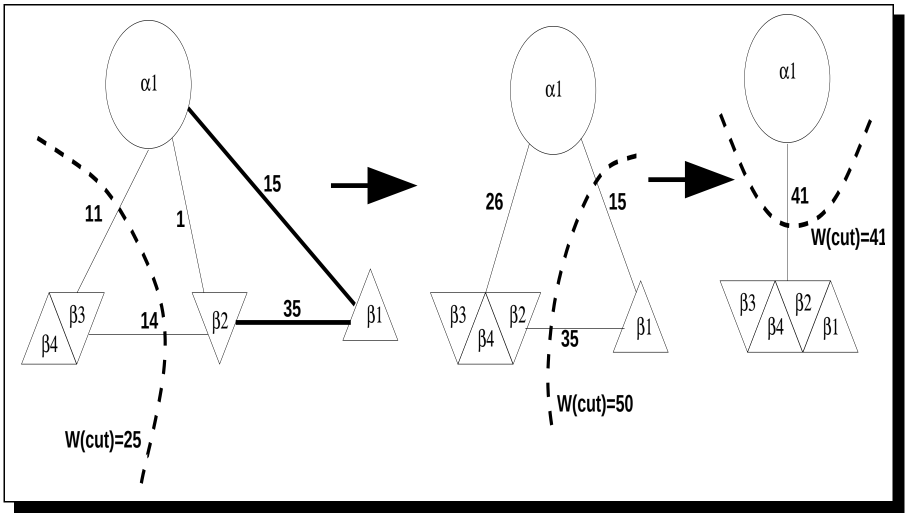

Fig. 1.4. SW Algorithm for Mincut of Protein 2IGD

The SW algorithm works iteratively by merging vertices, until only one merged vertex remains. In each phase i, SW starts with an arbitrary vertex $Y = \{a\}$, and adds the most highly connected vertex $z \notin Y$ to the current set Y, given as $z = argmax_z\{\sum_{x \in Y} W(z, x)\}$. This process is repeated until $Y = V$. At this stage the *cut-of-the-phase*, denoted C_i is calculated as the cut that separates the vertex added last to Y (i.e., the vertex t) from the rest of the current graph. At the end of each phase, the two vertices added last to Y, say s and t, are merged into a single node st, i.e., edges connecting them are removed, and for any $x \in V, W(x, st) = W(x, s) + W(x, t)$. The global mincut is the minimum cut over all phases, given as $C = argmax_i\{W(Ci)\}$.

As an example, consider the WSG for 2IGD shown in Figure 1.3. Let's assume that the starting vertex is $a = \alpha_1$, i.e., $Y = \{\alpha_1\}$. The next SSE to be picked is β_1 since it has the highest weight of connection to α_1 (thus, $Y = \{\alpha_1, \beta_1\}$). Out of the remaining vertices, β_2 has the highest weight of connection to Y ($W(\beta_2, Y) = 36$), so $Y = \{\alpha_1, \beta_1, \beta_2\}$. The last two vertices to be added to Y are $s = \beta_3$ and $t = \beta_4$. At this point phase 1 is over, and the weight of phase 1 cut is $W(C_i) = \sum_{x \in V} W(\beta_4, x) = 36$. We now merge β_3 and β_4 to get a new st node, as shown in Figure 1.4 (left). We next proceed through three more phases (again assuming we start at vertex $a = \alpha_1$), as shown in Figure 1.4. The lowest mincut weight among all the phases is $W(C) = 25$, corresponding to the mincut $C = \{(\beta_2, \beta_3), (\alpha_1, \beta_4)\}$, which partitions the WSG into two components $V_1 = \{\alpha_1, \beta_1, \beta_2\}$, and $V_2 = \{\beta_3, \beta_4\}$.

```
//$G$ is a graph with weight function $W$
UNFOLD ($G = (V, E), W : E \rightarrow \Re^+$):
    $C$ = SW-MinCut($G, W$);
    $G_1 = (V_1, E_1); G_2 = (V_2, E_2)$;
    if ($|V_1| > 1$) UNFOLD($G_1, W$);
    if ($|V_2| > 1$) UNFOLD($G_2, W$);

SW-MinCut($G = (V, E), W : E \rightarrow \Re^+$):
    while ($|V| > 1$)
        $W(C_i)$ = MinCutPhase($G, W$);
    return $C = \text{argmin}_i\{W(C_i)\}$;

MinCutPhase($G = (V, E), W : E \rightarrow \Re^+$):
    $Y = \{\text{some } a \in V\}$;
    while ($|Y| \neq |V| - 2$)
        $Y = Y \cup \{z = \text{argmax}_z\{\sum_{x \in Y} W(z, x)\}\}$;
    Shrink $G$ by merging $s, t \in G - Y$;
    return cut-of-the-phase (from $t$);
```

Fig. 1.5. The UNFOLD Algorithm

. An unfolding event according to our model, is a set of edges that form a mincut in the WSG $G = (V, E)$ for a protein. Our algorithm to predict the unfolding event is called UNFOLD, and it works as follows. First, a mincut C for the initial WSG is determined; ties are broken arbitrarily. This gives the first event in the unfolding process. The edges that form this cut are deleted from the WSG yielding two new connected subgraphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, where V_1 and V_2 are the two partitions resulting from the mincut C, and $E_i = \{(u, v) \in E | u, v \in V_i\}$. We recursively process each subgraph to yield a sequence of mincuts, corresponding to the unfolding events. This sequence when reversed produces our prediction for the folding pathway for the given protein. Figure 1.5 shows the pseudo-code for the complete UNFOLD algorithm to determine the unfolding events for a given protein.

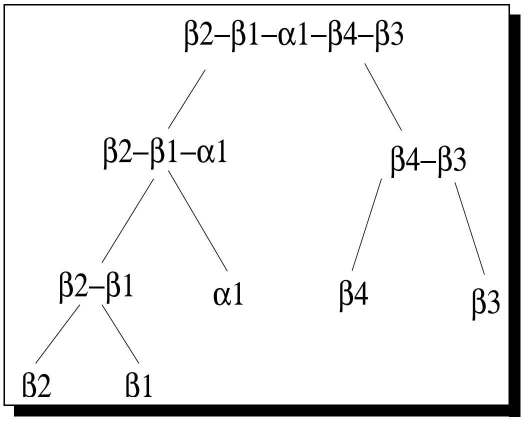

Fig. 1.6. UNFOLD 2IGD

As an example of how UNFOLD works, consider again protein 2IGD; we determined that the first unfolding event (mincut) partitions its WSG into two groups of SSEs $V_1 = \{\beta_2, \beta_1, \alpha_1\}$, and $V_2 = \{\beta_4, \beta_3\}$. After recursive processing UNFOLD produces a sequence of mincuts which can easily be visualized as a tree shown in Figure 1.6. Here each node represents a set of vertices comprising a graph obtained in the recursive application of UNFOLD, and the children of a node are the partitions resulting from the mincut whose value appears in brackets next to the node. For example, the node $\beta_2\beta_1\alpha_1$ is partitioned into $\beta_2\beta_1$ and α_1, which has a mincut value of 25. If we proceed from the leaf nodes of the tree to the root, we obtain the predicted folding pathway of 2IGD. We find that SSEs β_2 and β_1 fold to form a anti-parallel β-sheet. Simultaneously SSEs β_3 and β_4 may also form a parallel β-sheet. SSE α_1 then forms a $\beta_2\alpha_1\beta_1$ arrangement, and then the whole protein comes together by forming a parallel β-sheet between β_2

and β_3. We should be careful not to impose a *strict* linear timeline on the unfolding events predicted by UNFOLD; rather allowance should be made for several folding events to take place simultaneously. However, there may be intermediate stages that must happen before higher order folding can take place. We show that our approach is particularly suited to provide insights into such intermediate folding states.

1.3.2 Detailed Example: Dihydrofolate Reductase (4DFR)

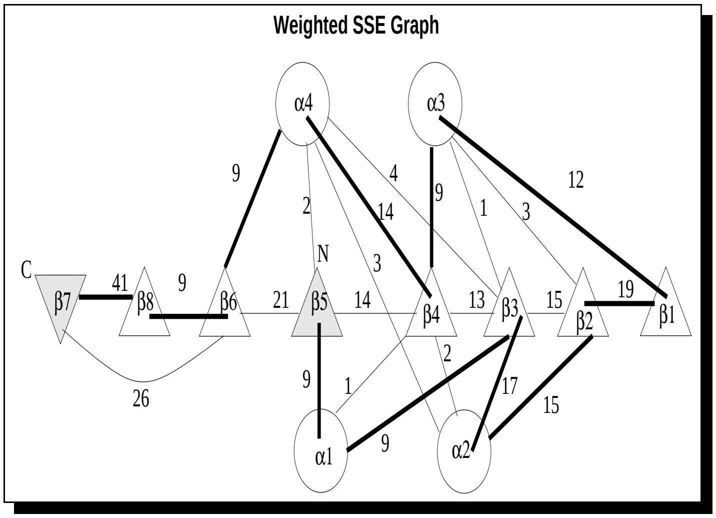

Fig. 1.7. Dihydrofolate Reductase (4DFR): Weight SSE Graph

Although no one has determined the precise order of appearance of secondary structures for any protein, there is evidence that supports intermediate stages in the pathway for several well-studied proteins, including specifically for the protein Dihydrofolate Reductase (PDB 4DFR; 159 residues), a two-domain α/β enzyme that maintains pools of tetrahydrofolate used in nucleotide metabolism [9, 11, 12].

Experimental data indicate that the adenine-binding domain, which encompasses the two tryptophans Trp-47 and Trp-74, is folded, and is an intermediate essential in the folding of 4DFR, and happens early in the folding [11]. Figure 1.7, shows the WSG, unfolding sequence, and a series of intermediate stages in the folding pathway of protein 4DFR. Trp-47 and Trp-74 lie in SSEs α_2 and β_1, respectively. According to our mincut based UNFOLD algorithm, the vertex set $\{\beta_2, \alpha_2, \beta_3, \beta_1\}$ lies on the folding pathway, in agreement with the experimental results!

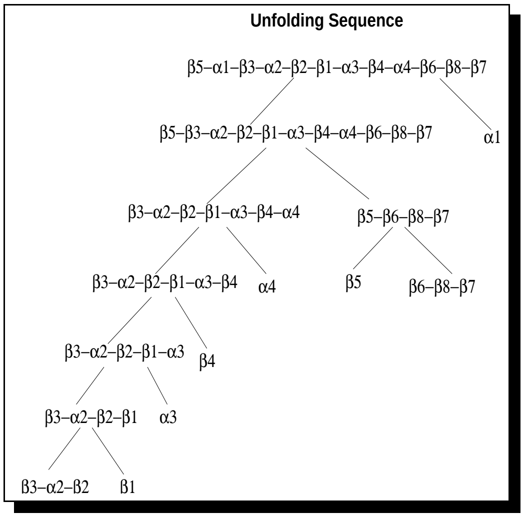

Fig. 1.8. Dihydrofolate Reductase (4DFR): Unfolding Sequence

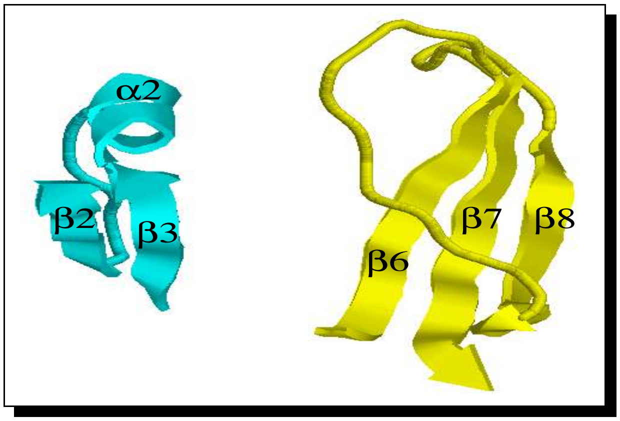

Fig. 1.9. Dihydrofolate Reductase (4DFR): Early Stages in the Folding Pathway

We can see from Figure 1.7, that 4DFR has four α-helices and eight β-strands. The WSG shows the interactions weights among the different SSEs (the bold lines indicate the backbone). Applying UNFOLD to 4DFR yields the sequence of cuts shown. For clarity the unfolding sequence tree has been stopped when there are no more than 3 SSEs in any given node. The remaining illustrations show some selected intermediate stages on the folding pathway by reversing the unfolding sequence.

We find that SSE group $\beta_2\alpha_2\beta_3$ and $\beta_6, \beta_8, \beta_7$ are among the first to fold (Figure 1.9), suggesting that they might be the folding initiation sites. Next β_1 joins $\beta_2\alpha_2\beta_3$, in agreement with the experimental results [9], as shown

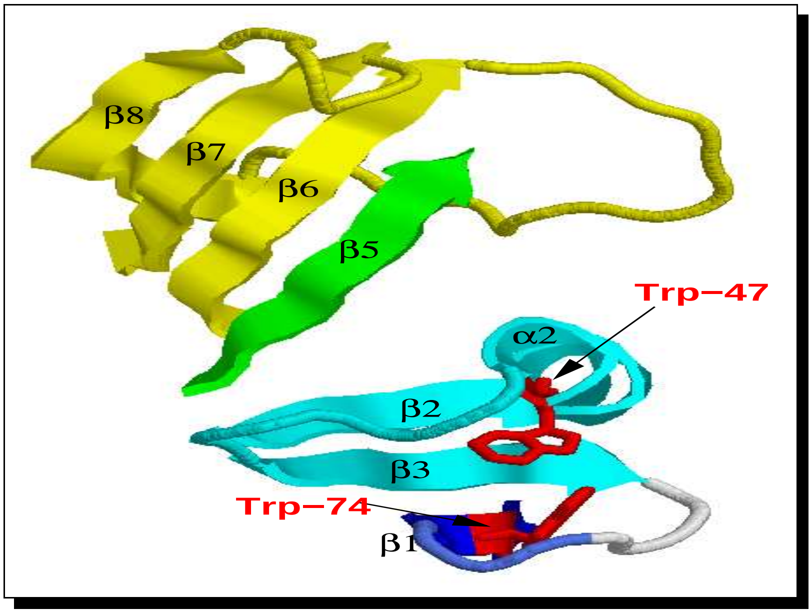

Fig. 1.10. Dihydrofolate Reductase (4DFR): Intermediate Stages in the Folding Pathway

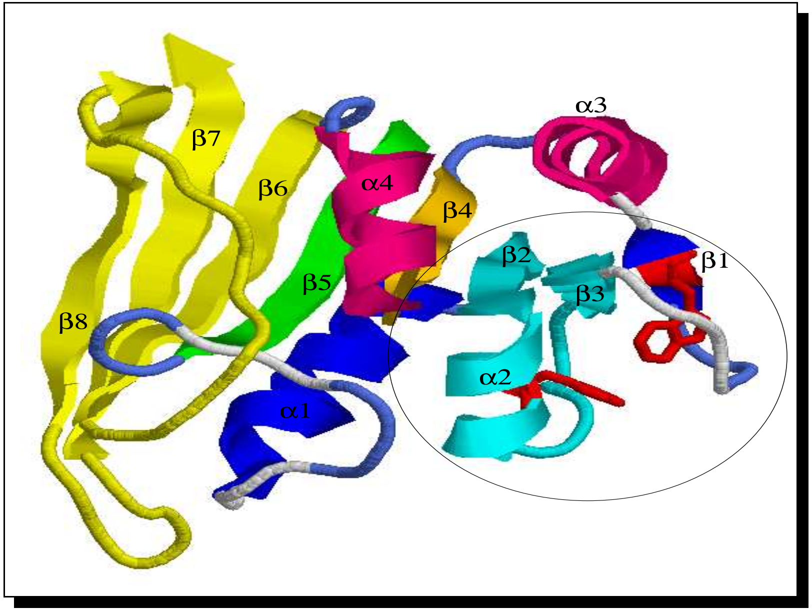

Fig. 1.11. Dihydrofolate Reductase (4DFR): Final Stages and Native Structure of the Folding Pathway

in Figure 1.10; the Trp-47 and Trp-74 interaction is also shown, and the other group now becomes $\beta_5, \beta_6, \beta_8, \beta_7$. The final native structure including $\alpha_3\beta_4\alpha_4$ and α_1 is shown in Figure 1.11. We again underscore that the results should not be taken to imply a *strict* folding timeline, but rather as a way to understand major events that are mandatory in the folding pathway. Once such experimentally verified case is the $\{\beta_2, \alpha_2, \beta_3, \beta_1\}$-group that is known to fold early, and our approach was able to predict that.

1.4 Pathways for other Proteins

To establish the utility of our methodology we predict the folding pathway for several proteins for which there are known intermediate stages in the folding pathway.

Bovine Pancreatic Trypsin Inhibitor (PDB 6PTI; 58 residues) is a small protein containing 2 α-helices and 2 β-strands [13]. It is known that the unfolding pathway of this protein involves the loss of the helix structure followed by the beta structure. Applying UNFOLD to 6PTI, we found that indeed $\beta_2\beta_3$ remain together until the end.

Chymotrypsin Inhibitor 2 (PDB 2CI2; 83 residues) is also a small protein with 1 helix and 4 strands, arranged in sequence as follows $\beta_1\alpha_1\beta_4\beta_3\beta_2$. Previous experimental and simulation studies have suggested an early displacement of β_1, and a key event in the disruption of the hydrophobic core formed primarily by α_1 and the strands β_3 and β_4 [16]. UNFOLD predicts that β_1 is the first to go, while $\beta_3\beta_4$ remain intact until the end.

The activation domain of Human Procarboxypeptidase A2 (PDB 1O6X) has 81 residues, with two α and three β strands arranged as follows $\beta_2\alpha_1\beta_1\alpha_2\beta_3$. The folding nucleus of 1O6X is made by packing of α_2 with $\beta_2\beta_1$ [28]. We found that the unfolding sequence indeed retains $\beta_2\beta_1\alpha_2$ and then finally $\beta_2\beta_1$.

The pathway of cell-cycle protein p13suc1 (PDB 1SCE; 112 residues) shows the stability of $\beta_2\beta_4$ interaction even though β_4 is the strand involved in domain swapping [2]. 1SCE has 4 domains, with 7 SSEs (3 α and 4 β). β_{4C} of domain C interacts with β_2 of domain A, and vice versa (the same is true for domains B and D). We found that $\beta_1\beta_2\beta_{4C}$ is the last to unfold.

β-Lactoglobulin (PDB 1CJ5; 162 residues) contains 10 strands and 3 helices. Beta strands F, G and H are formed immediately once the refolding starts [15], which was thus identified as the folding core of 1CJ5. In the predicted unfolding sequence obtained for 1DV9, we found that the SSEs $\beta_8, \beta_9, \beta_{10}$ corresponding to the F,G and H beta strands remain together till the last stages of unfolding.

Interleukin-1β (PDB 1I1B; 153 residues) is an all-β protein with 12 β-strands. Experiments indicate that strands $\beta_6\beta_7\beta_8$ are well folded in the intermediate state and $\beta_4\beta_5$ are partially formed [9]. We found $\beta_4\beta_5$ and $\beta_6\beta_7$ to be among the last unfolding units, including $\beta_8\beta_9$.

Myoglobin (PDB 1MBC; from sperm whale; 153 residues) and Leghemoglobin (PDB 1BIN; from Soybean; 143 residues), both belonging to the globin family of heme binding proteins, share a rather low sequence similarity, but share highly similar structure. Both are all-α proteins with 8 helices, denoted $\alpha_1(A)\alpha_2(B)\alpha_3(C)\alpha_4(D)\alpha_5(E)\alpha_6(F)\alpha_7(G)\alpha_8(H)$. In [19], they observed that the main similarity of their folding pathways is in the stabilization of the G and H helices in the burst phase folding intermediates. However, the details of the folding pathways are different. In 1MBC intermediate additional stabilizing interactions come from helices A and B, while in 1BIN they come form part of E helix. Running UNFOLD on 1MBC indeed finds that $\alpha_7(G)\alpha_8(H)$ remain together until the very last. For 1BIN we found a pathway passing through $\alpha_1(A)\alpha_2(B)\alpha_7(G)\alpha_8(H)$. UNFOLD was thus able to detect the similarity in the folding pathways, but not the details. For that we ran UNFOLD multiple times with different contact thresholds and we enumerated all exact mincuts and those mincuts within some ϵ of a mincut. From these different pathways we counted the number of times a given group of SSEs appears together. We found that $\alpha_5(E)$ showed a tendency to interact with $\alpha_8(H)$ in 1BIN, but never for 1MBC. This seems to hint at the results from experiments [19].

Protein Acylphosphatase (PDB 2ACY; 98 residues), with two α and five β SSEs ($\beta_2\alpha_1\beta_4\beta_3\alpha_2\beta_1\beta_5$), displays a transition state ensemble with a marked tendency for the β-sheets to be present, particularly β_3 and β_4, and while α_2 is present, it is highly disordered relative to rest of the structure [27]. UNFOLD finds that $\beta_2\beta_1$ remain intact until the end of unfolding, passing through a stage that also includes $\beta_3, \beta_4, \alpha_2$. To gain further insight we ran UNFOLD multiples times (as described for 1MBC and 1BIN), and we found that there was a marked tendency for $\beta_3\beta_4$ to be together in addition to $\beta_2\beta_1$, and β_3 also interacted with α_2.

Twitchin Immunoglobulin superfamily domain protein (PDB 1WIT; 93 residues) has a β-sandwich consisting of nine β-strands, and one very small helix. The folding nucleus consists of residues in the structural core $\beta_3\beta_4\beta_7\beta_9\beta_{10}$ centered around β_3 and β_9 on opposite sheets [8]. We found in multiple runs of UNFOLD this group does indeed have a very high tendency to remain intact.

1.5 Conclusions

In this paper we developed automated techniques to predict protein folding pathways. We construct a weighted SSE graph for a protein, where each vertex is a SSE, and each edge represents the strength of contacts between two SSEs. We use a repeated mincut approach (via the UNFOLD algorithm) on the WSG graph to discover strongly inter-related groups of SSEs and we then predict an (approximate) order of appearance of SSEs along the folding pathway.

Currently we consider interactions only among the α-helices and β-strands. In the future we also plan to incorporate the loop regions in the WSG, and see what effect it has on the folding pathway. Furthermore, we plan to test our folding pathways on the entire collection of proteins in the PDB. We would like to study different proteins from the same family and see if our method predicts consistent pathways; both similarities and dissimilarities may be of interest. We also plan to make our software available online so other researchers may first try the UNFOLD predictions before embarking on time-consuming experiments and simulations.

One limitation of the current approach is that our UNFOLD algorithm (arbitrarily) picks only one mincut out of perhaps several mincuts which have the same capacity. It would be interesting to enumerate all possible mincuts recursively, and construct all the possible folding pathways. If some mincuts appear on several pathways, that might provide stronger evidence of intermediate states.

Another limitation is that all native interactions are considered energetically equivalent, and thus large stabilizing interactions are not differentiated. Nevertheless the simplified model is based on topology and it helps investigate how much of the folding mechanism can be inferred from the native structure alone, without worrying about energetic frustration. Further justification for our model comes from the fact that many independent lines of investigation indicate that protein folding rates and mechanisms are largely determined by the topology of the protein [5], which is captured by our WSG model.

Glossary

Protein Structure Prediction: Given a protein amino acid sequence (i.e., linear structure), determine its three dimensional folded shape (i.e., tertiary structure).

Protein Pathway Prediction: Given a protein amino acid sequence and its three dimensional structure, determine the time ordered sequence of folding events, called the folding pathway, that leads from the linear structure to the tertiary structure.

Protein Contact Map: A binary, symmetric matrix indicating for each pair of amino acids, whether they are in contact or not.

Secondary Structure Element (SSE): Either an α-helix or β-Strand, two of the most common secondary structures found in proteins.

Weighted SSE Graph: A graph representation of a protein, where the vertices are the SSEs and the edges denote strength of interaction between the secondary structures.

Acknowledgments

This work was supported in part by NSF Career Award IIS-0092978, DOE Career Award DE-FG02-02ER25538, NSF grant EIA-0103708 and NSF grant EIA-0229454.

References

1. Ravindra K. Ahuja, Thomas L. Magnanti, and James B. Orlin. *Network Flows: Theory, Algorithms, and Applications.* Prentice Hall, Englewood Cliffs, NJ, 1993.
2. Darwin OV Alonso, Eric Alm, and Valerie Daggett. The unfolding pathway of the cell cycle protein p13suc1: Implications for domain swapping. *Structure*, 8(1):101–110, December 2000.
3. S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17), 3389-402, 1997.
4. C. Anfinsen and H. Scheraga. Experimental and theoretical aspects of protein folding. *Advances in Protein Chemistry*, 29, 205-300, 1975.
5. D. Baker. A surprising simplicity to protein folding. *Nature*, 405:39–42, May 2000.
6. S. Bryant. Evaluation of threading specificity and accuracy. *Proteins*, 26(2), 172-85, 1996.
7. G. Chikenji and M. Kikuchi. What is the role of non-native intermediates of beta-lactoglobulin in protein folding? *Proceedings of the National Academy of Sciences, USA*, 97:14273–7, 2000.
8. J. Clarke, E. Cota, S.B. Fowler, and S.J. Hamill. Folding studies of immunoglobulin-like β-sandwich proteins suggest that they share a common folding pathway. *Structure*, 7:1145–1153, September 1999.
9. C. Clementi, P. A. Jennings, and J.N. Onuchic. How native-state topology affects the folding of dihydrofolate reductase and interleukin-1beta. *Proceedings of the National Academy of Sciences, USA*, 97(11):5871–6, 2000.
10. W. Colon and H. Roder. Kinetic intermediates in the formation of the cytochrome c molten globule. *Nature Structural Biology*, 3(12), 1019-25, 1996.
11. D.K. Heidary, Jr. J. C. O'Neill, M. Roy, and P.A. Jennings. An essential intermediate in the folding of dihydrofolate reductase. *Proceedings of the National Academy of Sciences, USA*, 97(11):5866–70, 2000.
12. P.A. Jennings, B. E. Finn, and et al. A reexamination of the folding mechanism of dihydrofolate reductase from escherichia coli: verification and refinement of a four-channel model. *Biochemistry*, 32(14):3783–9, 1993.
13. Steven L. Kazmirski and Valerie Daggett. Simulations of the structural and dynamical properties of denatured proteins: The molten coil state of bovine pancreatic trypsin inhibitor. *Journal of Molecular Biology*, 277:487–506, 1998.
14. D.K. Klimov and D. Thirumalai. Multiple protein folding nuclei and the transition state ensemble in two-state proteins. *Proteins*, 43:465–75, 2001.
15. Kazuo kuwata, Ramachandra Shastry, Hong Cheng, Masaru Hoshino, Carl A. Bhatt, Yuji Goto, and Heinrich Roder. Structural and kinetic characterization of early folding events of lactoglobnulin. *Nature*, 8(2):151–155, 2001.

16. T. Lazardis and M. Karplus. New view of protein folding reconciled with the old through multiple unfolding simulations. *Science*, 278:1928–1931, December 1997.
17. C. Levinthal. Are there pathways for protein folding? *Journal of Chemical Physics*, 65, 44-45, 1968.
18. Y. Mok, C. Kay, L. Kay, and J. Forman-Kay. NOE data demonstrating a compact unfolded state for an sh3 domain under non-denaturing conditions. *Journal of Molecular Biology*, 289(3):619–638, 1999.
19. C. Nishimura, S. Prytulla, H.J. Dyson, and P.E. Wright. Conservation of folding pathways in evolutionary distant globin sequences. *Nature Structural Biology*, 7(8):679–686, August 2000.
20. B. Nolting, R. Golbik, J. Neira, A. Soler-Gonzalez, G. Schreiber, and A. Fersht. The folding pathway of a protein at high resolution from microseconds to seconds. *Proceedings of the National Academy of Sciences, USA*, 94(3), 826-30, 1997.
21. K.T. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *Journal of Molecular Biology*, 268(1), 209-25, 1997.
22. M. Sippl. Helmholtz free energy of peptide hydrogen bonds in proteins. *Journal of Molecular Biology*, 260(5), 644-8, 1996.
23. J. Skolnick, A. Kolinski, and A. Ortiz. Derivation of protein-specific pair potentials based on weak sequence fragment similarity. *Proteins*, 38(1), 3-16, 2000.
24. M.J.E. Sternberg and J.M. Thornton. On the conformation of proteins: The handedness of the connection between parallel beta strands. *Journal of Molecular Biology*, 110:269–283, 1977.
25. M. Stoer and F. Wagner. A simple min-cut algorithm. *Journal of the ACM*, 44(4):585–91, July 1997.
26. M. Vendruscolo, E. Kussell, and E. Domany. Recovery of protein structure from contact maps. *Folding and Design*, 2(5):295–306, September 1997.
27. M. Vendruscolo, E. Paci, C.M. Dobson, and M. Karplus. Three key residues form a critical contact network in a protein folding transition state. *Nature*, 409:641–645, February 2001.
28. V. Villegas, J.C. Martinez, F.X. Aviles, and L. Serrano. Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain. *Journal of Molecular Biology*, 283:1027–1036, 1998.
29. D. R. Westhead, T. W. F. Slidel, T. P. J. Flores, and J. M. Thornton. Protein structural topology: automated analysis, diagrammatic representation and database searching. *Protein Science*, 8:897–904, 1999.