BMC Bioinformatics

Research article

Predicting protein folding pathways at the mesoscopic level based on native interactions between secondary structure elements

Qingwu Yang[1] and Sing-Hoi Sze*[1,2]

Address: [1]Department of Computer Science, Texas A&M University, College Station, TX 77843, USA and [2]Department of Biochemistry & Biophysics, Texas A&M University, College Station, TX 77843, USA

Email: Qingwu Yang - qingwu-yang@neo.tamu.edu; Sing-Hoi Sze* - shsze@cs.tamu.edu

* Corresponding author

Published: 23 July 2008

BMC Bioinformatics 2008, **9**:320 doi:10.1186/1471-2105-9-320

Received: 5 April 2008
Accepted: 23 July 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/320

Abstract

Background: Since experimental determination of protein folding pathways remains difficult, computational techniques are often used to simulate protein folding. Most current techniques to predict protein folding pathways are computationally intensive and are suitable only for small proteins.

Results: By assuming that the native structure of a protein is known and representing each intermediate conformation as a collection of fully folded structures in which each of them contains a set of interacting secondary structure elements, we show that it is possible to significantly reduce the conformation space while still being able to predict the most energetically favorable folding pathway of large proteins with hundreds of residues at the mesoscopic level, including the pig muscle phosphoglycerate kinase with 416 residues. The model is detailed enough to distinguish between different folding pathways of structurally very similar proteins, including the streptococcal protein G and the peptostreptococcal protein L. The model is also able to recognize the differences between the folding pathways of protein G and its two structurally similar variants NuG1 and NuG2, which are even harder to distinguish. We show that this strategy can produce accurate predictions on many other proteins with experimentally determined intermediate folding states.

Conclusion: Our technique is efficient enough to predict folding pathways for both large and small proteins at the mesoscopic level. Such a strategy is often the only feasible choice for large proteins. A software program implementing this strategy (SSFold) is available at http://faculty.cs.tamu.edu/shsze/ssfold.

Background

As early studies revealed that an unfolded protein can fold spontaneously to a three-dimensional structure under suitable environmental conditions [1,2], traditional approaches to understanding protein folding have focused on the prediction of the native structure. As more studies showed the existence of intermediates and interaction among residues during the protein folding process [3,4], there is substantial interest to understand the time order of events during the formation of the tertiary structure. From the free energy point of view, each conformation of a protein is associated with a free energy and the protein folds from the high-energy denatured conformation to its folded structure along a funnel-like energy landscape [5,6].

Although advances in experimental techniques allow the investigation of protein folding pathways at the microsecond timescale [7,8], experimental determination of protein folding pathways remains difficult. Most studies are only able to identify general characteristics of the folding pathway without much details and are limited to analyzing small proteins. Computational techniques are often used to simulate protein folding and the problem is transformed to energetic optimization problems, that is, computational search for global energy minimum over all possible conformations. The most accurate computational techniques utilize molecular dynamics to determine the order of events that lead to the tertiary structure through atomic-level simulations [9-12]. Due to the extremely large conformation space, these approaches suffer from well-known problems accompanying high dimensionality, including computational expensiveness and ease of trapping in local minima, and are applicable only to small proteins in a short time course.

By omitting some details, proteins can be represented at the level of amino acids. Kolinski and Skolnick [13] performed Monte Carlo simulations of protein folding on a reduced lattice representation of the protein α-carbon backbone. Yue and Dill [14] limited the conformation space to a discrete subset of possibilities and used a branch-and-bound procedure to search for near-optimal conformations. Alm and Baker [15] and Muñoz and Eaton [16] further observed that the availability of the known native structure can dramatically reduce the search space. Alm and Baker [15] took into account only native interactions among residues and used a sequential binary collision model to predict protein folding mechanisms from the perspectives of free energy landscapes, while Muñoz and Eaton [16] used a slightly different approach of employing distinct free energy costs for different secondary structures. Amato and Song [17] represented a protein by the torsional angles of its residues and used the probabilistic roadmap technique with a biased sampling strategy around the native structure to predict folding pathways and secondary structure formation order. Liwo et al [18] and Kmiecik and Kolinski [19,20] showed that the use of reduced models of proteins is highly successful in characterizing folding pathways for small proteins at the mesoscopic level. Although these techniques are able to predict folding pathways very accurately for proteins with up to about 100 residues, the majority of proteins in the Protein Data Bank (PDB) [21] are much larger (Figure 1).

The problem with representing a protein at the amino acid level is that even with the assumption that each residue has only two states (ordered or disordered), a protein with n residues still has 2^n possible conformations [15]. To overcome this problem, several recent approaches represent a protein at the level of secondary structure elements (SSEs), in which each element corresponds to one helix or one β-strand. By adopting the framework model in which secondary structures are thought to fold relatively independently of the tertiary structure [22], each SSE is treated as an indivisible unit that interacts with other SSEs as a whole. Since the number of SSEs in a protein is small (Figure 1), this model is much more tractable to simulate. Eyrich et al [23] assumed that the SSEs are fixed and used a branch-and-bound algorithm to search for near-optimal tertiary structures. Apaydin et al [24] assumed that each SSE of a protein is already in native conformation and moves as a unit, and used the probabilistic roadmap approach to predict folding pathways. Zaki et al [25] proposed an algorithm to predict unfolding pathways based on applying a minimum cut procedure to a weighted graph that represents a protein's contact map or interaction strength between SSEs. Although the underlying assumption that intermediate secondary structures

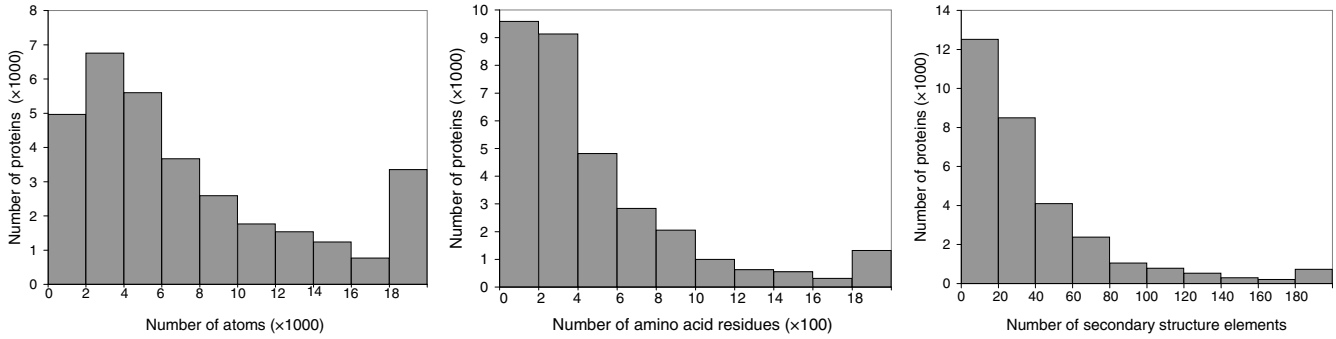

Figure 1
The distribution of the number of atoms, the number of amino acid residues, and the number of secondary structure elements among 32237 protein structures in the Protein Data Bank (PDB) [21]. Each bar (except the rightmost one in each chart) shows the number of proteins that have values falling between the indicated lower and upper limits. The rightmost bar in each chart shows the number of proteins that have values of at least the indicated lower limit.

are fully folded before the formation of tertiary structures is not satisfied for most proteins, these studies show that such a strategy is sufficient to study protein folding pathways at the mesoscopic level.

In this paper, our goal is to further reduce the conformation space without sacrificing prediction accuracy. This is achieved by assuming that SSEs that do not yet interact with each other are independent and can be treated separately. A conformation is represented by a collection of fully folded structures in which each of them contains a set of interacting SSEs. By using a steepest descent strategy, we show that it is possible to predict the most energetically favorable folding pathway of large proteins with hundreds of residues at the mesoscopic level and this model is detailed enough to distinguish between different folding pathways of structurally very similar proteins. In difference from the technique in [24], we do not consider the spatially moving process before the SSEs form native contacts, and thus we are able to achieve much better computational efficiency.

Methods

Assume that the native structure of a protein is known. The protein folding pathway prediction problem is to find an ordered sequence of intermediate conformations to fill the gap between the unfolded state and the native tertiary structure. At the secondary structure level, a protein can be viewed as an ordered sequence of secondary structure elements (SSEs) interspersed with irregular turns or loops, where each SSE is either a helix or a β-strand, and each β-sheet consists of a variable number of β-strands that are not necessarily consecutive on the primary sequence. We represent each protein by $t_0 s_1 t_1 \ \ s_k t_k$, where k is the number of SSEs, s_i denotes the ith SSE, t_j denotes the jth turn, and these elements are in the same order as they appear on the primary sequence. Given the three-dimensional structure of a protein, the assignment of SSEs can be obtained directly from the Protein Data Bank (PDB) [21] or using programs such as DSSP [26].

Following [24] and [25], we consider each SSE as an indivisible unit that folds independently of the others according to the contacts present in the native structure. This is based on the framework model that assumes that extensive intermediate secondary structures exist before they are assembled into the tertiary structure [22], and our goal is to predict the interaction order of SSEs during folding. Based on the observation in [15] and [16] that a model using only native interactions can explain most experimental results, we assume that the interactions between SSEs or turns are the same as the ones present in the native structure. Although these assumptions are often not satisfied as there are many proteins in which there are no clear secondary structures before the formation of tertiary structures or there are no clear preservations of secondary structures throughout folding, such a strategy is sufficient for studying folding pathways at the mesoscopic level and is often the only feasible choice for large proteins.

We represent a conformation of a protein on the folding pathway by $C = \{S_1, ..., S_k\}$, where each S_i represents a structure consisting of a set of fully folded SSEs and there are no interactions between two different sets S_j and $S_{j'}$ (see Figure 2 for an illustration). Since our focus is on the SSEs, turns are not included in the conformation but will be utilized when computing energies (see below). The protein folding problem is transformed to identifying a

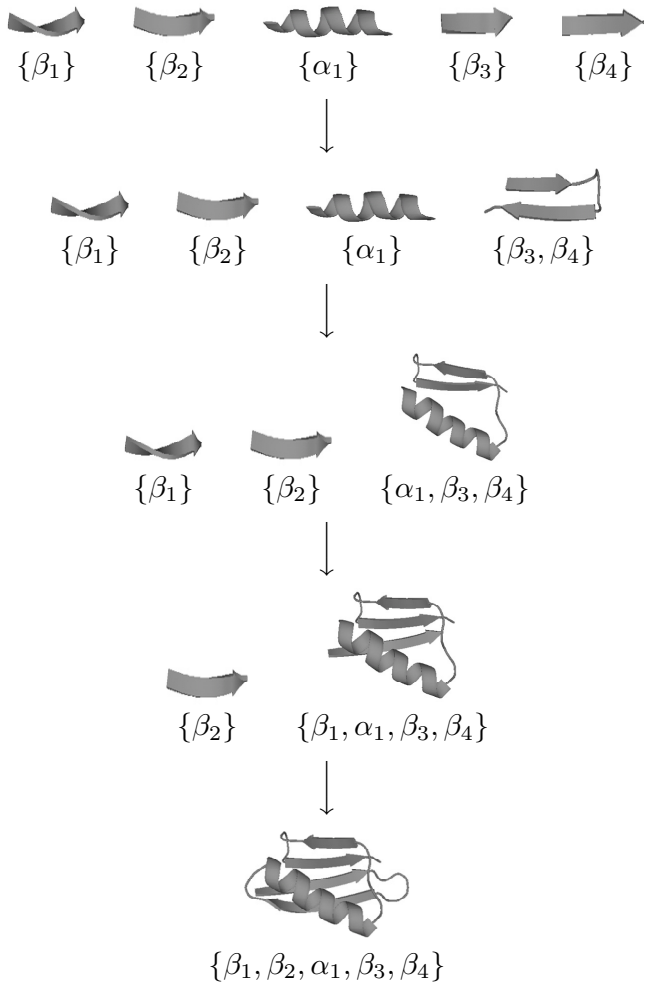

Figure 2
Illustration of the folding pathway prediction for GB1. The starting conformation {{β_1}, {β_2}, {α_1}, {β_3}, {β_4}} corresponds to the initial state. There are three intermediate conformations in the predicted folding pathway, including {{β_1}, {β_2}, {α_1}, {β_3, β_4}}, {{β_1}, {β_2}, {α_1, β_3, β_4}}, and {{β_2}, {β_1, α_1, β_3, β_4}}. The ending conformation {{β_1, β_2, α_1, β_3, β_4}} corresponds to the native state.

sequence of conformational changes that start from an initial state with fully folded SSEs but no interactions between SSEs through some intermediate conformations and ending in the native structure (Figure 2). Each conformational change corresponds to finding a new pair of interactions that merges two smaller structures of SSEs into a bigger one. Figure 2 illustrates the folding pathway prediction on the B1 domain of the streptococcal protein G (GB1). In the prediction, β_3 and β_4 interact first, then α_1 is added, followed by β_1 and β_2.

Folding pathway predictions are obtained through the computation of free energies of intermediate conformations. For an intermediate conformation $C = \{S_1, ..., S_k\}$, the free energy $E(C)$ of C is defined as:

$$E(C) = \sum_{i=1}^{k} E(S_i),$$

where each S_i is viewed as an isolated entity and each $E(S_i)$ is obtained separately by extracting the three-dimensional coordinates of its residues from the Protein Data Bank (PDB) [21] and using the Rosetta software [27] to compute its free energy. The original Rosetta energy function is used, which is obtained by representing each side chain by a centroid that is located at the center of mass, and computing a weighted sum of the binned probability descriptions of multiple effects, including the solvation and electrostatic effects based on observed distributions in known protein structures, the secondary structure packing effects that include strand pairing, strand arrangement into sheets and helix-strand packing, and the effects of steric repulsion and Van der Waals interactions (more details are available in [28] and in Table I of [27]). To take the backbone into consideration, a turn t_j is included in the computation of $E(S_i)$ if both of its adjacent SSEs s_j (if it exists) and s_{j+1} (if it exists) are included in S_i.

Since the interactions that favor folding usually decrease the free energy while the interactions that destabilize the native structure increase the free energy, our goal is to find the most energetically favorable folding pathway by identifying the conformational change that decreases the free energy the most in each step so that the protein can get to lower energy states as quickly as possible. Figure 3 illustrates our SSFold algorithm that uses a steepest descent strategy to choose a new pair of interactions that leads to a conformation with the lowest free energy in each iteration. This procedure is very efficient since only k - 1 iterations are needed. Within each iteration, $O(k^2)$ comparisons are needed to find the best pair of interactions that results in the lowest free energy. This leads to an overall time complexity of $O(k^3t)$, where k is the number of SSEs in a protein and t is the time to compute the free energy of a potentially partial protein that contains only some of the SSEs and turns.

input: a protein with k SSEs $s_1, \ldots, s_k$;
output: prediction of interaction order of SSEs;

$C \leftarrow \{\{s_1\}, \ldots, \{s_k\}\}$;
while $|C| > 1$ do {
  choose $S \in C$ and $S' \in C$ such that the energy $E((C - \{S, S'\}) \cup \{S \cup S'\})$ is minimized;
  $C \leftarrow (C - \{S, S'\}) \cup \{S \cup S'\}$; }

Figure 3
Algorithm SSFold to predict the most energetically favorable interaction order of SSEs that corresponds to a folding pathway. Each iteration corresponds to a conformational change that results from a new pair of interactions. Within a folded structure, a turn is included in the energy computations only when adjacent SSEs are included in the structure.

Results

We test our strategy on proteins from the Protein Data Bank (PDB) [21] that have known intermediate folding states from experimental data. We illustrate that our model is detailed enough to distinguish between subtle differences in the folding pathways of the streptococcal protein G, the peptostreptococcal protein L, and variants NuG1 and NuG2 of protein G, which are all structurally very similar proteins. We demonstrate that our approach is applicable to large proteins with hundreds of residues by testing it on the 416 residue pig muscle phosphoglycerate kinase (PGK). We further test it on proteins studied in [29] and [25] to validate that our model has very good accuracy.

Proteins GB1, LB1, NuG1 and NuG2

The 56 residue B1 immunoglobulin binding domain of streptococcal protein G (GB1, PDB: 1GB1) and the 62 residue B1 immunoglobulin binding domain of peptostreptococcal protein L (LB1, PDB: 2PTL) have been used extensively as model systems for studying protein folding mechanisms [30-37]. Both GB1 (see Figure 2) and LB1 consist of one β-sheet with four strands and one α-helix. Strands 1 and 2 form an N-terminal β-hairpin, while strands 3 and 4 form a C-terminal β-hairpin. Although GB1 and LB1 have very similar tertiary structures, they have different folding pathways. As suggested by [29], a detailed model is needed to distinguish between them.

Figure 4 shows our folding pathway predictions for GB1 and LB1 (see also Figure 2 for GB1).

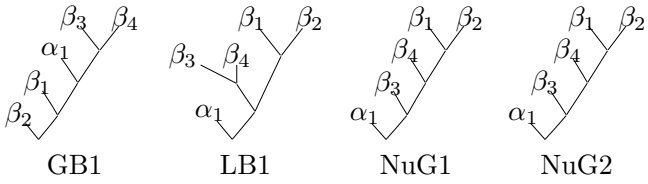

Figure 4
Folding pathway predictions for GB1, LB1, NuG1 and NuG2. Each internal node represents a new pair of interactions and nodes that are higher in the tree indicate earlier interactions. Also compare to Figure 2 for GB1.

Experimental results showed that the C-terminal β-hairpin in GB1 is formed in the transition state of the folding pathway and serves as the starting point on which the rest of the protein can fold [35]. Similar results were obtained using the diffusion-collision model [38]. Our prediction is consistent with these results. In contrast, experimental results showed that only the N-terminal β-hairpin in LB1 is mainly formed in the transition state and non-random structures can be detected in the region [34,39]. Our algorithm also predicts that the N-terminal β-hairpin forms earlier than the C-terminal β-hairpin in LB1.

Two protein G variants, NuG1 (PDB: 1MHX) and NuG2 (PDB: 1MI0), were designed to have a different folding mechanism from protein G by replacing some residues of protein G [36]. In NuG1 and NuG2, the stability of the N-terminal β-hairpin is enhanced while the stability of the C-terminal β-hairpin is reduced, with the N-terminal β-hairpin forming contacts earlier than the C-terminal β-hairpin in both cases [36].

Thomas et al [40] showed that it is more difficult to distinguish between the folding pathways of protein G and its variants NuG1 and NuG2 than to distinguish between the folding pathways of protein G and protein L. In our predictions in Figure 4, NuG1 and NuG2 have the same folding pathway, with the N-terminal β-hairpin folded first. This is consistent with the experimental results in [41] and the predictions in [40].

Figure 5 shows the free energy profiles of GB1, LB1, NuG1 and NuG2 in our predictions. Our predicted folding pathway of GB1 is a non-frustrated curve, similar to the average macroscopic folding pathway given by [37]. When compared to GB1, NuG1 and NuG2 have similar profiles and higher initial free energy, but their native structures have lower free energy and are more stable, which is consistent with the analysis in [41].

Pig muscle PGK: a large protein

Phosphoglycerate kinase (PGK) from various organisms has been used as a model system for studying domain-

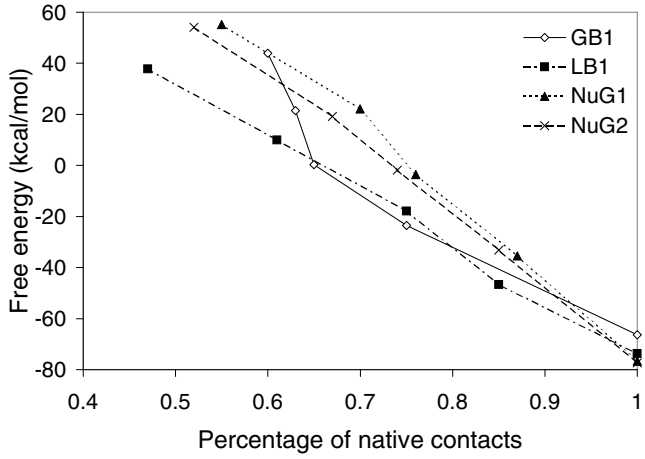

Figure 5
Free energy profiles of GB1, LB1, NuG1 and NuG2 in our predictions. A native contact is defined to be a pair of amino acids that have their α-carbon atoms within 7 Å of each other. Each starting point corresponds to the initial state in which each SSE has already completed its native fold independently and there are no interactions between SSEs.

domain interactions of multiple-domain proteins [42-44]. The pig muscle PGK (PDB: 1KF0) [43] is a large two-domain protein with 416 residues, with the N-terminal domain consisting of residues 1 to 155 and the C-terminal domain consisting of residues 156 to 416. There are 21 α-helices and 17 β-strands, which belong to four different β-sheets A, B, C and D, arranged as follows on the primary sequence: $\alpha_1 \beta_{A4} \alpha_2 \alpha_3 \beta_{A3} \alpha_4 \beta_{A1} \alpha_5 \beta_{A2} \alpha_6 \beta_{B1} \beta_{B2} \alpha_7 \beta_{A5} \alpha_8 \alpha_9 \beta_{A6} \alpha_{10} \beta_{C3} \alpha_{11} \alpha_{12} \beta_{C2} \alpha_{13} \alpha_{14} \alpha_{15} \beta_{C1} \beta_{D2} \beta_{D1} \beta_{D3} \alpha_{16} \beta_{C4} \alpha_{17} \alpha_{18} \beta_{C5} \alpha_{19} \beta_{C6} \alpha_{20} \alpha_{21}$.

Figure 6 shows our folding pathway prediction for the pig muscle PGK, in which β-sheet D is formed first, followed by the formation of β-sheet C interspersed with α-helices in the C-terminal domain. After most SSEs of the C-terminal domain are formed, the SSEs of the N-terminal domain begin to form, with β-sheet A formed before β-sheet B interspersed with α-helices in the N-terminal domain.

Szilágyi and Vas [45] suggested a sequential domain refolding mechanism for the pig muscle PGK, in which folding of the C-terminal domain is independent of the N-terminal domain and takes place first, and folding of the N-terminal domain starts after most of the C-terminal domain folds. The authors also suggested that an intermediate consists of a folded C-terminal domain and a still unfolded N-terminal domain. Our prediction is consistent with these experimental results.

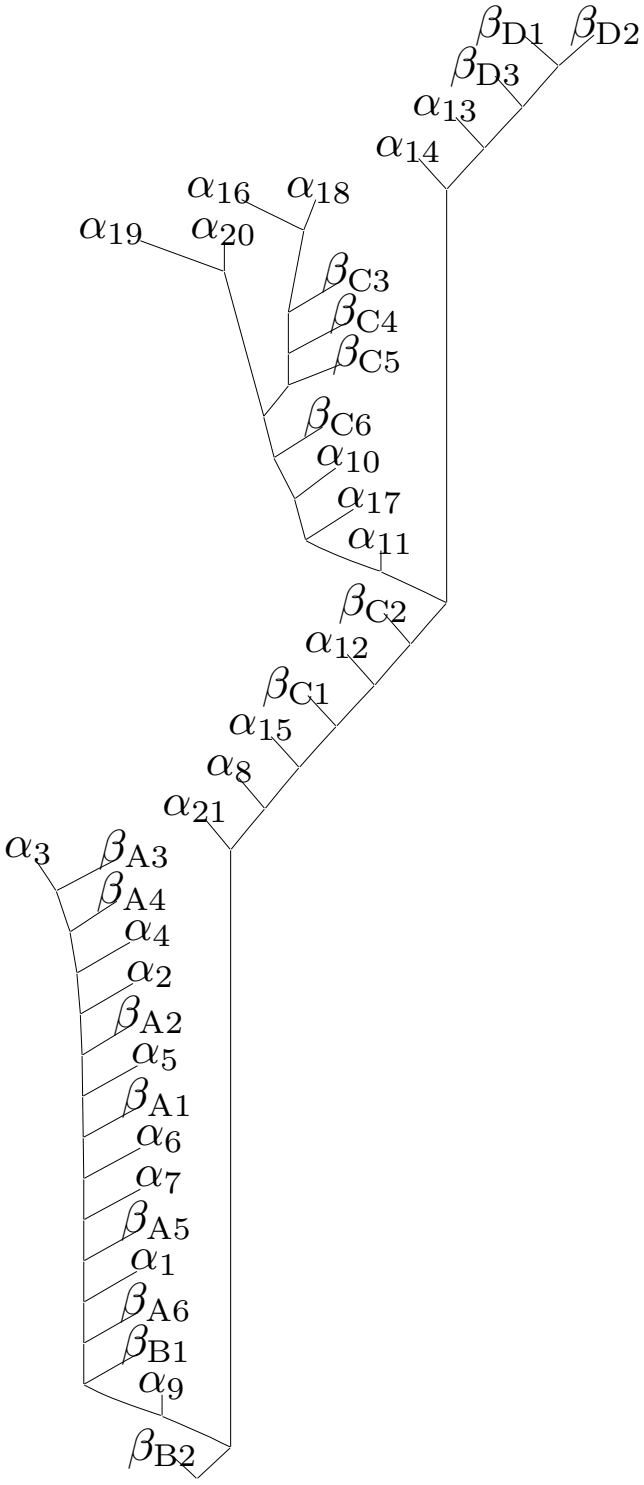

Figure 6
Folding pathway prediction for the pig muscle PGK.

Other proteins

Figure 7 shows folding pathway predictions for various small proteins that have known intermediate folding states from biological experiments. The proteins 1BDD and 2CRT were studied in [29], while the proteins 1BIN, 1MBC, 2CI2 and 6PTI were studied in [25].

The B domain of *Staphylococcus aureus* protein A (PDB: 1BDD) consists of three α-helices. In our prediction, α_2 and α_3 interact first, then α_1 is added. This is consistent with the result of the out-exchange experiment in [46] and experimental results under high temperature [47].

Although two members of the globin protein family, leghemoglobin A (PDB: 1BIN) and myoglobin (PDB: 1MBC), have very low sequence similarity, they both consist of eight α-helices and have very similar tertiary structures. Nishimura et al [48] compared their folding pathways experimentally. For leghemoglobin A, α_G, α_H, and part of α_E form stable structures first, while α_A and α_B form in the later stages of the folding pathway. For myoglobin, α_A, α_G and α_H form stable contacts first. The main difference between the two folding pathways is that α_A and α_B form earlier in the folding pathway of myoglobin than in the folding pathway of leghemoglobin A [48]. Our predictions are able to distinguish between these subtle differences. For leghemoglobin A, α_G and α_H are predicted to interact first, then α_E is added, with α_B and α_A added later. For myoglobin, α_G and α_H are also predicted to interact first, then α_A is added, followed by α_E and α_B.

There are two crystal structures for chymotrypsin inhibitor 2 (PDB: 1COA and 2CI2). While 2CI2 consists of 83 residues, 1COA is a fragment of 2CI2 from residues 20 to 83. They both consist of one α-helix and four β-strands, which are arranged as $\beta_1\alpha_1\beta_2\beta_3\beta_4$ in 1COA and $\beta_1\alpha_1\beta_4\beta_3\beta_2$ in 2CI2. In our predictions, 1COA and 2CI2 have the same folding pathway, with the middle two β-strands interacting first, then the α-helix is added, followed by the C-terminal β-strand, and the N-terminal β-strand is added last. For 1COA, simulation by [49] demonstrated that β_2 and β_3 form contacts first, then α_1 is added to form a folding nucleus. The coalescence of β_1 is the rate-limiting step and is completed at the end of the folding process. This is consistent with the result of the out-exchange experiment in [46] that showed that β_2, β_3 and α_1 form contacts first. Our prediction is consistent with these results.

The all β-sheet protein cardiotoxin III (PDB: 2CRT) consists of five strands. While β_1 and β_2 form a double-stranded domain, β_3, β_4 and β_5 form a triple-stranded domain. By the amide proton pulse exchange experiment, Sivaraman et al [50] showed that the triple-stranded domain forms earlier than the double-stranded domain during the refolding process. The carbonyl groups in β_3 and the amide groups in β_5 form hydrogen bonding part-

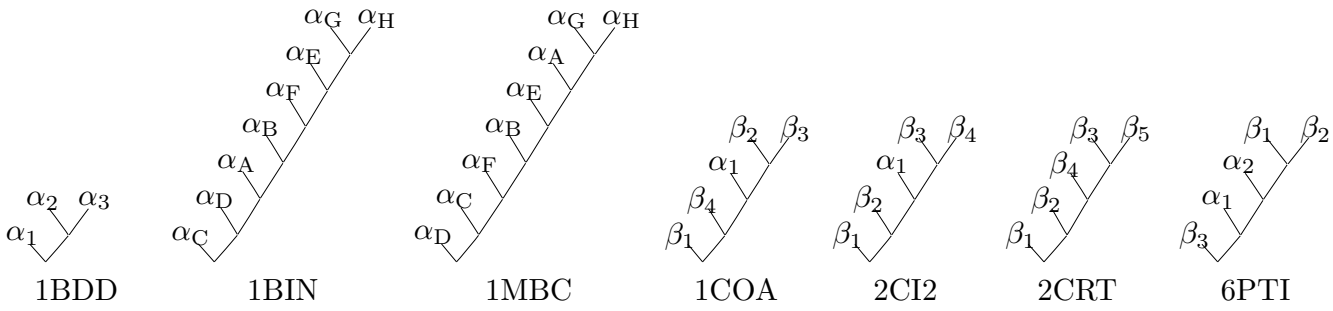

Figure 7
Folding pathway predictions for *Staphylococcus aureus* protein A domain B (PDB: 1BDD), leghemoglobin A (PDB: 1BIN), myoglobin (PDB: 1MBC), chymotrypsin inhibitor 2 structure 1 (PDB: 1COA), chymotrypsin inhibitor 2 structure 2 (PDB: 2CI2), cardiotoxin III (PDB: 2CRT), and bovine pancreatic trypsin inhibitor BPTI (PDB: 6PTI).

ners, which are important for the formation of a hydrophobic cluster [50]. Our prediction is consistent with these results, with β_3 and β_5 interacting first, then β_4 is added to form the triple-stranded domain, followed by β_2 and β_1 in the double-stranded domain.

Bovine pancreatic trypsin inhibitor BPTI (PDB: 6PTI) is a globular protein with two α-helices and three β-strands, which are arranged as $\alpha_1\beta_2\beta_1\beta_3\alpha_2$. Three disulfide bonds between residues 5 and 55, 14 and 38, and 30 and 51 play an important role in stabilizing the native structure [51], and their formation order was studied in [52]. In our prediction, β_1 and β_2 interact first, then α_2 is added. This brings residues 30 and 51 close together and helps to form the disulfide bond between them. Then α_1 is added and this helps to form the disulfide bond between residues 5 and 55, and 14 and 38. Our prediction that β_1 and β_2 interact earlier than the two α-helices is consistent with the result in [53].

Discussion

While our strategy corresponds most closely to the diffusion-collision model that allows folding to proceed independently in different parts of a protein [54], it is possible to use a modified strategy for other models. For example, to simulate the nucleation-propagation model [55] or the nucleation-condensation model [56], in which the existence of a nucleus facilitates further folding, one can iteratively add a SSE that results in the lowest free energy to the nucleus. Since energy computations can still be slow and can take hours, which account for significant amount of computation time in our algorithm, it is also possible to use lower resolution methods to compute energy.

While our strategy finds the most energetically favorable protein folding pathway, there are evidences that multiple folding pathways exist [5,57]. The ability to analyze multiple folding pathways will also allow the study of protein misfolding [58]. Our approach can be generalized to study the entire free energy landscape [5] as follows: construct a graph in which each vertex represents a biologically plausible conformation and each edge represents a feasible conformation change, which is similar to the roadmap graph in [24] and [17] and the protein folding network in [59] except that we consider each SSE as an indivisible unit. Various graph-theoretic algorithms can then be used to generate predictions of alternative folding pathways.

Conclusion

We have shown that our procedure has sufficient accuracy to distinguish between subtle differences and our strategy can be applied to large proteins due to its speed. An important future direction is to consider cooperative folding of secondary structures without too much sacrifice in speed, that is, when folding in one secondary structure affects folding in others.

Authors' contributions

QY performed the research and implemented the algorithm. S–HS supervised the research. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by NSF grant DBI-0624077. We thank Yutu Liu for many helpful discussions and for drawing our attention to the problem.

References

1. Levinthal C: **Are there pathways for protein folding?** *J Chim Phys* 1968, **65:**44-45.
2. Anfinsen C B, Scheraga H A: **Experimental and theoretical aspects of protein folding.** *Adv Protein Chem* 1975, **29:**205-300.
3. Kim P S, Baldwin R L: **Intermediates in the folding reactions of small proteins.** *Ann Rev Biochem* 1990, **59:**631-660.
4. Matthews C R: **Pathways of protein folding.** *Ann Rev Biochem* 1993, **62:**653-683.
5. Dill K A, Chan H S: **From Levinthal to pathways to funnels.** *Nat Struct Biol* 1997, **4:**10-19.
6. Gruebele M: **Protein folding: the free energy surface.** *Curr Opin Struct Biol* 2002, **12:**161-168.

7. Eaton WA, Muñoz V, Thompson PA, Chan CK, Hofrichter J: **Submillisecond kinetics of protein folding.** *Curr Opin Struct Biol* 1997, **7:**10-14.
8. Nölting B, Golbik R, Neira J L, Soler Gonzalez A S, Schreiber G, Fersht A R: **The folding pathway of a protein at high resolution from microseconds to seconds.** *Proc Natl Acad Sci USA* 1997, **94:**826-830.
9. Levitt M: **Protein folding by restrained energy minimization and molecular dynamics.** *J Mol Biol* 1983, **170:**723-764.
10. Daggett V, Levitt M: **Realistic simulations of native-protein dynamics in solution and beyond.** *Ann Rev Biophys Biomol Struct* 1993, **22:**353-380.
11. Duan Y, Kollman P A: **Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution.** *Science* 1998, **282:**740-744.
12. Daggett V: **Molecular dynamics simulations of the protein unfolding/folding reaction.** *Acc Chem Res* 2002, **35:**422-429.
13. Kolinski A, Skolnick J: **Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme.** *Proteins* 1994, **18:**338-352.
14. Yue K, Dill K A: **Folding proteins with a simple energy function and extensive conformational searching.** *Protein Sci* 1996, **5:**254-261.
15. Alm E, Baker D: **Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures.** *Proc Natl Acad Sci USA* 1999, **96:**11305-11310.
16. Muñoz V, Eaton W A: **A simple model for calculating the kinetics of protein folding from three-dimensional structures.** *Proc Natl Acad Sci USA* 1999, **96:**11311-11316.
17. Amato N M, Song G: **Using motion planning to study protein folding pathways.** *J Comput Biol* 2002, **9:**149-168.
18. Liwo A, Khalili M, Scheraga HA: ***Ab initio* simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains.** *Proc Natl Acad Sci USA* 2005, **102:**2362-2367.
19. Kmiecik S, Kolinski A: **Characterization of protein-folding pathways by reduced-space modeling.** *Proc Natl Acad Sci USA* 2007, **104:**12330-12335.
20. Kmiecik S, Kolinski A: **Folding pathway of the B1 domain of protein G explored by multiscale modeling.** *Biophys J* 2008, **94:**726-736.
21. Berman H M, Westbrook J, Feng Z, Gilliland G, Bhat T N, Weissig H, Shindyalov I N, Bourne P E: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28:**235-242.
22. Nölting B, Andert K: **Mechanism of protein folding.** *Proteins* 2000, **41:**288-298.
23. Eyrich V A, Standley D M, Felts A K, Friesner R A: **Protein tertiary structure prediction using a branch and bound algorithm.** *Proteins* 1999, **35:**41-57.
24. Apaydin M S, Singh A P, Brutlag D L, Latombe J C: **Capturing molecular energy landscapes with probabilistic conformational roadmaps.** *Proceedings of the IEEE International Conference on Robotics and Automation* 2001:932-939.
25. Zaki M J, Nadimpally V, Bardhan D, Bystroff C: **Predicting protein folding pathways.** *Bioinformatics* 2004, **20 Suppl 1:**386-393.
26. Kabsch W, Sander C: **Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.** *Biopolymers* 1983, **22:**2577-2637.
27. Rohl C A, Strauss C E M, Misura K M S, Baker D: **Protein structure prediction using Rosetta.** *Methods Enzymol* 2004, **383:**66-93.
28. Simons K T, Ruczinski I, Kooperberg C, Fox B A, Bystroff C, Baker D: **Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins.** *Proteins* 1999, **34:**82-95.
29. Song G, Thomas S, Dill K A, Scholtz J M, Amato N M: **A path planning-based study of protein folding with a case study of hairpin formation in protein G and L.** *Pacific Symposium on Biocomputing* 2003:240-251.
30. Alexander P, Orban J, Bryan P: **Kinetic analysis of folding and unfolding the 56 amino acid IgG-binding domain of streptococcal protein G.** *Biochemistry* 1992, **31:**7243-7248.
31. Blanco FJ, Rivas G, Serrano L: **A short linear peptide that folds into a native stable β-hairpin in aqueous solution.** *Nat Struct Biol* 1994, **1:**584-590.
32. Gallagher T, Alexander P, Bryan P, Gilliland G L: **Two crystal structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR.** *Biochemistry* 1994, **33:**4721-4729.
33. Blanco FJ, Serrano L: **Folding of protein G B1 domain studied by the conformational characterization of fragments comprising its secondary structure elements.** *Eur J Biochem* 1995, **230:**634-649.
34. Kim D E, Fisher C, Baker D: **A breakdown of symmetry in the folding transition state of protein L.** *J Mol Biol* 2000, **298:**971-984.
35. McCallister EL, Alm E, Baker D: **Critical role of β-hairpin formation in protein G folding.** *Nat Struct Biol* 2000, **7:**669-673.
36. Nauli S, Kuhlman B, Baker D: **Computer-based redesign of a protein folding pathway.** *Nat Struct Biol* 2001, **8:**602-605.
37. Tunnicliffe R B, Waby J L, Williams R J, Williamson M P: **An experimental investigation of conformational fluctuations in proteins G and L.** *Structure* 2005, **13:**1677-1684.
38. Islam S A, Karplus M, Weaver D L: **The role of sequence and structure in protein folding kinetics: the diffusion-collision model applied to proteins L and G.** *Structure* 2004, **12:**1833-1845.
39. Yi Q, Scalley Kim M L, Alm E J, Baker D: **NMR characterization of residual structure in the denatured state of protein L.** *J Mol Biol* 2000, **299:**1341-1351.
40. Thomas S, Tang X, Tapia L, Amato N M: **Simulating protein motions with rigidity analysis.** *J Comput Biol* 2007, **14:**839-855.
41. Nauli S, Kuhlman B, Le Trong I, Stenkamp R E, Teller D, Baker D: **Crystal structures and increased stabilization of the protein G variants with switched folding pathways NuG1 and NuG2.** *Protein Sci* 2002, **11:**2924-2931.
42. Parker M J, Spencer J, Jackson G S, Burston S G, Hosszu L L, Craven C J, Waltho J P, Clarke A R: **Domain behavior during the folding of a thermostable phosphoglycerate kinase.** *Biochemistry* 1996, **35:**15740-15752.
43. Kovári Z, Flachner B, Náray Szabó G, Vas M: **Crystallographic and thiol-reactivity studies on the complex of pig muscle phosphoglycerate kinase with ATP analogues: correlation between nucleotide binding mode and helix flexibility.** *Biochemistry* 2002, **41:**8796-8806.
44. Osváth S, Köhler G, Závodszky P, Fidy J: **Asymmetric effect of domain interactions on the kinetics of folding in yeast phosphoglycerate kinase.** *Protein Sci* 2005, **14:**1609-1616.
45. Szilágyi A N, Vas M: **Sequential domain refolding of pig muscle 3-phosphoglycerate kinase: kinetic analysis of reactivation.** *Fold Des* 1998, **3:**565-575.
46. Li R, Woodward C: **The hydrogen exchange core and protein folding.** *Protein Sci* 1999, **8:**1571-1590.
47. Itoh K, Sasai M: **Flexibly varying folding mechanism of a nearly symmetrical protein: B domain of protein A.** *Proc Natl Acad Sci USA* 2006, **103:**7298-7303.
48. Nishimura C, Prytulla S, Dyson H J, Wright P E: **Conservation of folding pathways in evolutionarily distant globin sequences.** *Nat Struct Biol* 2000, **7:**679-686.
49. Lazaridis T, Karplus M: **"New view" of protein folding reconciled with the old through multiple unfolding simulations.** *Science* 1997, **278:**1928-1931.
50. Sivaraman T, Kumar T K, Chang D K, Lin W Y, Yu C: **Events in the kinetic folding pathway of a small, all β-sheet protein.** *J Biol Chem* 1998, **273:**10181-10189.
51. Weissman J S, Kim P S: **A kinetic explanation for the rearrangement pathway of BPTI folding.** *Nat Struct Biol* 1995, **2:**1123-1130.
52. Zhang J X, Goldenberg D P: **Mutational analysis of the BPTI folding pathway: I. Effects of aromatic leucine substitutions on the distribution of folding intermediates.** *Protein Sci* 1997, **6:**1549-1562.
53. Kazmirski SL, Daggett V: **Simulations of the structural and dynamical properties of denatured proteins: the "molten coil" state of bovine pancreatic trypsin inhibitor.** *J Mol Biol* 1998, **277:**487-506.
54. Karplus M, Weaver D L: **Protein folding dynamics: the diffusion-collision model and experimental data.** *Protein Sci* 1994, **3:**650-668.
55. Abkevich V I, Gutin A M, Shakhnovich E I: **Specific nucleus as the transition state for protein folding: evidence from the lattice model.** *Biochemistry* 1994, **33:**10026-10036.

56. Fersht A R: **Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications.** *Proc Natl Acad Sci USA* 1995, **92:**10869-10873.
57. Viguera A R, Serrano L, Wilmanns M: **Different folding transition states may result in the same native structure.** *Nat Struct Biol* 1996, **3:**874-880.
58. Dobson C M: **Protein folding and misfolding.** *Nature* 2003, **426:**884-890.
59. Rao F, Caflisch A: **The protein folding network.** *J Mol Biol* 2004, **342:**299-306.