Report eFold

Dr. Jérôme Waldispühl and David Becerra

October 2, 2014

Contents

1 Abstract

2 Introduction

The protein folding problem entails advances in understanding the structural basis of protein interactions, as well as in the elucidation, characterization and annotation of protein function. These advances are supported by the understanding of protein post-translational modifications and folding intermediates, the identification of novel protein folds, and potential targets for drug design and treatments for many hereditary diseases [1, 2]. In contrast to how genes are studied, it is more challenging to study protein structure with high-throughput methods. Furthermore, post-translational modifications, regulatory mechanisms and environmental factors can result in proteins with multiple forms and structures [3]. Therefore, a detailed knowledge of protein 3D structures and structured folding pathways have facilitated the development of novel protein folding modelling methods [4].

The protein folding (PF) problem is interested in determining a protein tertiary structure from its amino acid sequence trying to understand the folding path which leads the folding process. Historically, the PF problem has been split in two related problems, the protein structure prediction problem (PSP) and the pathway prediction problem (PPP). The aim of PSP is at determining the configuration of a folded protein regardless of the folding process. On the other hand, the PPP is to determine the time ordered sequence of folding events (also know as the folding pathway). The PSP has widely acknowledged as an open problem and it has received more attention than the PPP problem. Furthermore, the importance of pathway prediction to get valuable insights into the folding process and to guide the search of the conformation space have been neglected. It is clear that the ability to predict folding pathways can greatly enhance structure prediction methods, however, most of the PPP methods starts from a known protein structure (i.e., 3D structure). The PPP problem is also interesting in itself given that protein misfolding and aggregation have been identified as the cause of several pathological conditions.

Functional proteins undergo natural selection processes preserving their function hence their structure. Simultaneously, they must also have good folding dynamic properties that enable them to fold quickly from an unfolded state to the native structure. A functional protein can be characterized by natural selection and/or folding properties. Then, the theory of evolution and the laws of physics are the principles on which the techniques of protein structure prediction are based. Comparative and fold recognition methods for protein structure prediction belong to the first characterization and they rely on the similarity between a target protein and a set of known protein structures at the fold level. By contrast, ab-initio methods focus on the second aspect and predict protein structure based on laws of physics, biology and chemistry without

considering any related structure as template.

A number of computational and experimental techniques for protein folding pathway prediction already exists in the literature, but most of them are limited by the required amount of time and resources, or the restrictive assumptions imposed during the modelling process. Despite its reliable predictions, molecular dynamics techniques have a extremely high computational cost and only predict one pathway. Some Monte Carlo simulations have been proposed rendering the simulations many orders of magnitude faster than molecular dynamics simulations, but simulations are still prohibitive expensive if custom-designed supercomputers are not used [5]. Probabilistic and Stochastic Roadmaps [6] are able to predict intermediate configurations on the folding pathway using a reasonable amount of computer resources, however the protein sampling process is highly hampered in these approaches due to the need of an a priori native conformation, the inefficiency due to the size of the configuration space and the lack of biological significance from the generated samples. One different approach to enumerate folding pathways is to start with a folded protein and unfold the protein in an ordered sequence of steps to its unfolded state [7, 8].

The protein folding problem is an NP-complete problem even in simple lattice models [9, 10] with tremendous running time requirements. Reliable predictions and critical features of protein foldings have been produced through custom-designed supercomputers and time-consuming molecular dynamics MD simulations [11, 12, 13, 14], however these computational approaches are hardly limited by the required amount of computational resources. State of the art methods are currently unable to compute, and even approximate, the complete 3D conformational landscape for all protein targets. Then, there is a tangible need in the structural biology research field to develop efficient and effective protein folding methods. The ensemble modelling [15] and evolutionary information content based methods [16] belong to a newer and promising group of approaches that aims to offer a better trade-off between efficiency and accuracy for predicting structures and folding pathways.

Many current obstacles presented in the protein folding problem have been already addressed by research in RNA. Specifically, the development of structural ensemble prediction algorithms have allowed the accurate computation of RNA secondary structure energy landscapes and sample structures from sequence information alone [17, 18]. Although, those approaches can not directly be mapped to proteins, they have been an excellent starting point to model more accurate and complex scenarios [19, 20, 21]. With respect to the protein structure scheme, we have already introduced a structural ensemble predictor for transmembrane β-barrel (TMB) proteins [15] continuing earlier work on molecular structure modelling [22, 23]. Recently, we introduced a method for modelling the folding process of large β-sheet proteins using sequence data alone [24].

Ensemble modelling methods employ a coarse-grained structural model that enables us to

efficiently compute the complete protein conformational landscape and apply statistical mechanics techniques. The prediction obtained by these methods describes the "ensemble" of protein conformational variants mimicking the ability of proteins to adopt different conformational states in vivo. Particularly, by using the Boltzmann partition function, the significance of all protein conformations based on strand residue interactions and their likelihood of occurrence can be estimated. The ensemble modelling has been proved to be accurate and novel to a variety of protein structural prediction problems. Specifically, structural ensemble predictors for transmembrane barrel TMB proteins [22, 15] and modelling the folding process of large single β-sheet proteins [24] have been proposed.

The prediction of 3D protein structures using evolutionary sequence information is a novel statistical approach in which evolutionary constraints are inferred from a set of sequences belonging to an iso-structural protein family [16]. These methods use the information gleaned from statistical analysis of multiple sequence alignments to reduce the space of 3D protein conformation where the 'native' structure can be identified. The first works in the area combined a few number of inferred residue contacts with protein structure information to predict the structure of small proteins [25, 26, 27]. The evolutionary sequence variation methods have been criticized for their little use in protein structure prediction due to their low accuracy. However, their usefulness debate has received new momentum with the rise of novel and accurate approaches, which could be based on homology modelling [28], or *de novo* modelling, i.e., do not use template-derived contacts or sequence-similar fragments from known structures [16, 29, 30].

The recent assessment of evolutionary sequence information prediction methods as accurate *de novo* models, allows their systematic application to 3D structure prediction studies. It follows that our ensemble modelling framework will highly benefit from the information deciphered from evolutionary records. In particular because the statistical potential energy functions used in our previous models contain a very weak signal. It can be hypothesized that the synergy between these models will improve the protein conformational sampling process, creating a balance between exploration and exploitation of the vast space of protein conformations, the primary obstacle of protein structure prediction.

In this work, we introduce `efold`, a new protein folding pathway prediction framework that combines ensemble modelling techniques with evolutionary sequence information methods. `efold` is a general framework that enables efficient simultaneous prediction of the protein folding mechanism and structure using only the primary sequence as input. Protein folding is modelled through the efficient enumeration of folding pathways using an ensemble methodology, where each folding step (Starting from an unfolded state) is represented by the addition of one topologically possible conformational with one less degree of freedom (i.e., an additional secondary structure).

`efold` represents a plausible advance in the PF state of the art because it makes feasible

the enumeration of folding pathways starting with an unfolded protein and consider the various possibilities for the protein to fold. Furthermore, `efold` studies protein folding as an ab-initio framework that models the dynamics of protein folding, instead of focussing solely on the native conformation. `efold` also expands our previous protein folding prediction frameworks in several directions while keeping its low CPU-time requirements. First `efold` models α-helices and multiple β-sheets. Next, `efold` algorithm applies memoization techniques and computes the conformational landscape of all β-sheet topology i.e. number of β-strand with their relative positions at once, hence avoiding redundant calculations and decreasing the computational complexity. Finally, to the best of our knowledge, for the first time the residue contact information is integrated in the Boltzmann sampling process performed by ensemble methods to predict protein pathways. The latter is important because statistical potentials have a limited accuracy and better scoring scheme are required to develop accurate folding pathways predictors. We found that the evolutionary sequence information stored in co-variation model has the potential to significantly increase the accuracy of our previous ensemble techniques.

3 Materials

3.1 Study Case Benchmark

This benchmark is composed by the protein G and Ubiquitin. These proteins have played a central role in protein folding studies being the system of choice in a vast body of experimental and theoretical studies. These small protein domains have represented ideal candidates for the elucidation of their folding pathways [31, 32].

The B1 domain of protein G, generally called GB1 or proteinG, has represented an ideal candidate for a vast number of different studies because of its small size and its simple and highly symmetrical topology. GB1 is a 56 amino acids length, regular a/b structure. The fold consists of a 4-stranded β-sheet and an α-helix tightly packed against the sheets [33]. Protein G folds through three pathways, all of which pass through an intermediate, to a single transition state (TS). The three intermediates feature a near-native helix along with hairpin 1 (I_1 intermediate), hairpin 2 (I_2), or the $\beta 1 - \beta 4$ sheet (I_3). The work [34, 35] reported an early formation of the second hairpin ($\beta 3 - \ turn - \beta 4$) and its fundamental role in the folding process. Additionally, this second hairpin centers around known nucleation points W43, Y50, F54 that are strongly stabilized by three hydrophobic residues W43, Y45 , F52 [32]. Namely, three folding pathways are observed, each involving formation of its own assembly: helix-first hairpin, helix-second hairpin, and $\beta 1 - \beta 4$ sheet. All pathways appeared to converge to the same folding nucleus.

Ubiquitin is a small protein (76 residues in length) that has a highly structured native

state which is very stable. Its high stability may be linked with the function of ubiquitin, which becomes covalently attached to lysine side chains in proteins thereby targeting them for degradation by the proteasome. It is likely that there is some residual structure in the denatured state of ubiquitin in the region of the first β-hairpin and the α-helix. The folding of ubiquitin is two-state under most conditions, however, an intermediate can be stabilized and become populated during folding using a number of methods, for example, by the use of a stabilizing salt such as sodium sulfate [36].

3.2 EvFold Benchmark

The original evFold benchmark is composed by 15 protein structures ranging from 48 to 258 amino acids in size. The evFold benchmark proteins were selected based on the following criteria: (i) Proteins that belong to a protein family composed by more then 1000 sequences per protein family; (ii) Proteins that include all of the main protein fold families, such as all$-\alpha$, α/β, $\alpha+\beta$ and all-β; (iii) Proteins with availability of experimentally derived (PDB) structures for at least one family member. Each PFAM family was assumed to be iso-structural, so that all protein structures in a family form a tight and distinct cluster in protein structure space. The essential components of the evFold method for the prediction of a 3D protein structure using evolutionary sequence information without the use of structural templates are [16]:

1. Protein sequence alignment for the protein family containing the target protein.
2. Formulation of a global statistical model for sequences in a protein family.
3. Derivation of parameters that maximize entropy in this model, using direct coupling analysis (DCA).
4. Derivation of a ranked set of evolutionarily inferred contacts!(EICs).
5. Secondary structure prediction using well established methods.
6. Implementation of weighted distance restraints from inferred contacts.
7. Application of distance geometry and constrained molecular dynamics.
8. Automated ranking of predicted structures to nominate a single predicted structure and a set of lower ranked alternatives.

A subset of 6 proteins are selected out of the 15 protein structures. The criteria to filter the set of proteins are to build a benchmark with proteins shorter than 250 aminoacid length; proteins belonging to the folding groups α/β, $\alpha+\beta$ and all-β; and proteins with less than six strands. Regarding the components of the evFold method, the 500 best ranked evolutionary

inferred contacts are selected as input for our algorithm. This mean that out of the 8 components of evFold, our method only runs the first four. Then, only the statistical analysis of co-variation in the protein sequences is used to infer residue-residue proximity within an isostructural protein family. In other words, `efold` does not make use of the secondary structure prediction and distance geometry and simulated annealing calculations performed by EvFold.

3.3 916 Benchmark

The BetaSheet916 dataset was extracted from the Protein Data Bank of May 2004 by Cheng [37]. This benchmark contains 916 chains (corresponding to 187516 residues) determined by X-ray diffraction having resolution better than 2.5Å. All the protein chains contain standard amino acids with a length greater than 50 amino acids. The redundancy in the dataset is guarantee to have a sequence identity of $15 - 20\%$. 48 996 are β-residues participating in 31638 interstrand residue pairs. The dataset has 10745β-strands with an average length of 4.6 residues and 8172 β-strand pairs, including 4519 antiparallel pairs, 2214 parallel pairs and 1439 pairs involving isolated β-bridges. These strand pairs form 2533 β-sheets. The average sequence separation between residue pairs and strand pairs is 43 and 40, respectively.

Regarding the proposed experimental framework, 125 proteins were selected out of the 916 data set. Specifically, only proteins that contains less than six strands were selected. The BetaSheet916 set is routinely adopted as benchmark set for β-sheet prediction methods. `efold` is not a method designed for secondary structure predictions alone, however, the BetaSheet916 represents a considerable corpus of proteins with low identity to validate the accuracy of `efold` through a big folding space.

3.4 Pfam Benchmark

112 Pfam families are identified when the complete set benchmark (i.e., Study Case plus EvFold plus 916 Benchmarks) is clustered. Three families (out of 112) are retained given that they contains 4 or more proteins and that there is experimental information about their folding pathways. These three families are studied in order to determine the existence of common folding intermediates between the members of a same Pfam family. The conservation of folding intermediates in evolutionary related proteins can unveil, throughout the identification of key regions, motifs and residue contacts, general kinetic and thermodynamical principles that govern protein folding.

3.4.1 PF0018 - SH3 Domain

Due to its small size and multiple homologues, SH3 has been widely studied to address various important aspects of protein folding, such as the the synergistic relationship between experi-

ments and simulations, the nature of protein folding transition states the relationship between protein topology and the folding pathway [38]. SH3 is composed of two orthogonally packed stranded β-sheets that form a single hydrophobic core [39]. The first sheet consists of the three central strands of the protein ($\beta 2 - \beta 3 - \beta 4$) and the second sheet of the two terminal strands ($\beta 1 - \beta 5$) and a portion of the RT loop. There is also a small 3_{10} helix between $\beta 4$ and $\beta 5$ [40]. It has been shown that the structure in the transition state ensemble is highly polarized with the hydrogen bonding network associated with two β-turns. The denaturation of the N and C termini, turns and loops, and a small amount of secondary structures located in the central $\beta 2 - \beta 3 - \beta 4$ are general features of the SH3 transition state ensembles (TSE) [39] . Particularly, the distal β-hairpin and the diverging turn are formed in the transition state and that all conformations in the TSE have the $\beta 2 - \beta 3 - \beta 4$ formed [41]. Experimental results have also shown that $\beta 2$, $\beta 3$, and to a lesser extent the $\beta 4$ strands are the most ordered regions of the TSE.

Protein engineering studies suggested that the folding pathways of SH3 domain may be evolutionary conserved and that its topology may play an important role in determining the folding pathway of this structure. Furthermore, L24 has been shown experimentally to be involved in the TSE and to be a highly conserved position in the SH3 fold family [39, 42].

3.4.2 Kunitz Domain

Kunitz domains are relatively small with a length of about 50 to 60 amino acids. Examples of Kunitz-type protease inhibitors are aprotinin (bovine pancreatic trypsin inhibitor, BPTI), Alzheimer's amyloid precursor protein (APP), and tissue factor pathway inhibitor (TFPI). From them, BPTI is one of the most extensively studied globular proteins and was the first case of well-documented disulfide folding pathway. Furthermore, its protein folding pathway and dynamics have been investigated in great detail. BPTI is a Kunitz-type protease inhibitor which comprises 58 amino acids and three disulfides-bonds in its native form. Its structure is a disulfide rich $\alpha + \beta$ fold. Disulfide-bonds occur between cysteine residues 5-55, 30-51 and 14-38.

The BPTI folding pathway is primarily a five state system including the unfolded and native forms. In the second state, the formation of the native disulfide 30-51 predominates. In the third state, non-native disulfides 5-14 and 5-38 rapidly interconvert between each other and the native 14-38, with 30-51 remaining stable. In the fourth state, BPTI must pass through the intermediate containing the native disulfides 30-51 and 5-55 [43]. NMR exchange data indicate the formation of a fully folded sheet with subsequent helix formation during the folding process. Then, the pathways seems to involve the full association of the 3-stranded sheet ($\beta 1$, $\beta 2$ and $\beta 3$), followed by the C terminal helix ($\alpha 2$), the N terminal helix ($\alpha 1$). The initial formation of the 30-51 disulfide can be in agreement with the early formation of the $\beta 1 \beta 2 \beta 3$ sheet and its association with $\alpha 2$. With the joining of $\alpha 1$ with the framework, the disulphide $5 - 55$ can be

formed. Finally, the disulphide $14-38$ will formed with the loops coming into place [43]. Then, secondary structures form early during the folding, which is followed by docking and packing of preformed secondary structural units to form the native tertiary structure [44].

3.4.3 LSm PF01423

Two sequence motifs (named Sm1 and Sm2) have been identified through the comparison of various LSm homologs. The size of the Sm1 and Sm2 motifs are 32 and 14 amino acids long, respectively. The Sm1 sequence motif corresponds to the $\beta 1$, $\beta 2$, $\beta 3$ strands, and the Sm2 sequence motif corresponds to the $\beta 4$ and $\beta 5$ strands. The sequence motifs are conserved and they are separated by a non-conserved region of variable length. This fact suggest that all LSm protein genes evolved from a single ancestral gene.

4 Methods

The free energy global optimization of a potential energy function is the classical physical approach for the prediction of protein structures in *the novo* approach. However, structures predicted from those algorithms may not represent the true structure, or even a suboptimal folding [45]. The free energy based algorithms are highly hampered by i-) the inaccuracy of the potential energy functions devised to represent the protein energy landscape, and ii-) the unfeasibility of adequately sampling the conformational landscape. Thus, many works have introduced variants to improve the methods for global optimization, the constraints in protein conformational searches and distributed computing technologies [46]. Additionally, some methods are not longer performing a search for an individual, lowest energy structure, but they aim the prediction of an ensemble of protein conformations and pathways. New approaches aim to make a better use of protein folding kinetics properties to improve their accuracy; where the idea of a single folding pathway is replaced by an energy landscape and a folding funnel model.

Many current obstacles presented in the protein structure prediction problem (such as the aforementioned) have been already addressed by research in RNA. Specifically, the development of structural ensemble prediction algorithms have allowed the accurate computation of RNA secondary structure energy landscapes and sample structures from sequence information alone [17, 18]. Although, those approaches can not directly be mapped to proteins, they have been an excellent starting point to model more accurate and complex scenarios [19, 20, 21]. With respect to the protein structure scheme, we have already introduced a structural ensemble predictor for transmembrane β-barrel (TMB) proteins [15] continuing earlier work on molecular structure modelling [22, 23]. Recently, we introduced a method for modelling the folding process of large β-sheet proteins using sequence data alone [24].

In this work, we expand the scope of our previous ensembles prediction techniques and improve their performance (i.e. speed and accuracy). Specifically, the proposed method is novel because: *i*) It allows the pure β, pure α and α/β interactions. *ii*) It uses a divide-and-conquer approach enhanced with memoization techniques to allow the efficient computation of the Boltzmann partition function over the set of all possible protein states. Additionally, the chosen data structure allows the modelling of a meaningful hierarchical assembly folding mechanism to simulate population folding dynamics. This assembly of protein topologies is based on the energy favorability of the protein schemas, instead of using a hard coded as in our previous implementations *iii*−) In order to circumvent the limitation of the scoring scheme of our previous techniques, this work exploit the evolutionary record in protein families adding information about evolutionary constrained interactions in the protein folding process. We will infer residue pair couplings and we will compute an enhanced statistical mechanical energy framework in the modelling of folding pathways transitions and population dynamics.

The proposed approach predicts protein structures and protein pathways in a single run. Then, it can be naturally divided in two main tasks:

1. **Modelling the ensembles:** The main goal of this task is to compute a set of protein states with the highest occurrence likelihood. Our approach is based in two steps:

   (a) **The forward step** of the algorithm computes the equilibrium partition function of all possible secondary structures: Using a divide-and-conquer approach and memoization techniques, we compute the Boltzmann partition function over the set of all possible protein states, where the protein states has been modelled through a coarse grain representation based on secondary structures. Particularly, each protein is presumed to fold into a complete set of unique structural states, with a single energetic value assigned according to a Boltzmann distribution and evolutionary contact prediction scores. Then, clusters of low-energy states with similar conformations are extracted using their relative energetics.

   (b) **The backward step** computes the probabilities of a set of statistically representative samples: We analyze the significance of the protein states generated in the forward step computing its associated occurrence likelihood.

2. **Modelling the Folding Dynamics:** The main goal of this task is to derive the likelihood of dynamic state-to-state transitions, and assemble a set of complete folding paths. The transition from a random coil to the native state was modelled as a path in a graph of varyingly folded protein conformation states. The dynamics of the system is calculated by treating the folding process as a continuous time discrete state Markov process.

A schematic pipeline and the flowchart of the proposed method can be seen in Figure 1. The specific details of the methodology are shown in the hereinafter subsections.

eFold

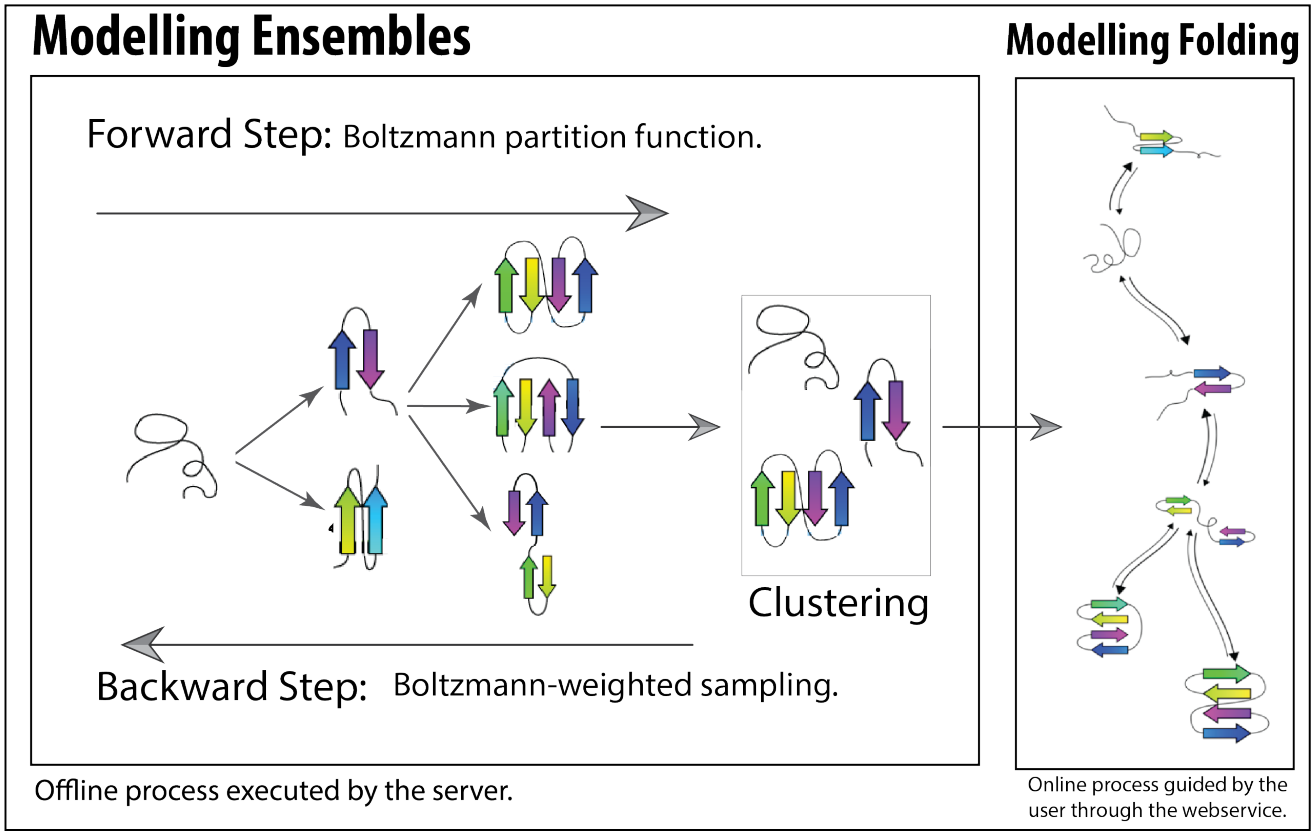

Figure 1: **efold:** is the proposed algorithm for predicting protein folding pathways and topologies using ensemble modelling and genomic variation. The algorithm is divided in two main phases, the modelling of ensembles and the modelling of the predicted folding dynamics. The first phase is computed off-line and it consist of a forward and backward traversal over the tree that model the hierarchical folding mechanism and that stores all the possible proteins states with its respective energies and likelihoods of occurrence. The second phase simulates the protein population dynamics based on the clusters computed in the previous phase. Specifically, the transition from a random coil to the native state was modelled thorough a hierarchical assembly folding mechanism and it is represented as a path in a graph of varyingly folded protein conformation states. In this graph, the vertices are represented by energetically accessible conformation states presented in the clusters. The edges in the graph represent the possible folding pathways and the existence of structural similarity between the connected vertices. The user can change the structural similarity cutoff in order to generate different predicted protein pathways.

4.1 Modelling the ensembles

4.1.1 The forward step.

The main task of the forward step in the modelling of ensembles, is to compute the partition function of secondary structures with arbitrary β-strand topologies. In order to accomplish this goal, a statistical mechanics framework to compute the set of all possible secondary structure conformations that a protein can attain was defined. This framework is characterized by the implementation of a protein representation, the generation of all the admissible β-sheet topologies following the proposed protein representation and the computation of the Boltzmann partition function over those topologies.

Computation of the Partition Function:

Conceptually, each protein structure was described by a coarse grain residue-level representation. Specifically, the structure was defined by the set of residue/residue contacts that form hydrogen bonds between β-strand backbones. The protein representation includes side-chain orientation and long-range contacts, that will enable us to develop an efficient strategy to enumerate all potential states. This representation sufficiently reduces the complexity of the conformational search, although, the number of protein conformations are still greatly flexible (E.g. permutation of strands, strand's size, orientation of side chains, secondary structure motifs, etc.), and the structures can take on various conformations that are vastly different between them, and the native conformation.

The protein generic topologies were encoded using a stepwise permutation algorithm through the labeled set of β-strands $\{1 \dots n\}$. For each permutation, the set of all β-strand/β-strand pairings were computed, such that each interaction in the β-topology is assigned to be parallel (P), anti-parallel (A) or none (N) (See Figure 2). It is important to stress that in order to avoid unrealistic general protein shapes, optimize computation resources and focus in valid motifs, we imposed that valid foldings must satisfy steric and biologically derived constraints. More specifically, we set a minimum and maximum strand length and minimum inter-strand loop size for the protein conformations.

Contrary to our previous implementations, the computation of all protein topologies is performed using a tree data structure, where each level of the tree contains all the topologies with a specific number of strands. Then, the first level (i.e., the root) of the tree correspond to the topologies containing the unfolded scheme, the second level of the tree contains the topologies with two strands, and so on until the leaves of the tree (i.e., n-level) are stored with topologies having n strands. The tree is a balanced tree, for which each node (except the root) has one node parent, $m-1$ sibling nodes and m children nodes. All the parent nodes share a structure with their children, where two topologies share their structures if they are identical

to each other, modulo the addition or removal of a single strand pairing.

Having a tree as a data structure is important for four main reasons: *i*) It guarantees the algorithm correctness, given that all the possible offsprings are traversed. Additionally, It ensures an exhaustive and non-overlapping count of all protein structures and it support a hierarchical assembly folding mechanism to narrow the conformation search (See the section Folding Dynamics for details) *ii*) A Boltzmann sampling procedure can be efficiently computed using a depth-first search approach (DFS). Furthermore, the tree should not be completely filled in order to perform the procedure (see Sampling subsection). *iii*) Pruning methods can be computed over many branches of the tree previously computed. The pruning of the three will keep the memory complexity in tractable terms, furthermore it will avoid the degradation of their performance (avoiding collisions and crossing the hash load factor). *iv*) The tree data structure can be traversed in different fashions allowing the analysis of a highly diverse set of experiments.

The tree structure is filled using a breadth-first approach (BFS). In other words, the level $i + 1$ would not be considered until all the instances of level i have been computed. The filling of the tree consists in the computation of the Boltzmann partition function Z for all the nodes of the tree (i.e., all admissible β-sheet schemas). Conceptually, each structure with a specific topology is described by the set of residue/residue contacts that form hydrogen bonds between β-strand backbones. Then, we compute for each conformation a pseudo-energy which is determined by the specific residues involved in contacts. The residue/residue contact energy is computed through a potential-energy scoring function derived from frequency observations of specific residue/residue interactions in experimental data [22]. Particularly, an energy $E_{i,j}$ is given to each residue/residue pair following Equation 1, where Z_c is a statistical re-centering constant and $p(i, j)$ is the likelihood of these two residues appearing in a β-sheet environment, as observed across all nonsequence-homologous solved structures in the PDB.

$$E_{i,j} = -RT[log(p(i, j)) - Z_c] \tag{1}$$

A predicted energy is then related to the sum of potentials for all residue/residue interactions (see Equation 2), where i, j represent the positions of the amino-acids being computed that belongs to all the possible residue pairs γ. Further, we assign separate likelihoods based on the hydrophobicity of the environment on either face of a β-sheet.

$$E(S_n) = \sum_{i,j \in \gamma} E_{i,j} \tag{2}$$

The Boltzmann partition function Z can be calculated over all protein structural states to characterize the energetic landscape of a specific ensemble (see Equation 3), where $E(S_i)$ is the free energy of the structure for the input sequence, R is the gas constant and T is the absolute

temperature.

$$Z = \sum_{i=1}^{n} \exp[-E(S_i)/RT] \tag{3}$$

With the partition function Z available, the Boltzmann probability for all the structures can then be computed using Equation 4. Therefore, the Boltzmann probability statistically characterizes the ensemble.

$$P(S_i) = \frac{\exp[-E(S_i)/RT]}{Z} \tag{4}$$

The enumeration of all possible structures is infeasible during the computation of the partition function. We have previously shown that a dynamic programming approach is an efficient method to compute arbitrary single $\beta-$sheet fold topologies. In this work, we propose a much more efficient method using a tree data structure and memoization techniques.

$$E(S_n) = E(S_{n-1}) + Pair(s_{n-1}, s_n) \tag{5}$$

Equation 5 represent the recursion to compute the energy of a structure with n strands, where $E(S_{n-1})$ is the interaction energy between the first $n-1$ strands, and $Pair(s_{n-1}, s_n)$ is the energy of the pairing of strand $n-1$ with strand n (See Figure 2a). The implemented recursion function exploits the shared sub-structures between schemes in the ensemble using a memoization approach. Each recursive call compute the energy function of a specific instance and store this value in a hash table indexed by an identifier. Subsequent recursive calls, which involves the same instance, will perform a search in the tree and a table lookup instead of re-computing the value of the recursion.

A hash table maps *keys* to *values*. In our implementation, *keys* are lists of four indices i_i, i_2, i_3, i_4. These indices partition the protein structures based on the boundaries of region occupied by the strands (See Figure 2b). The *values* correspond to an array that contains information about the templates, the best computed Boltzmann partition function Z and a value representing the relative abundance (likelihood) of the structure. These likelihoods are finally weighted using an evolutionary contact prediction method in order to circumvent the inherent limitation of potential energy scoring schemes.

The used evolutionary contact prediction method [16] is based on a maximum entropy approach to perform an unsupervised inference of residue-residue contacts from multiple sequence alignments (MSAs). Specifically, the method derives a set of essential residue pair couplings through a maximum entropy approach and a direct coupling analysis. The minimal set of pairs predicted to co-vary due to evolutionary constraints is returned as output of the algorithm and it is connected as an heuristic to our ensemble approach.

In our ensemble pipeline, the set of predicted couplings are ranked by their numerical values and they are codified in an $N \times N$ binary matrix C, also know as a predicted contact map, whose element $C(i, j) = 1$ if the predicted direct information of residues i and j is greater

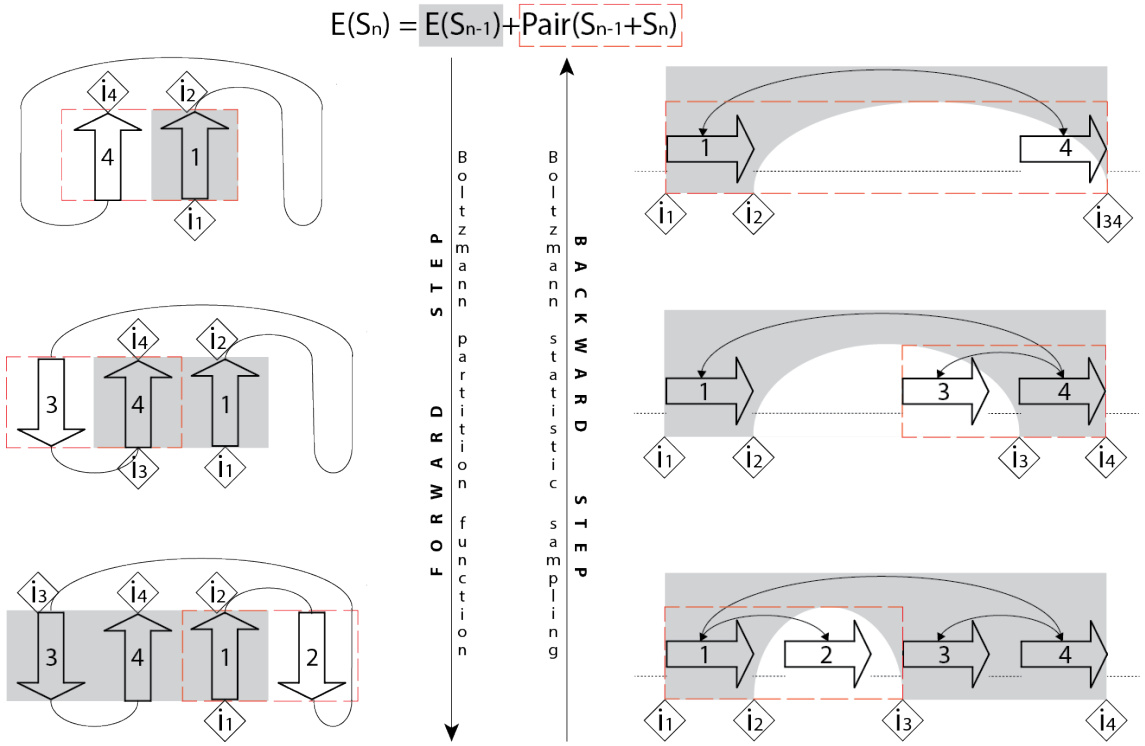

Figure 2: `efold` **Topologies:** *3A4P1A2* is the protein topology for the proteinG

than a threshold value t. In our approach, t was chosen as the direct information of the 500 hundred best ranked prediction. This parameter was determined as a good threshold to predict 3D structures with correct spatial arrangement of α helices and β-strands for our benchmark proteins, as compared to their experimentally determined structures.

The predicted contact map C is used to numerically compute residue pairs involved in secondary structure motifs. Particularly, those motifs can be recognized in the matrix C identifying a cluster of contacts using geometric knowledge of α-helices and β-strands. Then, we can add α-helices template information to our permutable β-template procedure to enable the modelling of pure β, pure α and α/β interactions. Now, the different sampled structures can be penalized or rewarded depending on the modelled motif. The last procedure builds a selective constraint which can intensify the signal of β-strand interactions during the modelling of pathway kinetic.

4.1.2 The backward step.

A characterization of the full ensemble of protein structures using the complete enumeration of secondaries structures is restrictive. Then, during the backward step, we compute a statistically representative sample of secondary structures. Additionally, clusters of these secondary struc-

tures are built based on their topological and structural similarities to work with a tractably sized system. This system is used as input for the prediction of folding dynamics.

During the sampling process, a statistical sampling over the protein conformations generated in the forward step is performed. Particularly, a recursive statistical algorithm to sample from the Boltzmann ensembles of secondary structures using the tables constructed to compute the partition function is used. We take advantage of the tree structure and the memoization tables to randomly draw secondary structures according to the probabilities given by equation 4.

Since the final structure of the protein is not known, the proposed approach samples configurations from all possible $\beta-$sheet topologies (i.e., all the nodes of the tree). Then, for each node, the sampling algorithm performs a recursive traceback through the partition function tables of its parents. For a specific node, the location of a single strand is sampled from the region indicated by the indices i_2, i_3 (See figure 2 for an example).

4.2 Predicting Folding Dynamics

In order to simulate population dynamics, we use ensemble predictions and a hierarchical assembly folding mechanism to narrow the conformation search. In this process, the secondary structure is formed according to the primary structure of the protein. Specifically, the first step in the process is represented by the unfolded state, next the secondary structures are formed and they fluctuate around their equilibrium positions. Finally, the secondary structures interact between them and they create a folding pattern that will find the native conformation. The proposed approach try to separate conformational transitions that are critical to folding from those that could simply result from minor structural fluctuations.

Our approach predicts coarse folding transitions as described in previous models [47]. Specifically, the transition from a random coil to the native state was modelled as a path in a graph of varyingly folded protein conformation states. In this graph, the vertices are represented by energetically accessible conformation states which have been previously generated by the proposed Boltzmann ensemble sampling method (See subsection Sampling Process). The edges in the graph represent the possible folding pathways and the existence of structural similarity between the connected vertices. Specifically, for every pair of states we add an transition edge if (1) the states have compatible topologies, and further, (2) the states show structural similarity. Two states are compatible if they are identical to each other, modulo the addition or removal of a single strand pairing. On the other hand, the structural similarity between two samples is estimated through a contact based metric, where two structures are structurally similar if the contact-based metric is below a transition threshold.

Given that two states are connected in the graph, the rate at which they interconvert is proportional to the difference between free energies of the states (ΔG). Since we sample thousands of states from each strand topology and in order to work with a tractably sized system, we

partition the state space into macro states using clustering. We cluster protein configurations according to contact distance metrics, and associate each cluster with a intermediate folding state. Under this approximation, we consider two clusters to be connected if the minimum distance between any two states from each is less than a threshold value. We define the ensemble free energy difference ΔG_{ij} between two macro states i and j by summing over the states from which they are composed (See Equation 6).

$$\Delta G_{ij} = E(\chi_i) - E(\chi_j) = \sum_{x \in \chi_i} E(x) - \sum_{x \in \chi_j} E(x) \tag{6}$$

Given the previous graph, the transition rates r_{ij} between states i and j is calculated using the Kawasaki rule (with parameter t_0 to scale the time dimension (See Equation 7)). Then, the change in the probability of the system being in state i at time t can be calculated from the total flux into and out of state i (see Equation 8, where p_i is the probability of state i, X is the state space).

$$r_{ij} = r_0 \exp(-\Delta G_{ij}/2RT) \tag{7}$$

$$\frac{dp_i}{dt} = \sum_{j \in X} r_{ij} p_j(t) \tag{8}$$

Finally, the dynamics of the system are calculated by treating the folding process as a continuous time discrete state Markov process. Given the matrix of folding rates R, where $R_{ij} = r_{ij}$ and initial state density $p(0)$, the distribution overs states $p(t)$ of the system at time t is given by the explicit solution to the system reported by Equation 9. Then, the distribution of conformations over folding time is estimated by solving this system.

$$p(t) = \exp(Rt)p(0) \tag{9}$$

5 Experimental Framework

In order to understand the performance of the proposed method, two main experiments were performed to study the modelling ensemble (protein structure prediction) and modelling folding phases (protein pathway prediction). The first phase is performed off line and it contributes most of the complexity of the algorithm. The second part is computed online and it runs using a web-service as interface with the user. `efold` is run having as input only the amino-acid sequence and a set of parameters. Then, the algorithm runs until all the folding pathways and structures have been computed. Based on the experimental framework and user experiences with our previous techniques, we have fixed the limits of our algorithm (constrained in the web

service interface) to predict the folding pathways for small proteins (less than 200 amino acids), and to model proteins with up to 7 different β-sheet strands.

The first experiment quantify the ability of `efold` to predict protein topologies through the prediction of residue-residue contacts, secondary structures and topologies. For the residue-residue contact prediction, we sample 150 configurations for each protein of the benchmark, and use these ensembles to compute a stochastic contact map. The contact map represents the probability of observing a given contact. Contacts are defined as all C_α atoms less than $8\mathring{A}$ apart in the PDB file. The precision (i.e., $\frac{no.\ of\ correctly\ predicted\ contacts}{no.\ of\ predicted\ contacts}$), sensitivity (i.e., $\frac{no.\ of\ correctly\ predicted\ contacts}{no.\ of\ observed\ contacts}$) and F-measure i.e., $\frac{2\times precision\times sensitivity}{precision+sensitivity}$ were the chosen measures to evaluate the quality of the predicted contacts. These metrics are evaluated when predicted contacts are within exact or ± 2 residues of an observed contact, and are more than 0, 12 and 24 residues apart. The secondary structure predictions are studied following the same methodology than residue-residue contacts. A pair of residues are considered to be part of β-strand interactions if the predicted residues are contacts and those contacts are observed to be involved in a β-sheet interaction in its corresponding PDB file. Finally, the topology prediction is studied based on the ranking of the Boltzmann probabilities of each ensemble of secondary structures. Particularly, the position of the topology reported by the PDB file with respect a sorting of the Boltzmann probabilities of all the secondary structures is computed.

To study the efficacy of our technique for predicting protein folding pathways, we studied the folding landscape of proteins for which their pathways have been elucidated through many experimental studies and/or MD simulations (see Material section for more details). We studied the ability of `efold` to separate conformational transitions that are critical to folding from those transitions that could simply result from minor structural fluctuations.

The graph of the folding pathway was constructed by considering all pairs of clusters computed during the modelling ensemble phase. If the minimum distance between two clusters was less than the transition threshold, we considered that there was exchange between the two states.

This ability is studied on the different

folding intermediates nuclei conserved residues

Boltzmann probability that statistically characterized the ensemble the complete set of experiments.

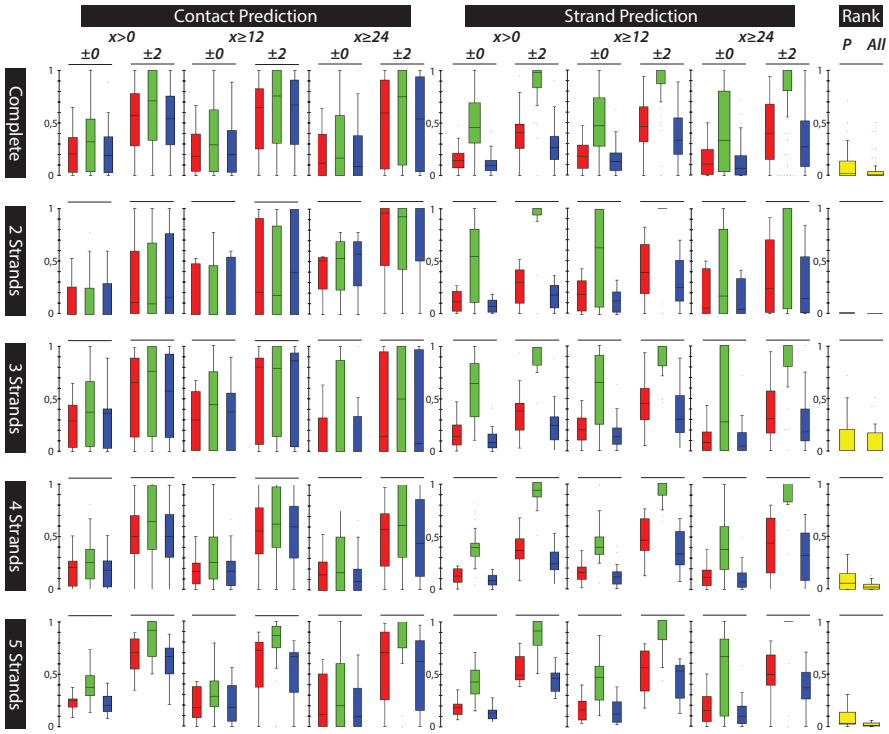

Figure 3

6 Results

6.1 Study Case Benchmark

6.1.1 Protein G

- Check the probability of all the intermediates.
  - During the forward step, all the intermediate steps are considered as a first folded event in at least one of the runs. Specifically, the hairpin 1 was selected 56% of the times as a first folding event, hairpin 2 was selected 4% of the times as the first folding event, B1 B4 sheet was select 40% of the times as the first folding event.
  - 100% of the runs contain the topology 1A2 (Needed for the Intermediate hairpin 1 and hairpin 2). 100% of the runs contain the topology 1P2 (Needed for the intermediate B1 B4). 100% of the runs contain the topology 4P1A2 (Needed for the interaction of the intermediates hairpin1 + B1 B4 Sheet). 78% of the runs contains the topology 3A4P1 (needed for the interaction of the intermediates hairpin2 + B1-B4 Sheet) 73% of the runs contains the topology 3A4N1A2 (needed for the interaction of the intermediates hairpin1 + hairpin2). 58.2% of the runs contain the

topologies 1A2 & 1P2 & 4P1A2 & 3A4P1 & 3A4N1A2 & 3A4P1A2 (All possible pathways).

- 65.45% of the runs contain the intermediate pathway 1P4 ⇒ 4P1A2 (interaction of the intermediates B1 B4 Sheet + hairpin1). 74.54% of the runs contain the intermediate pathway 1P4 ⇒ 3A4P1 (needed for the interaction of the intermediates B1-B4 Sheet + hairpin2). 98.18% of the runs contain the intermediate pathway 1A2 ⇒ 4P1A2 (interaction of the intermediates hairpin1 + B1 B4 Sheet). 70.9% of the runs contain the intermediate pathway 1A2 + 3A4 ⇒ 3A4N1A2 (needed for the interaction of the intermediates harpin1 + hairpin2). 80% of the runs contain the intermediate pathway 4P1A2 ⇒ 3A4P1A2 (interaction of the intermediate intermediates hairpin1 + B1 B4 Sheet + native state). 20% of the runs contain the intermediate pathway 3A4P1⇒ 3A4P1A2 (interaction of the intermediate intermediates hairpin2 + B1 B4 Sheet + native state). 63.63% of the runs contain the intermediate pathway 3A4N1A2 ⇒ 3A4P1A2 (interaction of the intermediates harpin1 + hairpin2 + native state).
- 65.45% of the runs contain the intermediate pathway 1P4 ⇒ 4P1A2 (interaction of the intermediates B1 B4 Sheet + hairpin1). 74.54% of the runs contain the intermediate pathway 1P4 ⇒ 3A4P1 (needed for the interaction of the intermediates B1-B4 Sheet + hairpin2). 98.18% of the runs contain the intermediate pathway 1A2 ⇒ 4P1A2 (interaction of the intermediates hairpin1 + B1 B4 Sheet). 70.9% of the runs contain the intermediate pathway 1A2 + 3A4 ⇒ 3A4N1A2 (needed for the interaction of the intermediates harpin1 + hairpin2). 80% of the runs contain the intermediate pathway 4P1A2 ⇒ 3A4P1A2 (interaction of the intermediate intermediates hairpin1 + B1 B4 Sheet + native state). 20% of the runs contain the intermediate pathway 3A4P1⇒ 3A4P1A2 (interaction of the intermediate intermediates hairpin2 + B1 B4 Sheet + native state). 63.63% of the runs contain the intermediate pathway 3A4N1A2 ⇒ 3A4P1A2 (interaction of the intermediates harpin1 + hairpin2 + native state).
- 63.63% of the runs contains the pathway 1A2 + 3A4⇒3A4N1A2⇒3A4P1A2. 80% of the runs contains the pathway 1A2 ⇒3P1A2⇒3A4P1A2. 52.72% of the runs contains the pathway 1P4 ⇒4P1A2⇒3A4P1A2. 3.63% of the runs contains the pathway 3A4 ⇒3A4P1⇒3A4P1A2. 16.36% of the runs contains the pathway 1P4 ⇒ 3A4P1⇒ 3A4P1A2.
- 5.45% of the runs contain all the possible pathways in the same graph.

- Check the nucleus residues.

- The nuclei residues identify experimentally are Y3,L5 (Hairpin1), F30 (Helix), W43,Y45,F52 (Hairpin2). The runs contains a total of 4220 structures, for which the nuclei residues are present in the following proportions Y3(22.46%), L5(88.67%), W43(9.36%), Y45(13.29%) and F52(85.1%).

- Compare the nucleus residues with ubiquitin family.

  - Y3, L5, F30, W43, and F52 show low sequence entropy over aligned sequences in the ubiquitin superfamily. The aligned sequences in the ubiquitin superfamily (with respect to the protein G) that show a frequency greater than 0 are: L5:K6(99.52%), F30:V26(82.54%), W43:L43(85.37%), and Y45:F45(42.45%). The sequence and structural alignment was taken from (Similarity of Protein G and Ubiquitin). It is important to stress that all the matches belongs to the secondary structures and that the atoms in the matched residues can be superimposed. The protein reference was 1UBQ.

6.1.2 Protein G Mutant A

- Check the probability of all the intermediates.

  - During the forward step, all the intermediate steps are considered as a first folded event in at least one of the runs. Specifically, the hairpin 1 was selected 56% of the times as a first folding event, hairpin 2 was selected 4% of the times as the first folding event, B1 B4 sheet was select 40% of the times as the first folding event.

  - 100% of the runs contain the topology 1A2 (Needed for the Intermediate hairpin 1 and hairpin 2). 100% of the runs contain the topology 1P2 (Needed for the intermediate B1 B4). 100% of the runs contain the topology 4P1A2 (Needed for the interaction of the intermediates hairpin1 + B1 B4 Sheet). 78% of the runs contains the topology 3A4P1 (needed for the interaction of the intermediates hairpin2 + B1-B4 Sheet) 73% of the runs contains the topology 3A4N1A2 (needed for the interaction of the intermediates hairpin1 + hairpin2). 58.2% of the runs contain the topologies 1A2 & 1P2 & 4P1A2 & 3A4P1 & 3A4N1A2 & 3A4P1A2 (All possible pathways).

  - 65.45% of the runs contain the intermediate pathway 1P4 $\Rightarrow$ 4P1A2 (interaction of the intermediates B1 B4 Sheet + hairpin1). 74.54% of the runs contain the intermediate pathway 1P4 $\Rightarrow$ 3A4P1 (needed for the interaction of the intermediates B1-B4 Sheet + hairpin2). 98.18% of the runs contain the intermediate pathway 1A2 $\Rightarrow$ 4P1A2 (interaction of the intermediates hairpin1 + B1 B4 Sheet). 70.9% of the runs contain the intermediate pathway 1A2 + 3A4 $\Rightarrow$ 3A4N1A2 (needed

for the interaction of the intermediates harpin1 + hairpin2). 80% of the runs contain the intermediate pathway 4P1A2 ⇒ 3A4P1A2 (interaction of the intermediate intermediates hairpin1 + B1 B4 Sheet + native state). 20% of the runs contain the intermediate pathway 3A4P1⇒ 3A4P1A2 (interaction of the intermediate intermediates hairpin2 + B1 B4 Sheet + native state). 63.63% of the runs contain the intermediate pathway 3A4N1A2 ⇒ 3A4P1A2 (interaction of the intermediates harpin1 + hairpin2 + native state).

- 65.45% of the runs contain the intermediate pathway 1P4 ⇒ 4P1A2 (interaction of the intermediates B1 B4 Sheet + hairpin1). 74.54% of the runs contain the intermediate pathway 1P4 ⇒ 3A4P1 (needed for the interaction of the intermediates B1-B4 Sheet + hairpin2). 98.18% of the runs contain the intermediate pathway 1A2 ⇒ 4P1A2 (interaction of the intermediates hairpin1 + B1 B4 Sheet). 70.9% of the runs contain the intermediate pathway 1A2 + 3A4 ⇒ 3A4N1A2 (needed for the interaction of the intermediates harpin1 + hairpin2). 80% of the runs contain the intermediate pathway 4P1A2 ⇒ 3A4P1A2 (interaction of the intermediate intermediates hairpin1 + B1 B4 Sheet + native state). 20% of the runs contain the intermediate pathway 3A4P1⇒ 3A4P1A2 (interaction of the intermediate intermediates hairpin2 + B1 B4 Sheet + native state). 63.63% of the runs contain the intermediate pathway 3A4N1A2 ⇒ 3A4P1A2 (interaction of the intermediates harpin1 + hairpin2 + native state).
- 63.63% of the runs contains the pathway 1A2 + 3A4⇒3A4N1A2⇒3A4P1A2. 80% of the runs contains the pathway 1A2 ⇒3P1A2⇒3A4P1A2. 52.72% of the runs contains the pathway 1P4 ⇒4P1A2⇒3A4P1A2. 3.63% of the runs contains the pathway 3A4 ⇒3A4P1⇒3A4P1A2. 16.36% of the runs contains the pathway 1P4 ⇒ 3A4P1⇒ 3A4P1A2.
- 5.45% of the runs contain all the possible pathways in the same graph.

- Check the nucleus residues.
  - The nuclei residues identify experimentally are Y3,L5 (Hairpin1), F30 (Helix), W43,Y45,F52 (Hairpin2). The runs contains a total of 4220 structures, for which the nuclei residues are present in the following proportions Y3(22.46%), L5(88.67%), W43(9.36%), Y45(13.29%) and F52(85.1%).

- Compare the nucleus residues with ubiquitin family.
  - Y3, L5, F30, W43, and F52 show low sequence entropy over aligned sequences in the ubiquitin superfamily. The aligned sequences in the ubiquitin superfamily (with respect to the protein G) that show a frequency greater than 0 are:

L5:K6(99.52%), F30:V26(82.54%), W43:L43(85.37%), and Y45:F45(42.45%). The sequence and structural alignment was taken from (Similarity of Protein G and Ubiquitin). It is important to stress that all the matches belongs to the secondary structures and that the atoms in the matched residues can be superimposed. The protein reference was 1UBQ.

6.2 Pfam Benchmark

6.2.1 SH3 Domain

- Check the probability of all the intermediates.
  - During the forward step, the interaction B1-B2 was selected 100% of the times as a first folding event. The interaction B1B2B3 was selected 98.33% of the times as the topology produced by the adding of a third strand. For the addition of a four strand, the interaction B1B2B3B4 was selected in 56.6% of the opportunities, meanwhile the topology B5B1B2B3 was selected in the remaining 43.3%.
  - 100% of the runs contain the topology 1A2 (needed for all the folding events). 100% of the runs contain the topology 1A2A3 (Needed for the creation of the packed three-stranded b -sheets that form a single hydrophobic core). 98.3% of the runs contains the topology 5A1 (Needed for the creation of the packed three-stranded b -sheets that form a single hydrophobic core) 100% of the runs contains the topology 1A2A3A4(needed for the topology before aggregation). 98.3% of the runs contain the topology 5A1A2A3. 96.6% of the runs contain all the possible intermediates to build all the reported pathways.
  - 100% of the runs contain the intermediate pathway 1A2 $\Rightarrow$ 1A2A3 (interaction needed for the creation of the packed three-stranded b -sheets that form a single hydrophobic core). 98.33% of the runs contain the intermediate pathway 1A2 $\Rightarrow$ 5A1A2 (interaction needed for the creation of the packed two-stranded b -sheets that form a single hydrophobic core). 96.66% of the runs contain the intermediate pathway 1A2A3 $\Rightarrow$ 1A2A3A4 (interaction of the intermediate intermediates before aggregation). 68.33% of the runs contain the intermediate pathway 1A2A3$\Rightarrow$ 5A1N2A3A4 (interaction needed in the hypothetical state where both packed stranded b-sheets are folded independently). 83% of the runs contain the intermediate pathway 1A2A3A4 $\Rightarrow$ 5A1A2A3A4 (needed to obtain the native state without aggregation). It is important to note that the protein 1NEG present the most divergent results between the pfam family.
  - 83.33% of the runs contains the pathway 1A2 $\Rightarrow$1A2A3$\Rightarrow$1A2A3A4$\Rightarrow$5A1A2A3A4.

All the proteins has a probability ¿0.93 with the exeption of 1NEG(46.6%). 11.66% of the runs contains the pathway 1A2 ⇒5A1A2⇒5A1A2A3⇒5A1A2A3A4. 46.66% of the runs contains the pathway 1A2 ⇒1A2A3⇒5A1N2A3A4⇒5A1A2A3A4.

- 11.66% of the runs contain all the possible pathways in the same graph.

- Check the nucleus residues.
  - L24 (numbering from the src domain) is a highly conserved position in the SH3 fold family. It has been shown experimentally to be at least partially involved in the TSE ⇒ 1I0C (L18) = 19%, 1NEG(L24) = 3.5%, 1OOT(L18) 30%, 2HDA(L21) 25.2%.
  - Our results agrees with experimental results indicating that the second, third, and to a lesser extent the fourth b strands are the most ordered regions of the TSE.

6.2.2 PF00240

- Check the probability of all the intermediates.
  - During the forward step, all the intermediate steps are considered as a first folded event in at least one of the runs. Specifically, the hairpin 1 was selected 29.54% of the times as a first folding event, hairpin 2 was also selected 29.54% of the times as the first folding event, B1 B4 sheet was select 34% of the times as the first folding event.
  - 100% of the runs contain the topology 1A2 (Needed for the Intermediate hairpin 1 and hairpin 2). 100% of the runs contain the topology 1P2 (Needed for the intermediate B1 B4). 88.63% of the runs contain the topology 4P1A2 (Needed for the interaction of the intermediates hairpin1 + B1 B4 Sheet). 70.45% of the runs contains the topology 3A4P1 (needed for the interaction of the intermediates hairpin2 + B1-B4 Sheet). 70.45% of the runs contains the topology 3A4N1A2 (needed for the interaction of the intermediates harpin1 + hairpin2). 52.27% of the runs contain the topologies 1A2 & 1P2 & 4P1A2 & 3A4P1 & 3A4N1A2 & 3A4P1A2 (All possible pathways).
  - 89% of the runs contain the intermediate pathway 1P4 ⇒ 4P1A2 (interaction of the intermediates B1 B4 Sheet + hairpin1). 59.36% of the runs contain the intermediate pathway 1P4 ⇒ 3A4P1 (needed for the interaction of the intermediates B1-B4 Sheet + hairpin2). 68.88% of the runs contain the intermediate pathway 1A2 ⇒ 4P1A2 (interaction of the intermediates hairpin1 + B1 B4 Sheet). 73.17% of the runs contain the intermediate pathway 1A2 + 3A4 ⇒ 3A4N1A2 (needed for the interaction of the intermediates harpin1 + hairpin2). 72.85% of the runs contain the intermediate

pathway 4P1A2 ⇒ 3A4P1A2 (interaction of the intermediate intermediates hairpin1 + B1 B4 Sheet + native state). 48.1% of the runs contain the intermediate pathway 3A4P1⇒ 3A4P1A2 (interaction of the intermediate intermediates hairpin2 + B1 B4 Sheet + native state). 45.87% of the runs contain the intermediate pathway 3A4N1A2 ⇒ 3A4P1A2 (interaction of the intermediates harpin1 + hairpin2 + native state).

- 45.87% of the runs contains the pathway 1A2 + 3A4⇒3A4N1A2⇒3A4P1A2. 59.52% of the runs contains the pathway 1A2 ⇒3P1A2⇒3A4P1A2. 70.63% of the runs contains the pathway 1P4 ⇒4P1A2⇒3A4P1A2. 20% of the runs contains the pathway 3A4 ⇒3A4P1⇒3A4P1A2. 45.87% of the runs contains the pathway 1P4 ⇒ 3A4P1⇒ 3A4P1A2.
-

- Compare the nucleus residues with Protein G..
  - The aligned sequences in the ubiquitin superfamily (with respect to the protein G) that show a frequency greater than 0 are: L5:K306(92.77%), F30:V326(31%), W43:L343(30.93%), Y45:F345(27.43%) and F52:L367(61.2%). The sequence and structural alignment was taken from (Similarity of Protein G and Ubiquitin). It is important to stress that all the matches belongs to the secondary structures and that the atoms in the matched residues can be superimposed. The protein reference was 1UBQ and 1CMX.

6.2.3 PF01423 Kunitz Domain

- Check the probability of all the intermediates.
  - The SM1 sequence motif (which corresponds to B1,B2 and B3) are formed as first folding event in the 95.5% of the runs. The SM2 sequence motif (which corresponds to B4,B5) are formed as first folding event in the 4.5% of the runs. The two SM groups are formed and other folding events are limited.
  - 100% of the runs contain the topology 1A2 (Needed for the SM1 and SM2 sequence motifs). 97.78% of the runs contain the topology 1A2A3 (Needed for the SM1 motif). 88.89% of the runs contain the topology 1A2A3A4 and 66.67% contain the topology 5A1A2A3 (Those two topologies allow the formation of the SM1 motif followed of the SM2 motif)
  - 97.78% of the runs contain the intermediate pathway 1A2 ⇒ 1A2A3 (interaction needed for the formation of the SM1 motif). 66.67% of the runs contain the intermediate pathway 1A4 ⇒ 5A1A4 (interaction needed for the formation of the SM2

motif). 73.34% of the runs contain the intermediate pathway 1A2A3⇒1A2A3A4 (This interaction is needed for the formation of the SM1 motif and the start formation of the SM2 motif through the strand B4). 35.6% of the runs contain the intermediate pathway 1A2A3⇒5A1A2A3 (This interaction is needed for the formation of the SM1 motif and the start formation of the SM2 motif through the strand B5). 11.11% of the runs contain the intermediate pathway 5A1A2⇒5A1A2A4 (This interaction is needed for the formation of the SM2 motif and the start formation of the SM1 motif through the strand B1 and B2. 86.66% of the runs contain the intermediate 1A2A3A4⇒5A1A2A3A4 (Formation of the SM1 motif followed of the SM2 motif) and 33.33% of the runs contain the intermediate 5A1A2A3⇒ 5A1A2A3A4 (Formation of the SM1 motif followed of the SM2 motif OR formation of the SM2 motif followed of the SM1 motif).

- 76.7% of the runs contain the pathway 1A2 ⇒ 1A2A3⇒1A2A3A4⇒1A2A3A4A5 (Formation of the SM1 motif followed of the SM2 motif). 26.7% of the runs contain the pathway 1A2 ⇒ 1A2A3⇒5A1A2A3⇒1A2A3A4A5 (Formation of the SM1 motif followed of the SM2 motif). 3.33% of the runs contain the pathway 1A2⇒3A1A2⇒4A1A2A3⇒5A1 (Formation of the SM1 motif followed of the SM2 motif)
- There are no runs that contains all the possible pathways.

6.2.4 PF00240

- Check the probability of all the intermediates.
  - During the forward step, all the intermediate steps are considered as a first folded event in at least one of the runs. Specifically, the hairpin 1 was selected 29.54% of the times as a first folding event, hairpin 2 was also selected 29.54% of the times as the first folding event, B1 B4 sheet was select 34% of the times as the first folding event.
  - 100% of the runs contain the topology 1A2 (Needed for the Intermediate hairpin 1 and hairpin 2). 100% of the runs contain the topology 1P2 (Needed for the intermediate B1 B4). 88.63% of the runs contain the topology 4P1A2 (Needed for the interaction of the intermediates hairpin1 + B1 B4 Sheet). 70.45% of the runs contains the topology 3A4P1 (needed for the interaction of the intermediates hairpin2 + B1-B4 Sheet). 70.45% of the runs contains the topology 3A4N1A2 (needed for the interaction of the intermediates harpin1 + hairpin2). 52.27% of the runs contain the topologies 1A2 & 1P2 & 4P1A2 & 3A4P1 & 3A4N1A2 & 3A4P1A2 (All possible pathways).

- 89% of the runs contain the intermediate pathway 1P4 ⇒ 4P1A2 (interaction of the intermediates B1 B4 Sheet + hairpin1). 59.36% of the runs contain the intermediate pathway 1P4 ⇒ 3A4P1 (needed for the interaction of the intermediates B1-B4 Sheet + hairpin2). 68.88% of the runs contain the intermediate pathway 1A2 ⇒ 4P1A2 (interaction of the intermediates hairpin1 + B1 B4 Sheet). 73.17% of the runs contain the intermediate pathway 1A2 + 3A4 ⇒ 3A4N1A2 (needed for the interaction of the intermediates harpin1 + hairpin2). 72.85% of the runs contain the intermediate pathway 4P1A2 ⇒ 3A4P1A2 (interaction of the intermediate intermediates hairpin1 + B1 B4 Sheet + native state). 48.1% of the runs contain the intermediate pathway 3A4P1⇒ 3A4P1A2 (interaction of the intermediate intermediates hairpin2 + B1 B4 Sheet + native state). 45.87% of the runs contain the intermediate pathway 3A4N1A2 ⇒ 3A4P1A2 (interaction of the intermediates harpin1 + hairpin2 + native state).
- 45.87% of the runs contains the pathway 1A2 + 3A4⇒3A4N1A2⇒3A4P1A2. 59.52% of the runs contains the pathway 1A2 ⇒3P1A2⇒3A4P1A2. 70.63% of the runs contains the pathway 1P4 ⇒4P1A2⇒3A4P1A2. 20% of the runs contains the pathway 3A4 ⇒3A4P1⇒3A4P1A2. 45.87% of the runs contains the pathway 1P4 ⇒ 3A4P1⇒ 3A4P1A2.
-

- Compare the nucleus residues with Protein G..
  - The aligned sequences in the ubiquitin superfamily (with respect to the protein G) that show a frequency greater than 0 are: L5:K306(92.77%), F30:V326(31%), W43:L343(30.93%), Y45:F345(27.43%) and F52:L367(61.2%). The sequence and structural alignment was taken from (Similarity of Protein G and Ubiquitin). It is important to stress that all the matches belongs to the secondary structures and that the atoms in the matched residues can be superimposed. The protein reference was 1UBQ and 1CMX.

7 Conclusions

References

[1] O.N. Jensen. Interpreting the protein language using proteomics. *Nature Reviews Molecular Cell Biology*, 7(6):391–403, 2006.

[2] D. Kihara, Y. Zhang, H. Lu, A. Kolinski, and J. Skolnick. Ab initio protein structure prediction on a genomic scale: Application to the mycoplasma genitalium genome. *Proceedings of the National Academy of Sciences*, 99(9):5993, 2002.

[3] D.C. Liebler. *Introduction to proteomics: tools for the new biology*. Humana Pr Inc, 2002.

[4] A. Fiser. Protein structure modeling in the proteomics era. *Expert review of proteomics*, 1(1):97–110, 2004.

[5] Aashish N Adhikari, Karl F Freed, and Tobin R Sosnick. De novo prediction of protein folding pathways and structure using the principle of sequential stabilization. *Proceedings of the National Academy of Sciences*, 109(43):17442–17447, 2012.

[6] Nancy M Amato, Ken A Dill, and Guang Song. Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures. *Journal of Computational Biology*, 10(3-4):239–255, 2003.

[7] Mohammed J Zaki, Vinay Nadimpally, Deb Bardhan, and Chris Bystroff. Predicting protein folding pathways. *Bioinformatics*, 20(suppl 1):i386–i393, 2004.

[8] Vibin Ramakrishnan, Sai Praveen Srinivasan, Saeed M Salem, Suzanne J Matthews, Wilfredo Colón, Mohammed Zaki, and Christopher Bystroff. Geofold: Topology-based protein unfolding pathways capture the effects of engineered disulfides on kinetic stability. *Proteins: Structure, Function, and Bioinformatics*, 80(3):920–934, 2012.

[9] B. Berger and T. Leighton. Protein folding in the hydrophobic-hydrophilic (hp) model is np-complete. *Journal of Computational Biology*, 5(1):27–40, 1998.

[10] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, and M. Yannakakis. On the complexity of protein folding. *Journal of computational biology*, 5(3):423–465, 1998.

[11] D.E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R.O. Dror, M.P. Eastwood, J.A. Bank, J.M. Jumper, J.K. Salmon, Y. Shan, et al. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341–346, 2010.

[12] R.O. Dror, A.C. Pan, D.H. Arlow, D.W. Borhani, P. Maragakis, Y. Shan, H. Xu, and D.E. Shaw. Pathway and mechanism of drug binding to g-protein-coupled receptors. *Proceedings of the National Academy of Sciences*, 108(32):13118–13123, 2011.

[13] S. Pronk, P. Larsson, I. Pouya, G.R. Bowman, I.S. Haque, K. Beauchamp, B. Hess, V.S. Pande, P.M. Kasson, and E. Lindahl. Copernicus: A new paradigm for parallel adaptive molecular dynamics. In *High Performance Computing, Networking, Storage and Analysis (SC), 2011 International Conference for*, pages 1–10. IEEE, 2011.

[14] S. Piana, K. Lindorff-Larsen, and D.E. Shaw. Protein folding kinetics and thermodynamics from atomistic simulation. *Proceedings of the National Academy of Sciences*, 2012.

[15] J. Waldispühl, C.W. O'Donnell, S. Devadas, P. Clote, and B. Berger. Modeling ensembles of transmembrane β-barrel proteins. *Proteins: Structure, Function, and Bioinformatics*, 71(3):1097–1112, 2008.

[16] D.S. Marks, L.J. Colwell, R. Sheridan, T.A. Hopf, A. Pagnani, R. Zecchina, and C. Sander. Protein 3d structure computed from evolutionary sequence variation. *PloS one*, 6(12):e28766, 2011.

[17] Y. Ding and C.E. Lawrence. A statistical sampling algorithm for rna secondary structure prediction. *Nucleic acids research*, 31(24):7280–7301, 2003.

[18] J.S. McCaskill. The equilibrium partition function and base pair binding probabilities for rna secondary structure. *Biopolymers*, 29(6-7):1105–1119, 2004.

[19] B.C. Foat, A.V. Morozov, and H.J. Bussemaker. Statistical mechanical modeling of genome-wide transcription factor occupancy data by matrixreduce. *Bioinformatics*, 22(14):e141–e149, 2006.

[20] R.V. Chereji, D. Tolkunov, G. Locke, and A.V. Morozov. Statistical mechanics of nucleosome ordering by chromatin-structure-induced two-body interactions. *Physical Review E*, 83(5):050903, 2011.

[21] B. Fain and M. Levitt. A novel method for sampling alpha-helical protein backbones. *Journal of molecular biology*, 305(2):191–201, 2001.

[22] J. Waldispühl, B. Berger, P. Clote, and J.M. Steyaert. Predicting transmembrane β-barrels and interstrand residue interactions from sequence. *PROTEINS: Structure, Function, and Bioinformatics*, 65(1):61–74, 2006.

[23] J. Waldispühl and J.M. Steyaert. Modeling and predicting all-alpha transmembrane proteins including helix-helix pairing. *Theoretical computer science*, 335(1):67–92, 2005.

[24] S. Shenker, C. ODonnell, S. Devadas, B. Berger, and J. Waldispühl. Efficient traversal of beta-sheet protein folding pathways using ensemble models. In *Research in Computational Molecular Biology*, pages 408–423. Springer, 2011.

[25] J. Skolnick, A. Kolinski, and A.R. Ortiz. Monsster: a method for folding globular proteins with a small number of distance restraints1. *Journal of molecular biology*, 265(2):217–241, 1997.

[26] A.R. Ortiz, A. Kolinski, and J. Skolnick. Nativelike topology assembly of small proteins using predicted restraints in monte carlo folding simulations. *Proceedings of the National Academy of Sciences*, 95(3):1020, 1998.

[27] A.R. Ortiz, A. Kolinski, P. Rotkiewicz, B. Ilkowski, and J. Skolnick. Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins: Structure, Function, and Bioinformatics*, 37(S3):177–185, 1999.

[28] S. Wu, A. Szilagyi, and Y. Zhang. Improving protein structure prediction using multiple sequence-based contact predictions. *Structure*, 19(8):1182–1191, 2011.

[29] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D.S. Marks, C. Sander, R. Zecchina, J.N. Onuchic, T. Hwa, and M. Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

[30] T.A. Hopf, L.J. Colwell, R. Sheridan, B. Rost, C. Sander, and D.S. Marks. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, 2012.

[31] S. Kmiecik and A. Kolinski. Folding pathway of the b1 domain of protein g explored by multiscale modeling. *Biophysical journal*, 94(3):726–736, 2008.

[32] I.A. Hubner, J. Shimada, and E.I. Shakhnovich. Commitment and nucleation in the protein g transition state. *Journal of molecular biology*, 336(3):745–761, 2004.

[33] A.M. Gronenborn, D.R. Filpula, N.Z. Essig, A. Achari, M. Whitlow, P.T. Wingfield, GM Clore, et al. A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein g. *Science(Washington)*, 253(5020):657–661, 1991.

[34] F.J. Blanco, G. Rivas, and L. Serrano. A short linear peptide that folds into a native stable β-hairpin in aqueous solution. *Nature Structural & Molecular Biology*, 1(9):584–590, 1994.

[35] J. Kuszewski, G.M. Clore, and A.M. Gronenborn. Fast folding of a prototypic polypeptide: the immunoglobulin binding domain of streptococcal protein g. *Protein Science*, 3(11):1945–1952, 2008.

[36] Sophie E Jackson. Ubiquitin: a small protein folding paradigm. *Organic & biomolecular chemistry*, 4(10):1845–1853, 2006.

[37] Jianlin Cheng and Pierre Baldi. Three-stage prediction of protein β-sheets by neural networks, alignments and graph algorithms. *Bioinformatics*, 21(suppl 1):i75–i84, 2005.

[38] Ajazul Hamid Wani and Jayant B Udgaonkar. Revealing a concealed intermediate that forms after the rate-limiting step of refolding of the sh3 domain of pi3 kinase. *Journal of molecular biology*, 387(2):348–362, 2009.

[39] Isaac A Hubner, Katherine A Edmonds, and Eugene I Shakhnovich. Nucleation and the transition state of the sh3 domain. *Journal of molecular biology*, 349(2):424–434, 2005.

[40] Feng Ding, Weihua Guo, Nikolay V Dokholyan, Eugene I Shakhnovich, and Joan-Emma Shea. Reconstruction of the src-sh3 protein domain transition state ensemble using multiscale molecular dynamics simulations. *Journal of molecular biology*, 350(5):1035–1050, 2005.

[41] Viara P Grantcharova, David S Riddle, and David Baker. Long-range order in the src sh3 folding transition state. *Proceedings of the National Academy of Sciences*, 97(13):7084–7089, 2000.

[42] Jose C Martínez and Luis Serrano. The folding transition state between sh3 domains is conformationally restricted and evolutionarily conserved. *Nature Structural & Molecular Biology*, 6(11):1010–1016, 1999.

[43] G Chelvanayagam and P Argos. Prediction of protein folding pathways: Bovine pancreatic trypsin inhibitor. *Cytotechnology*, 11(1):S67–S71, 1993.

[44] Jui-Yoa Chang. Distinct folding pathways of two homologous disulfide proteins: bovine pancreatic trypsin inhibitor and tick anticoagulant peptide. *Antioxidants & redox signaling*, 14(1):127–135, 2011.

[45] A. Liwo, J. Lee, D.R. Ripoll, J. Pillardy, and H.A. Scheraga. Protein structure prediction by global optimization of a potential energy function. *Proceedings of the National Academy of Sciences*, 96(10):5482–5485, 1999.

[46] D. Becerra, A. Sandoval, D. Restrepo-Montoya, and L.F. Nino. A parallel multi-objective ab initio approach for protein structure prediction. In *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on*, pages 137–141. IEEE, 2010.

[47] M.T. Wolfinger, W.A. Svrcek-Seiler, C. Flamm, I.L. Hofacker, and P.F. Stadler. Efficient computation of rna folding dynamics. *Journal of Physics A: Mathematical and General*, 37:4731, 2004.