Report eFold

Dr. Jérôme Waldispühl and David Becerra

August 12, 2014

Contents

1 Abstract

2 Introduction

The protein structure and pathway prediction problems entail 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].

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. Comparative and fold recognition methods 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.

The protein folding problem is an NP-complete problem even in simple lattice models [5, 6] 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 [7, 8, 9, 10], 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 [11] and evolutionary information content based methods [12] 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.

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 [13, 11] and modelling the folding process of large single β-sheet proteins [14] 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 [12]. 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 [15, 16, 17]. 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 [18], or *de novo* modelling, i.e., do not use template-derived contacts or sequence-similar fragments from known structures [12, 19, 20].

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. The `efold` algorithm expands our previous `tfolder` program in several directions. First `efold` models α-helices and multiple β-sheets. Next, unlike `tfolder`, the `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. Folding pathways are predicted using, at a large scale, hierarchical folding and sequential stabilization principles. 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.

Many current obstacles presented in the protein structure prediction problem (such as the mentioned above) 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 [21, 22]. Although, those approaches can not directly be mapped to proteins, they have been an excellent starting point to model more accurate and complex scenarios [23, 24, 25]. With respect to the protein structure scheme, we have already introduced a structural ensemble predictor for transmembrane β-barrel (TMB) proteins [11] continuing earlier work on molecular structure modelling [13, 26]. Recently, we introduced a method for modelling the folding process of large β-sheet proteins using sequence data alone [14].

3 Materials

3.1 Study Case Benchmark

This benchmark is composed by two proteins for which there is a vast body of experimental and theoretical studies for which the pathways of these proteins have been elucidated [27, 28]. Then, to evaluate the performance and efficiency of our approach, the folding landscape and contact prediction maps were reconstructed for the B1 domain of protein G and the Ubiquitin.

The B1 domain of protein G , generally called GB1, has played a central role in protein folding studies being the system of choice in more than 200 publications carried out using a wide variety of experimental and theoretical approaches. Because of its small size and its simple and highly symmetrical topology, this small protein domain has represented an ideal candidate for a vast number of different studies. The B1 domain of protein G 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 [29]. It was previously shown that protein G folds through three pathways, all of which pass through an intermediate, to a single TS. The three intermediates feature a near-native helix along with hairpin 1 (I_1 intermediate), hairpin 2 (I_2), or the b1b 4 sheet (I_3). The work [30, 31] 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 [28]. Namely, three folding pathways are observed, each involving formation of its own assembly: helix-first hairpin, helix-second hairpin, and b 1b 4. All pathways appeared to converge to the same folding nucleus.

The extensive use of ubiquitin results from its favourable propertiesit is a small protein

(76 residues in length) that has a highly structured native state which is very stable. Its high stability has been known for some time and this 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 b-hairpin and the a-helix. Such structure will be transient, flickering in-and-out. By the rate-limiting transition state this structure has consolidatedthe helix is almost fully formed and the b-hairpin is partially structured. In this state, the secondary structure is stabilised by interactions between the b-hairpin and the helix. The C-terminal region remains relatively unstructured until after the TS barrier has been crossed. It is formed rapidly in a downhill process post-TS. 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 stabilising salt such as sodium sulfate.

3.2 EvFold Benchmark

The original evFold benchmark is composed by 15 protein structures ranging from 48 to 258 amino acids in size.

3.3 916 Benchmark

"The original dataset is extracted from the Protein Data Bank of May 2004. Only structures determined by X-ray diffraction and having resolution better than 2.5 are retained. Chains containing unknown or non-standard amino acids, backbone interruptions or whose length is ¡50 amino acids are excluded. The redundancy in the dataset is reduced by the UniqueProt (Mika and Rost, 2003) with a HSSP threshold of 0, which corresponds to sequence identity of roughly 1520%. The final dataset contains 916 chains corresponding to 187 516 residues. Of these, 26% (48 996) are ?-residues participating in 31 638 interstrand residue pairs. The dataset has 10 745 ?-strands with an average length of 4.6 residues and 8172 ?-strand pairs, including 4519 antiparallel pairs, 2214 parallel pairs and 1439 pairs involving isolated ?-bridges. These strand pairs form 2533 ?-sheets. The average sequence separation between residue pairs and strand pairs is 43 and 40, respectively. Sequence separation histograms are displayed in Figure 2a and 2b. Figure 2c and 2d shows that the number of interstrand residue pairs or strand pairs has a strong correlation with the number of ?-residues or strands in the chain, as expected."

3.4 Pfam Benchmark

3.4.1 SH3 Domain

SH3 presents an ideal system for studying the nucleation-condensation mechanism and highlights the synergistic relationship between experiment and simulation in the study of protein folding. SH3 is a widely studied protein that exhibits two state folding, and is composed of two orthogonally packed three-stranded b -sheets that form a single hydrophobic core. SH3 domains from many different proteins have been used as model systems to address various important aspects of protein folding, such as the nature of protein folding transition states,29, 30 and 31 and the relationship between protein topology and the folding pathway. The folding of SH3 domains has been studied by protein engineering methods NMR as well as by computational studies. "The first sheet consists of the three central strands of the protein (?2-?3-?4) and the second sheet of the two terminal strands (?1 and ?5) and a portion of the RT loop. There is also a small 310 helix between ?4 and ?5. Due to its small size and multiple homologues, it has been the target of extensive experimental and theoretical studies."

The general features of all observed SH3 TSEs include denaturation of the N and C termini, turns and loops, and a small amount of secondary structures located in the central b strands. L24 (numbering from the src domain) is a highly conserved position in the SH3 fold family, and has been shown experimentally to be at least partially involved in the TSE26.

"Through the preceding analysis, it is clear that the structure of the TSE, common to all domains, is formed by a small number of residues in the b 2-b 3 hairpin making non-local contacts to the RT loop/diverging turn and distal b -hairpin regions. The nucleus residues common to all three domains include L24, F26, E30, L32, W43, L44, A45, H46, and G51. This 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"

"Despite minor differences, the common behavior that we observe between all studied SH3 domains is the formations of the turn-RT loop, b 1-b 2, b 2-b 3, and b 3-b 4motifs immediately after the TSE. Contacts between the b 2 and b 3 strands begin forming around the transition region to various degrees, but form largely after nucleation."

3.4.2 Kunitz Domain

Around 1995, comparisons between the various LSm homologs identified two sequence motifs, 32 amino acids long and 14 amino acids long, that were very similar in each LSm homolog, and were separated by a non-conserved region of variable length. This indicated the importance of these two sequence motifs (named Sm1 and Sm2), and suggested that all LSm protein genes evolved from a single ancestral gene.

The SM1 sequence motif corresponds to the ?1, ?2, ?3 strands, and the SM2 sequence motif corresponds to the ?4 and ?5 strands.

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 [32]. 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 [33]. 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 mentioned above) 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 [21, 22]. Although, those approaches can not directly be mapped to proteins, they have been an excellent starting point to model more accurate and complex scenarios [23, 24, 25]. With respect to the protein structure scheme, we have already introduced a structural ensemble predictor for transmembrane β-barrel (TMB) proteins [11] continuing earlier work on molecular structure modelling [13, 26]. Recently, we introduced a method for modelling the folding process of large β-sheet proteins using sequence data alone [14].

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. $iii-$) 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 can be 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. 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. This likelihood is weighted using an evolutionary contact prediction method.

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 Figures 1 and 2. The specific details of the methodology are shown in the hereinafter subsections.

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, and the computation of the Boltzmann partition function for all admissible β-sheet schemas.

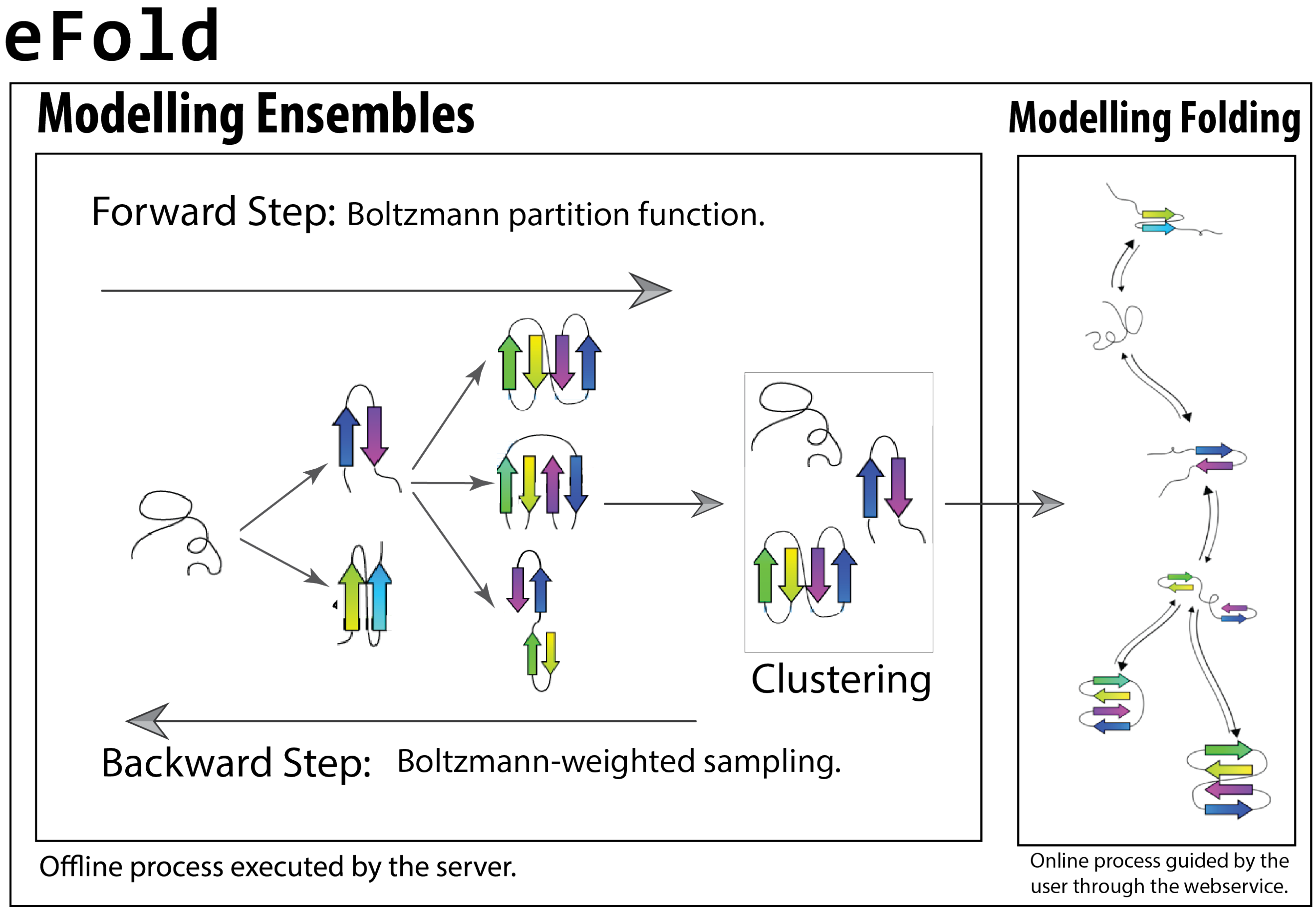

eFold
Modelling Ensembles
Modelling Folding
Forward Step: Boltzmann partition function.
Clustering
Backward Step: Boltzmann-weighted sampling.
Offline process executed by the server.
Online process guided by the user through the webservice.

Figure 1: `efold:` is the proposed algorithm for predicting protein folding pathways 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.

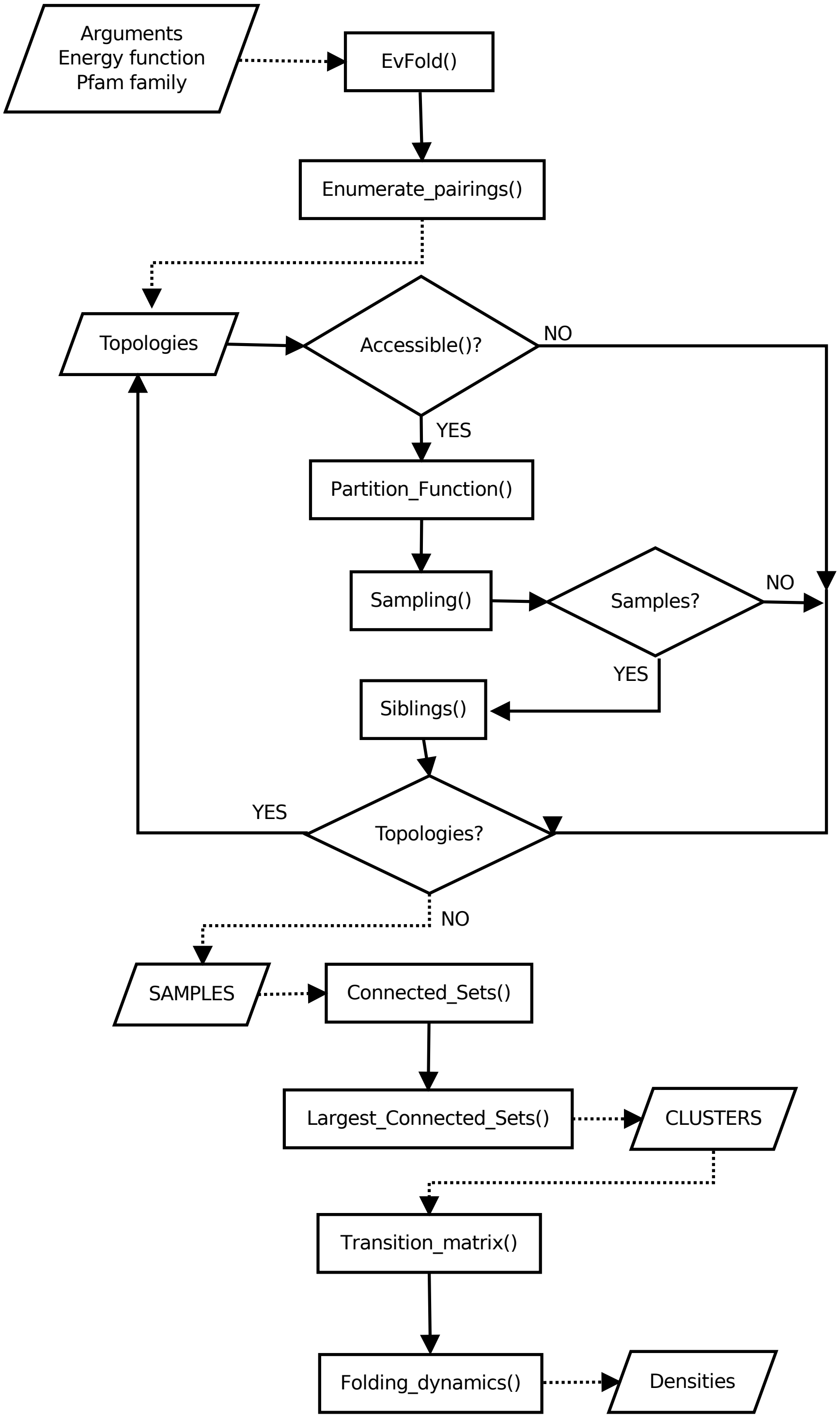


Figure 2: `efold` **Flowchart:** .

Protein Representation:

We model protein structures using a coarse grain residue-level representation, including 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 shapes were encoded using a stepwise permutation algorithm through the labeled set of β-strands $\{1 \ldots n\}$. 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, anti-parallel or none (See Figure **??**. 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.

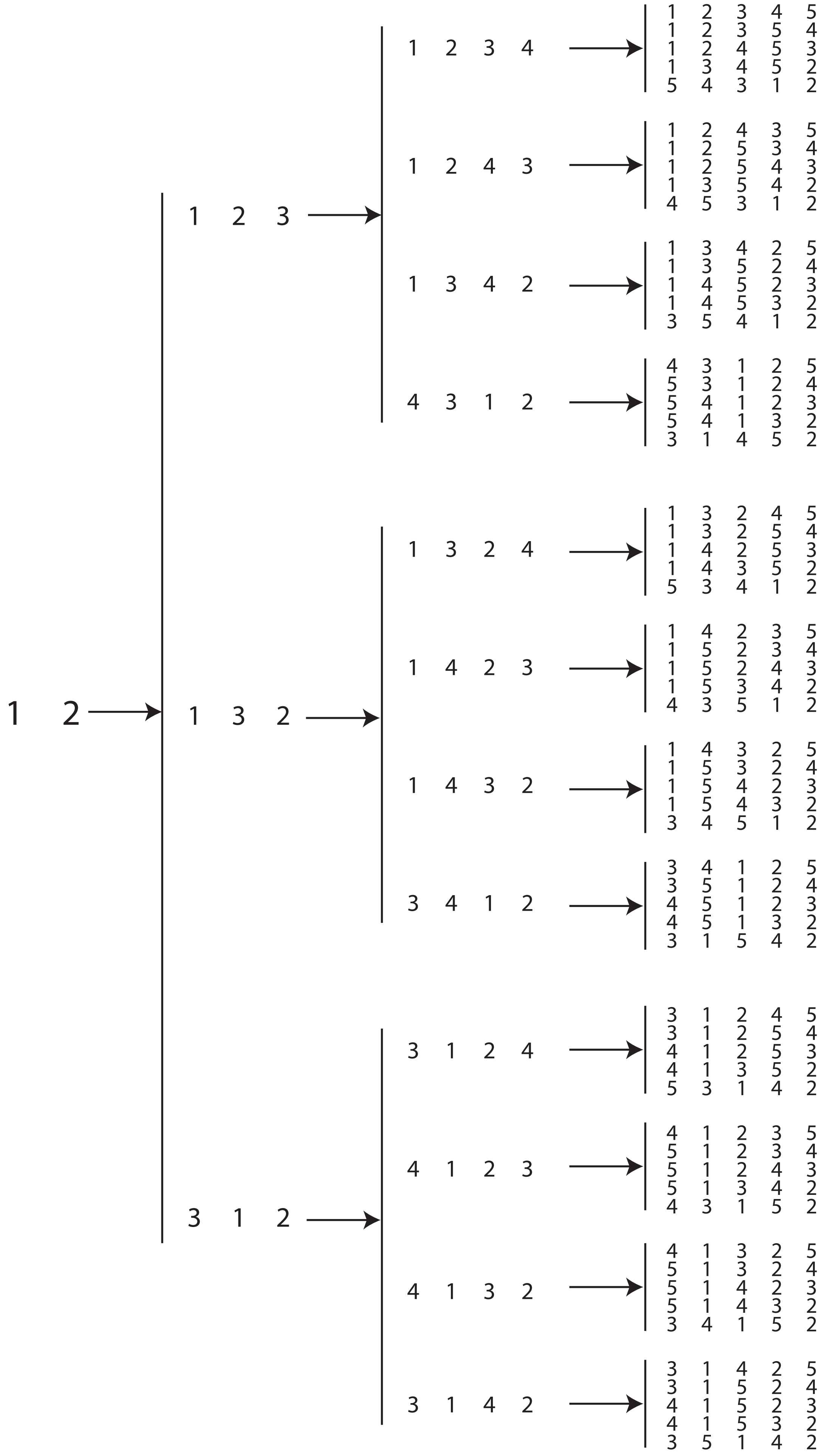
Contrary to our previous implementations, the computation of all permutable β-sheet schemas is performed through a tree, where each level of the tree contains all the instances with a specific number of strands. Then, the first level of the tree correspond to the instances that contains only two strand, the second level of the tree contains the instances with three strands, and so on until the leaves of the tree (i.e., m-level) are stored. The tree can be easily traversed given that each node has a pointer to its parent. A parent of an instance belonging to level i is defined as the instance belonging to level $i-1$ that share a sub-structure with that instance. Two instances share their structures if they are identical to each other, modulo the addition or removal of a single strand pairing.

The tree structure can be filled through a breadth-first or a deep-first approach. In the first case (i.e., BFS) the level $i+1$ is not considered until all the instances of level i have been computed. In the second case (i.e., DFS) one starts in the less-deep level of the tree (i.e., 2 strand interactions) and explores as far as possible along each branch before backtracking. 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*) The sampling procedure for the instances with i strands, can be performed before reaching the filling of the hash table. *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.

In order to develop an DFS or BFS approach, a recursion that guarantee the correct make up of the states in the graph is required. Specifically, It is desired that the offsprings present a compatible topology with the parent, that the structural similarity can be easily computed and that all the energies of intermediate structures have been already computed by the ancestors in order to compute the recursion. The recursive formulation for a tree was used to make up the nodes of the tree, where the offsprings are built considering the adding of a new strand ordered according to sequence (i.e., adding a new strand going from $i = 2$ to $i = sizeoftheparent$). The previous approach allowed us to built the tree shown in Figure 3. It is important to stress the following features of the tree: i-) It is balanced tree. ii-) All the nodes belonging to the same level have the same number of children. iii-) For each node, the number of offsprings is one more than the number of offsprings of his parent. iv-) Each added strand (i.e., last column in the template) is part of a ordered sequence of numbers that goes from $i = 2$ to $i =$size of the template.

The main idea of the recursion function is to exploits the shared sub-structures between instances in the ensemble using a memoization approach. Each recursive call must compute the energy function of a specific instance, this result is stored in a table indexed by an identifier. Subsequent recursive calls, which involves the same instance, will perform a table lookup instead of re-computing the value of the recursion. The proposed implementation is based on a memoization technique, where hash tables are used as the base data structures. A consecutive identifier is the key to access the hash table. The values stored in each node correspond to an array that contains the information of the templates, the structure of the parent and the computed energy value. The values represent the necessary information to traceback an specific protein instance.

Each node in the tree is labeled as a permutation of a set of β-strands (i.e., Template), however, each node also stores the values for all the pairings involving the template of the node. In other words, each node stores all the energy values for the schemas of the permutable β-sheets that consider the labeled template. Each node also stores the information (substructure energies) to compute each of the offsprings generated by the template. Given that the added strand in each offsprings is part of a ordered sequence of numbers (going from $i = 2$ to $i = sizeofthetemplate$), each specific node has to store the information needed as substructures for the next level. Particularly, all the configurations that will allow the addition of a new strand i, for $i = 2 \dots i =$goes from $i = 2$ to $i = sizeofthetemplate$, must be independently stored. Figure 4 shows how the computation of the Z score is performed for the template 12345. Specifically, it only needs the substructures stored (which have been already computed)


1 2
1 2 3
1 3 2
3 1 2
1 2 3 4
1 2 4 3
1 3 4 2
4 3 1 2
1 3 2 4
1 4 2 3
1 4 3 2
3 4 1 2
3 1 2 4
4 1 2 3
4 1 3 2
3 1 4 2
1 2 3 4 5
1 2 3 5 4
1 2 4 5 3
1 3 4 5 2
5 4 3 1 2
1 2 4 3 5
1 2 5 3 4
1 2 5 4 3
1 3 5 4 2
4 5 3 1 2
1 3 4 2 5
1 3 5 2 4
1 4 5 2 3
1 4 5 3 2
3 5 4 1 2
4 3 1 2 5
5 3 1 2 4
5 4 1 2 3
5 4 1 3 2
3 1 4 5 2
1 3 2 4 5
1 3 2 5 4
1 4 2 5 3
1 4 3 5 2
5 3 4 1 2
1 4 2 3 5
1 5 2 3 4
1 5 2 4 3
1 5 3 4 2
4 3 5 1 2
1 4 3 2 5
1 5 3 2 4
1 5 4 2 3
1 5 4 3 2
3 4 5 1 2
3 4 1 2 5
3 5 1 2 4
4 5 1 2 3
4 5 1 3 2
3 1 5 4 2
3 1 2 4 5
3 1 2 5 4
4 1 2 5 3
4 1 3 5 2
5 3 1 4 2
4 1 2 3 5
5 1 2 3 4
5 1 2 4 3
5 1 3 4 2
4 3 1 5 2
4 1 3 2 5
5 1 3 2 4
5 1 4 2 3
5 1 4 3 2
3 4 1 5 2
3 1 4 2 5
3 1 5 2 4
4 1 5 2 3
4 1 5 3 2
3 5 1 4 2

in the parent node and that are known to have enough space in the sequence to add a new strand in the position 5 (see the red circle area in template 1234). The underlined red circle areas in the other tree leves are drawn as a reference of the path followed by the substructures to compute the template 12345, however, it must be clear that no re-computation or look up table is performed over those tables.

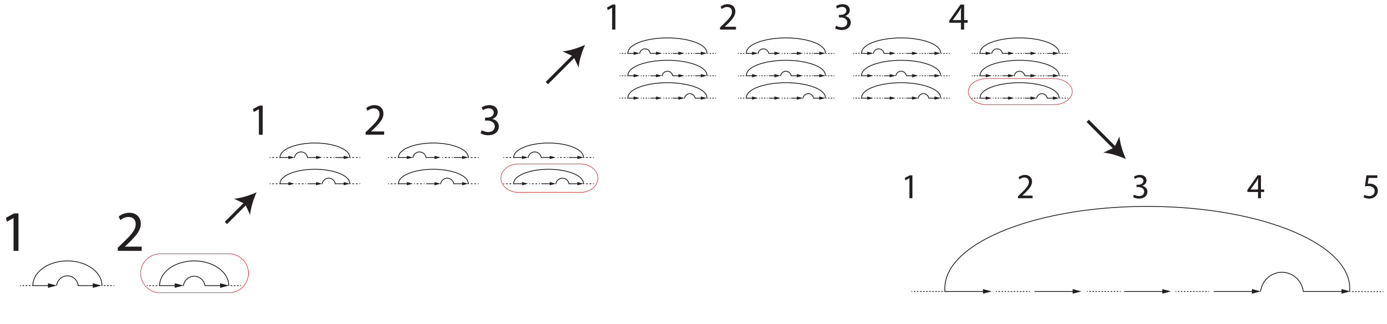


Figure 4

The proposed data structure presents the following improvements with respect to our previous implementations.

- `tfolder` re-computed all the permutable β-schemas of a set of instances with shared structures (Two instances share their structures if they are identical to each other, modulo the addition or removal of a single strand pairing). With the new implementation, this re-computation is avoided through a table lookup from the node of the parent node.

- `tfolder` fails to compute the complete set of β-schemas once a threshold of 5 strands has been gotten. With the new implementations, the computation of all β-schemas until a threshold of 6 strands is guarantee.

- `tfolder` computes the protein topologies following a specific order. Particularly, it builds the topologies adding strands from left to right. For example, in order to compute the template 1 2 3 4 5, `tfolder` will create the templates 1 2, then, it will add the left most strand (i.e., strand 3) to create the template 1 2 3. `tfolder` will go on with this process (adding in order the templates 4 and 5) until the final template is achieved. On the other hand, `efold` will create the topologies based on the most energy favourable templates (and not in a hard-coded order). Then, in order to create the template 1 2 3 4 5, `efold` can start the process choosing the most favourable template for the following topologies 1 2 or 2 3 or 3 4 or 4 5. Next, and depending on its selection, `efold` is able to add strands to the left or right of the previous structure.

- The computation of compatible topologies is simplified and it is derived from pruning procedures over the tree.

- The sampling procedure is performed as a lookup table procedure in the tree.

EvFold() A maximum entropy was used 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 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.

Computation of the Partition Function:

Conceptually, each protein structure was described by the set of residue/residue contacts that form hydrogen bonds between β-strand backbones. 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 [13]. 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 for all admissible β-sheet schemas (i.e., S_i, where $i \in \{1 \dots n\}$) 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.

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

4.1.2 The backward step.

In this step, the probabilities of the protein conformations generated in the forward step are computed. Particularly, we analyze the significance of the protein states computing its associated occurrence likelihood. These likelihoods are finally weighted using an evolutionary contact prediction method in order to circumvent the inherent limitation of potential energy scoring schemes. Then, clusters of low-energy states with similar conformations are extracted using their relative energetics. The structures belonging to those clusters are used to predict the folding dynamics (process described in the Section Folding Dynamics).

Sampling process:

A sampling process is performed over the total set of protein states in order to work with a tractably sized system. A linear-order algorithm was implemented to model a Boltzmann-weighted random selection procedure. Specifically, the distribution of structures in the Boltzmann ensemble is used to randomly generate low energy protein structures. The sampling algorithm is based on the approach developed for RNA secondary structures in [21].

$$weight_{sample} = \prod_{i=2}^{m} \frac{\exp[-E(S_{k,i})/RT]}{Z_i} \tag{5}$$

Once a set of structures has been sampled, a traceback procedure through the hash table is performed to generate a weight value for each structure. The main idea is to compute a value that represents the likelihood of the structure to be generated. Then, in our approach, all the inward compatible structures (i.e., structures that are identical to each other, modulo the removal of a single strand pairing) are found. Furthermore, the weight function defined in Equation 5 is computed for all the set of compatible structures, where where m is the number of strands, $E(S_{k,i})$ is the energy of the sampled topology k with i strands, and Z_i is the partition function computed for the sub topology composed by i strands.

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 [34]. 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-weighted 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 the 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, we partition the state space into macro states using clustering, in order to work with a tractably sized system. 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 is calculated. 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, the distribution of conformations over folding time is estimated by solving this system.

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

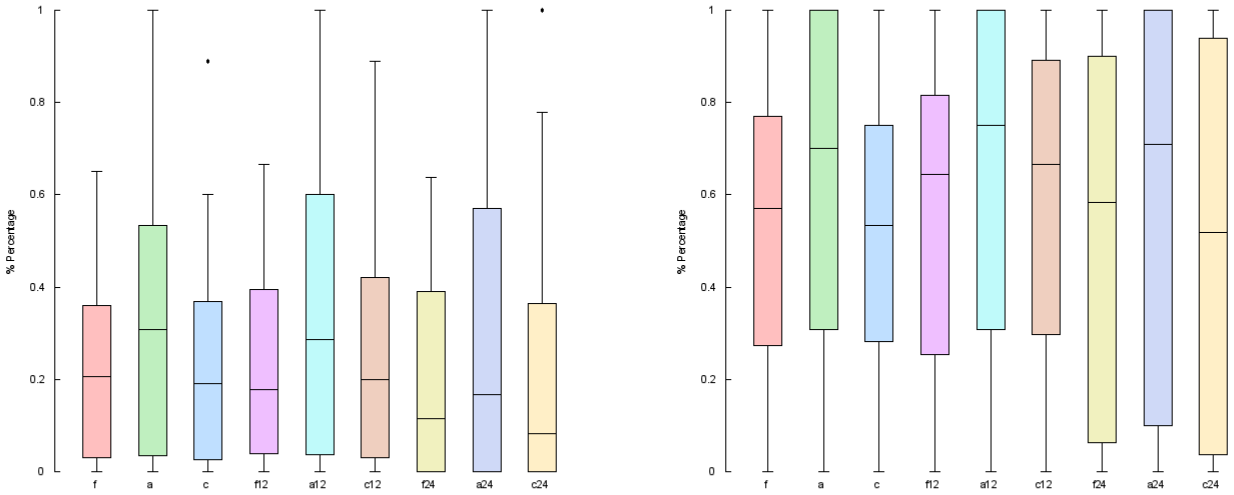
5 Experimental Framework

In order to try to give some insights with respect the limitations of the proposed method and the parameter optimization, the software architecture of the algorithm was thought to perform two main steps, the modelling ensemble and modelling folding phases (See Figure 1 and the Method Section). 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. This architecture is highly convenient because *i*) It allows us to compute the procedures with a heaviest load of resources in an off-line fashion, *ii*) It allows the user to simulate different folding predictions changing a variety of key parameters. This makes easier and more intuitive the understanding and manipulation of the folding pathway predictions; moreover, the optimization of these parameters can be done through heuristics or high level information introduced by the user.

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.

6 Results

6.1 All



(a) Best 1

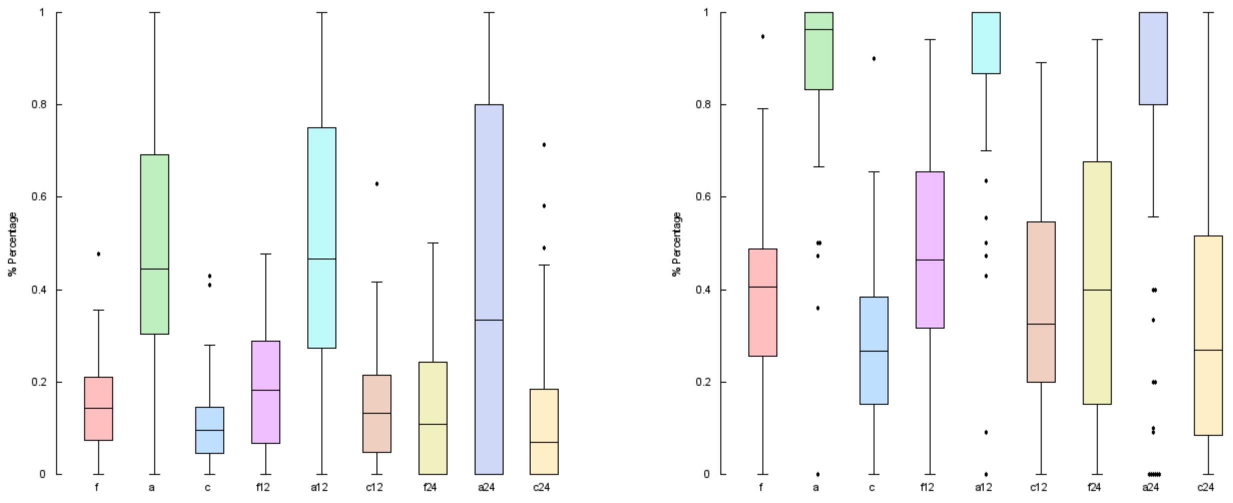
(b) Best 1 +2

Figure 5: Best 1 - All

6.2 Study Case Benchmark

6.2.1 Protein G

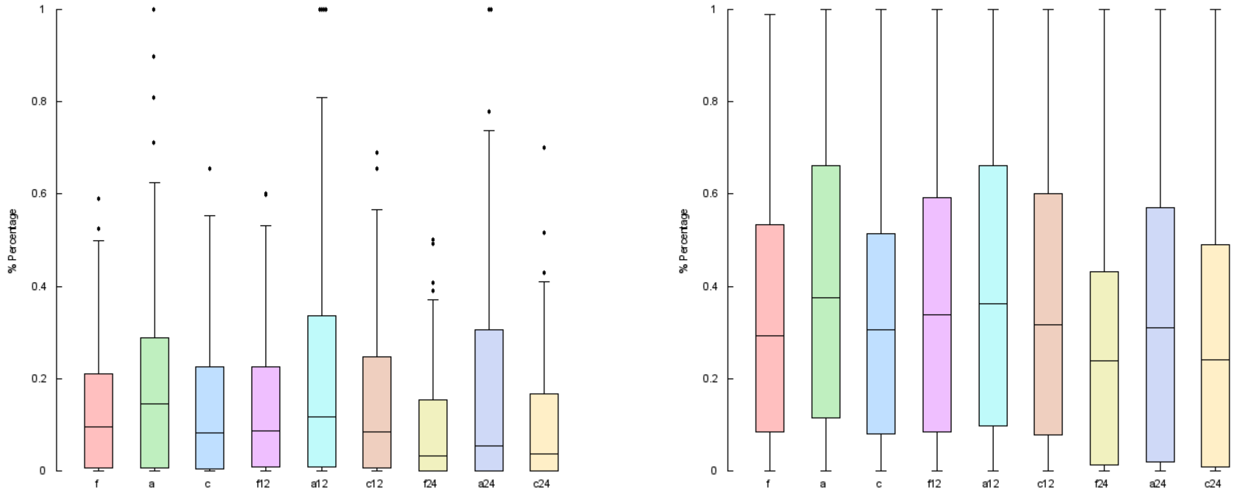
- Check the probability of all the intermediates.
  - During the forward step, all the intermediate steps are considered as a first folded event in at least one of the runs. Specifically, the hairpin 1 was selected 56% of the times as a first folding event, hairpin 2 was selected 4% of the times as the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



(a) Best 2

(b) Best 2 +2

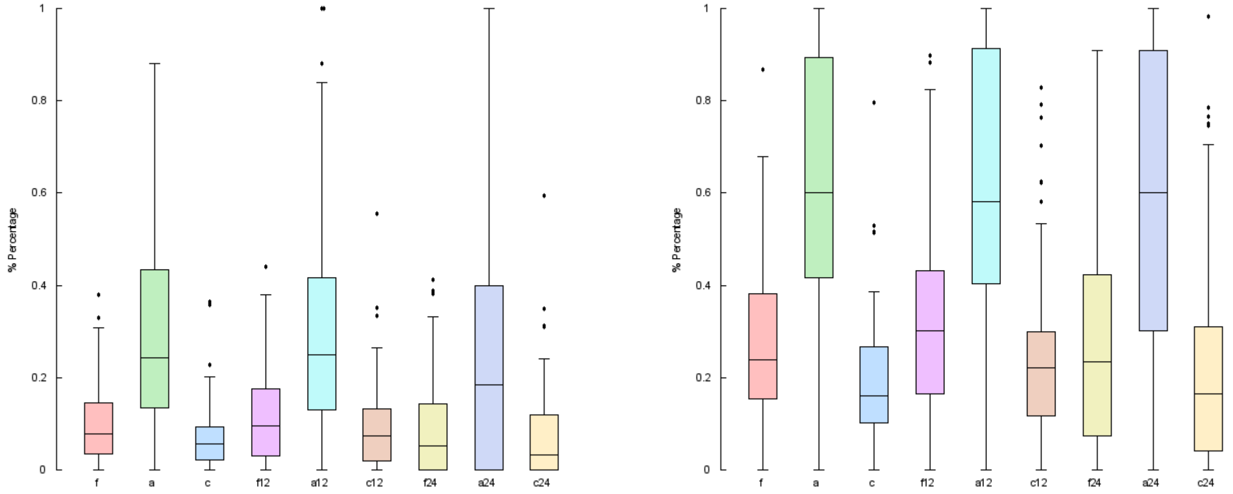
Figure 6: Best 2 - All



(a) Average Cluster 1

(b) Average Cluster 1 +2

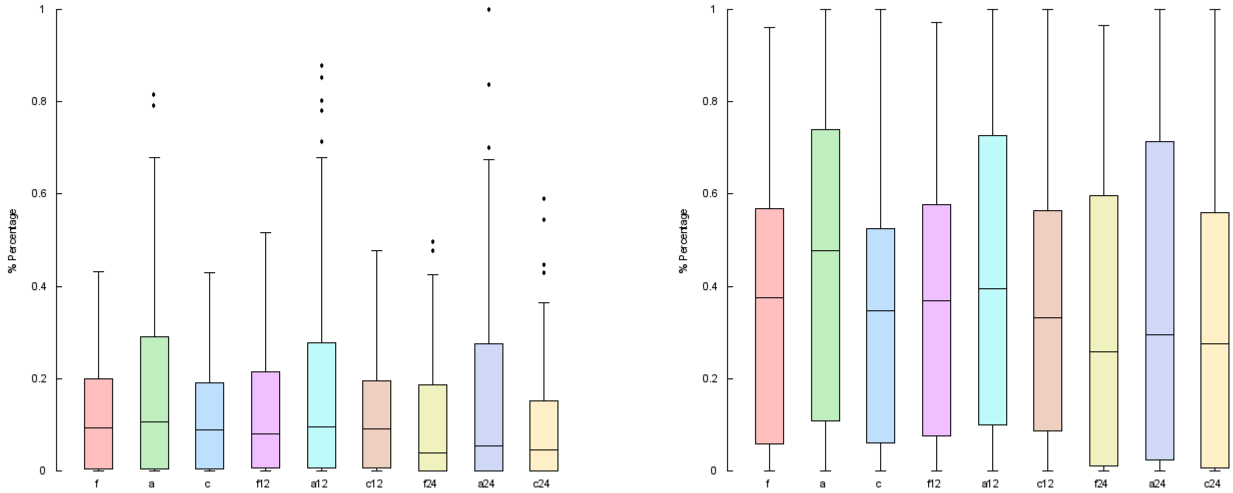
Figure 7: Average Cluster 1 - All



(a) Average Cluster 2

(b) Average Cluster 2 +2

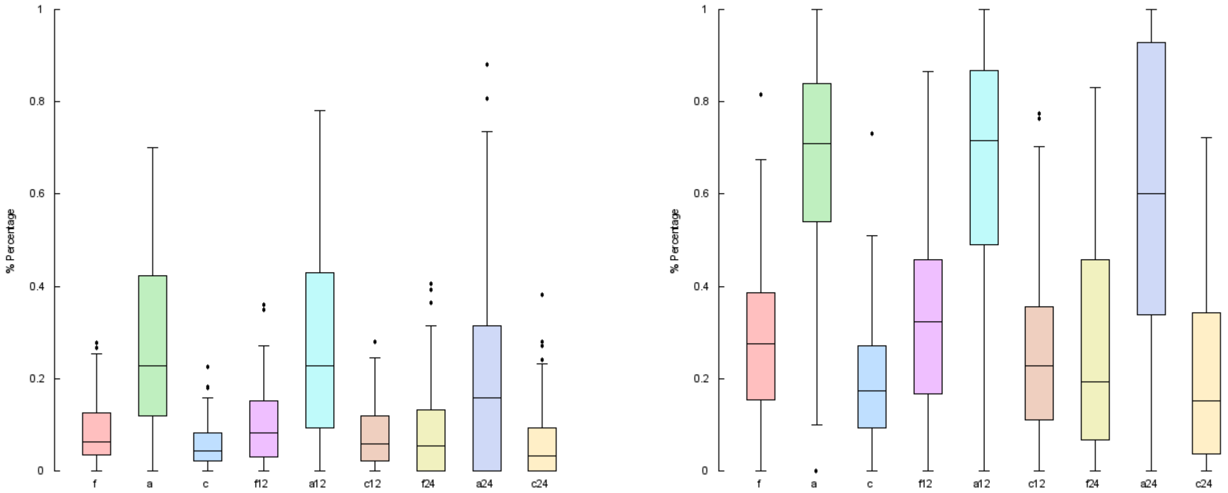
Figure 8: Average Cluster 2 - All



(a) Average 1

(b) Average 1 +2

Figure 9: Average 1 -All



(a) Average 2

(b) Average 2 +2

Figure 10: Average 2 - All

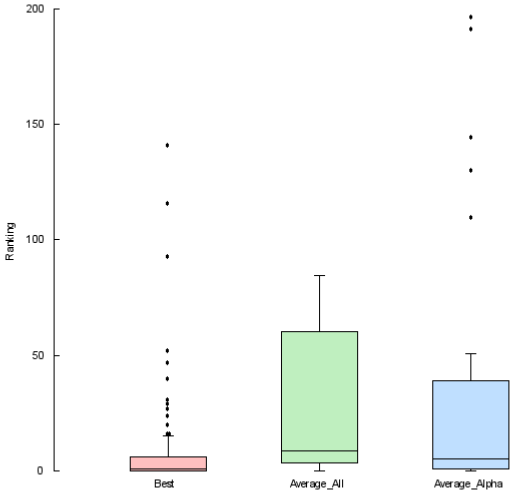


Figure 11: Ranking_All

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.

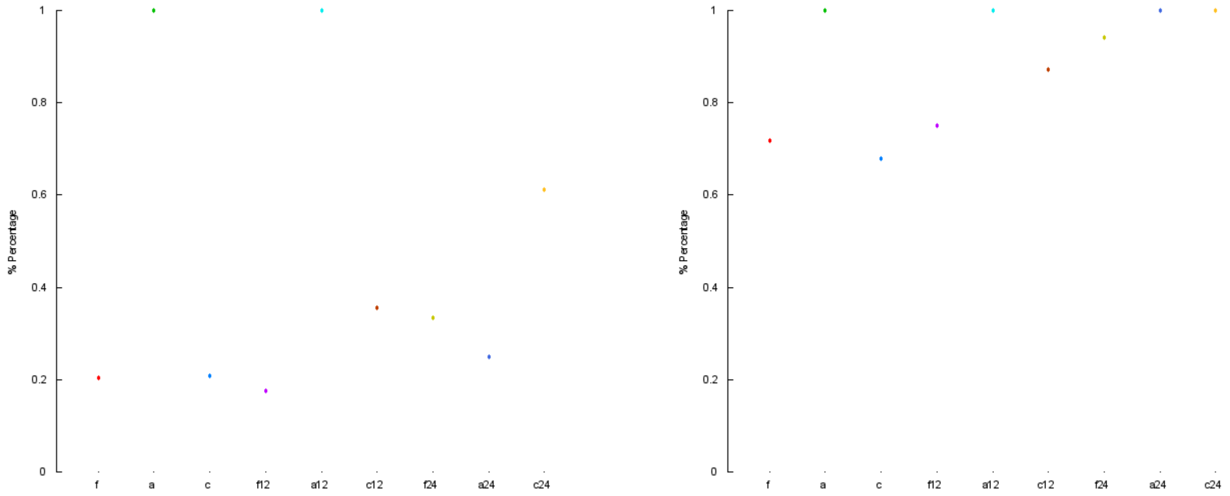
  - 5.45% of the runs contain all the possible pathways in the same graph.

- 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.



(a) ProteinG 1

(b) ProteinG 1 +2

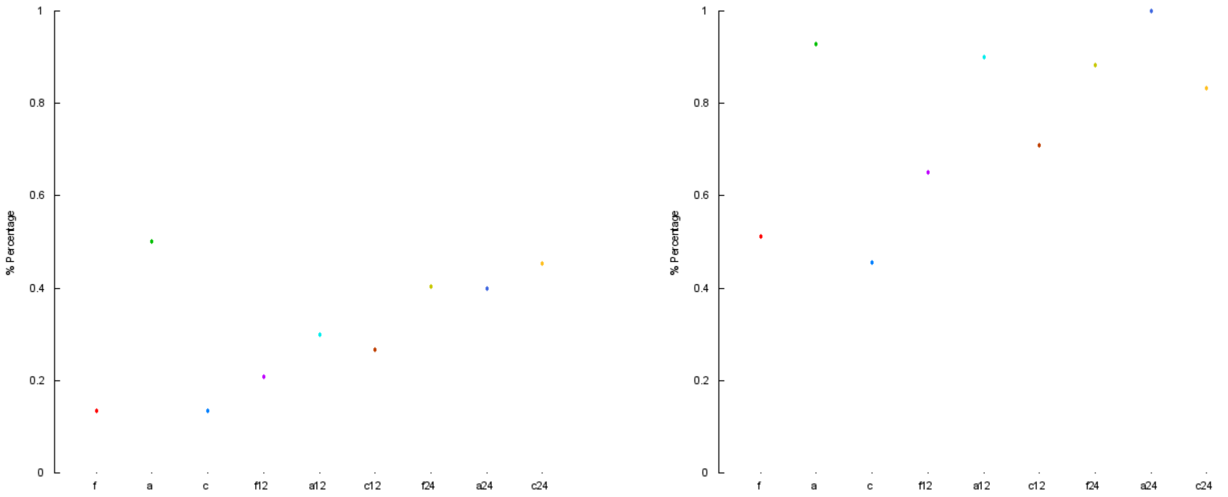
Figure 12: ProteinG 1

6.3 916 Benchmark

6.4 Pfam Benchmark

6.4.1 SH3 Domain

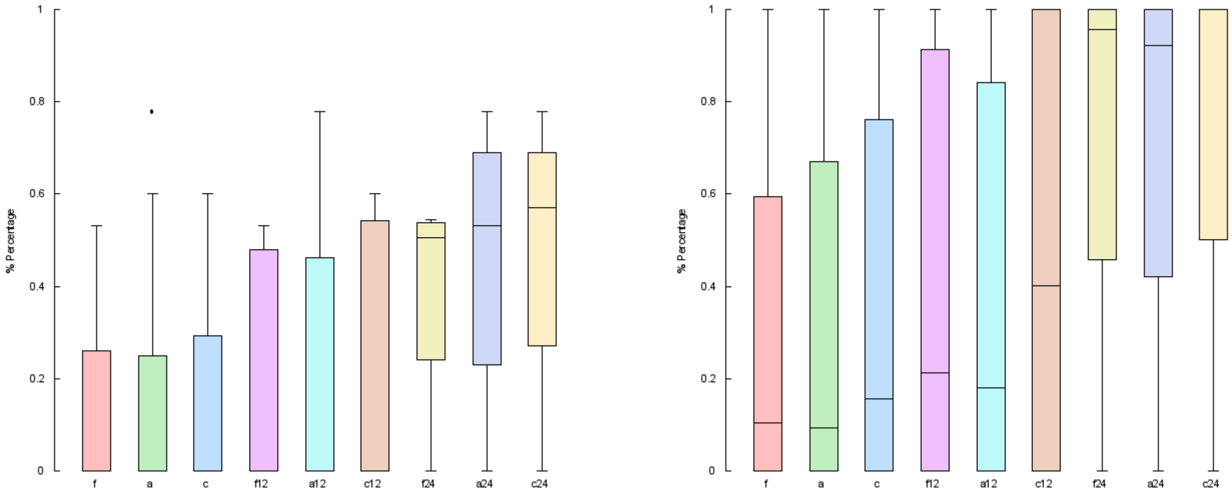
- Check the probability of all the intermediates.



(a) ProteinG 2

(b) ProteinG 2 +2

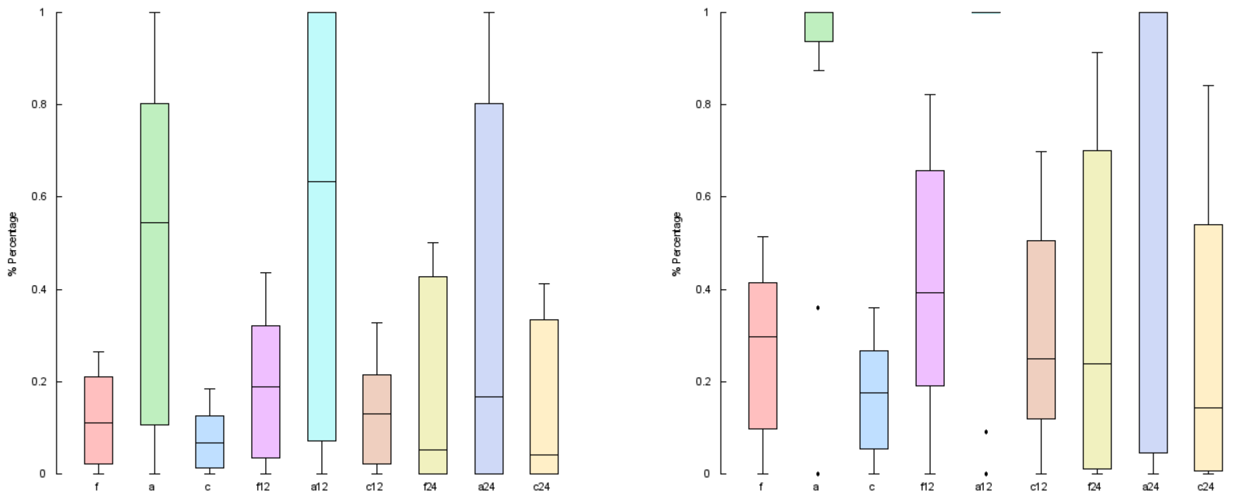
Figure 13: ProteinG 2



(a) Set2 1

(b) Set2 1 +2

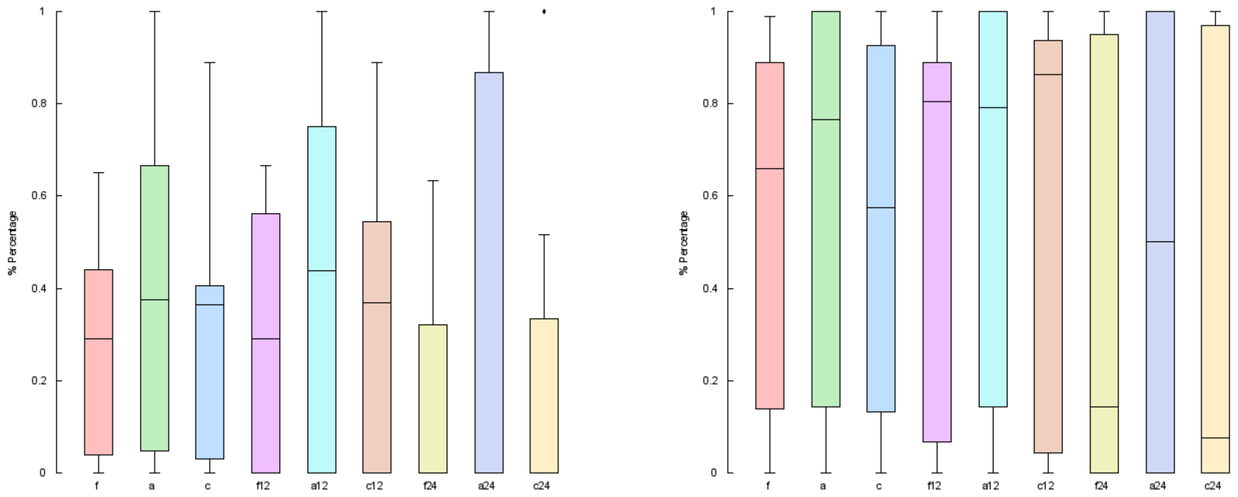
Figure 14: Set2 1



(a) Set2 2

(b) Set2 2 +2

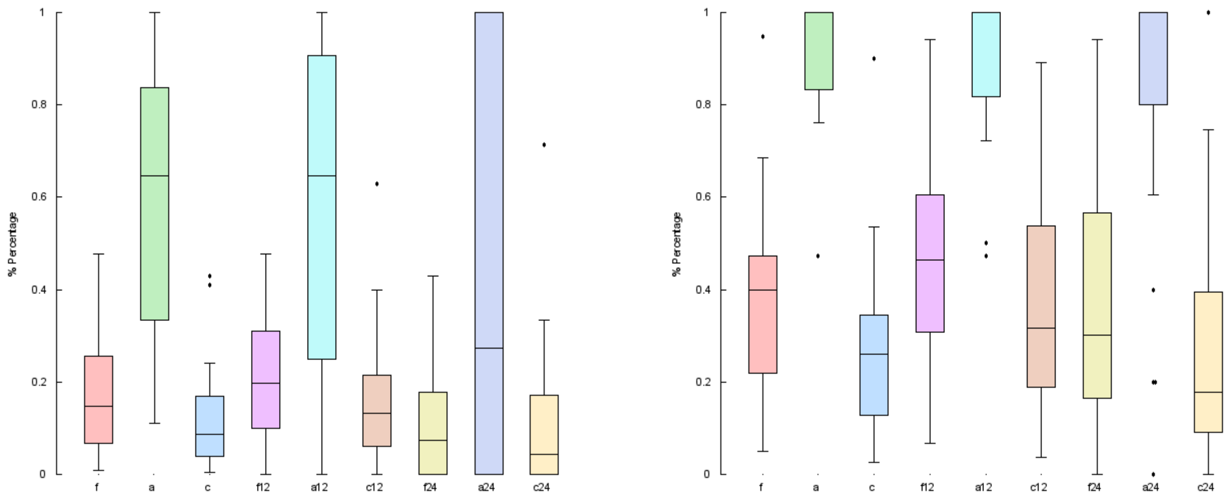
Figure 15: Set2 2



(a) Set3 1

(b) Set3 1 +2

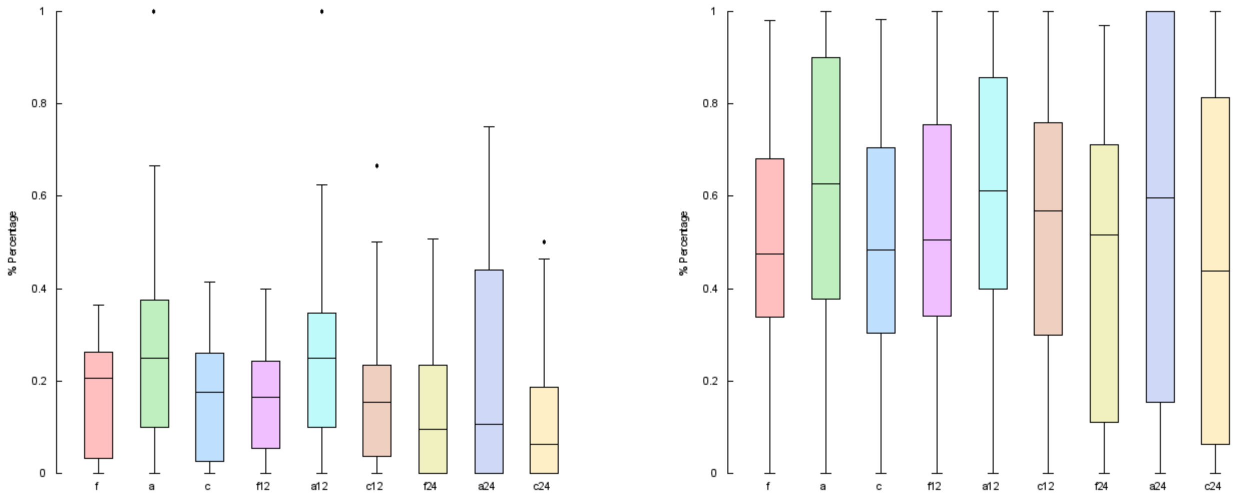
Figure 16: Set3 1



(a) Set3 2

(b) Set3 2 +2

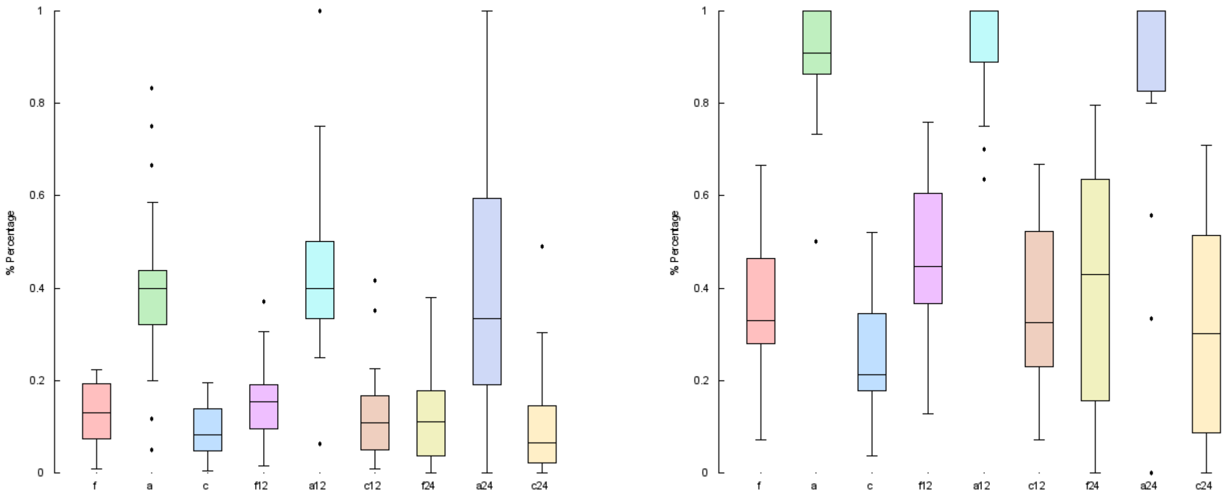
Figure 17: Set3 2



(a) Set4 1

(b) Set4 1 +2

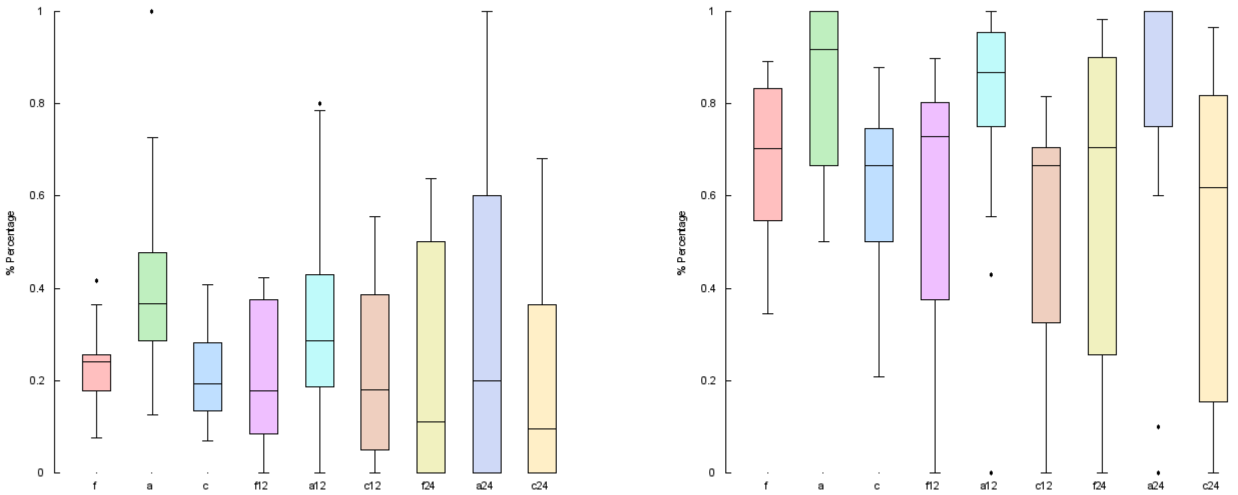
Figure 18: Set4 1



(a) Set4 2

(b) Set4 2 +2

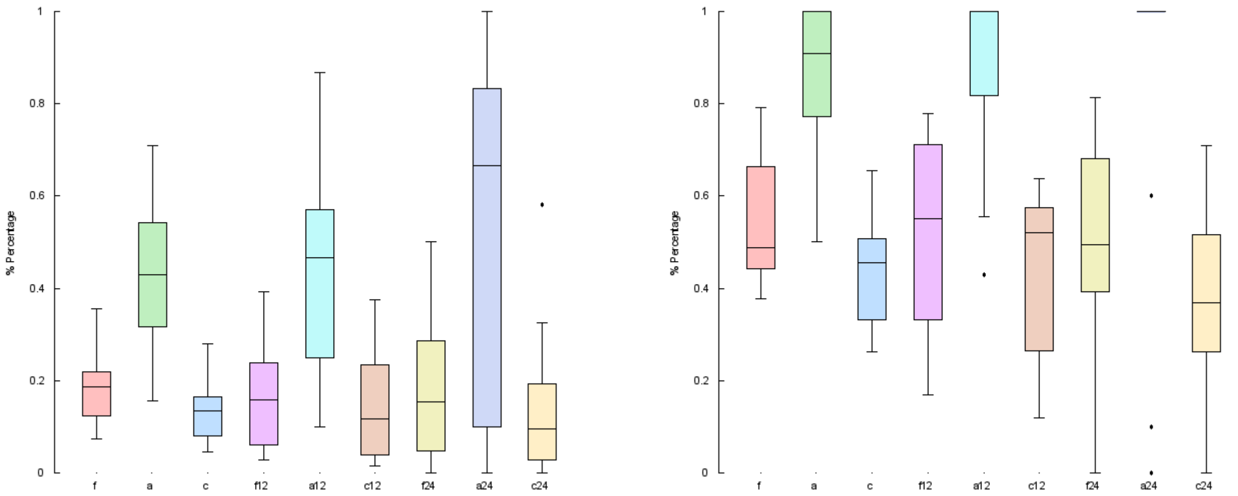
Figure 19: Set4 2



(a) Set5 1

(b) Set5 1 +2

Figure 20: Set5 1

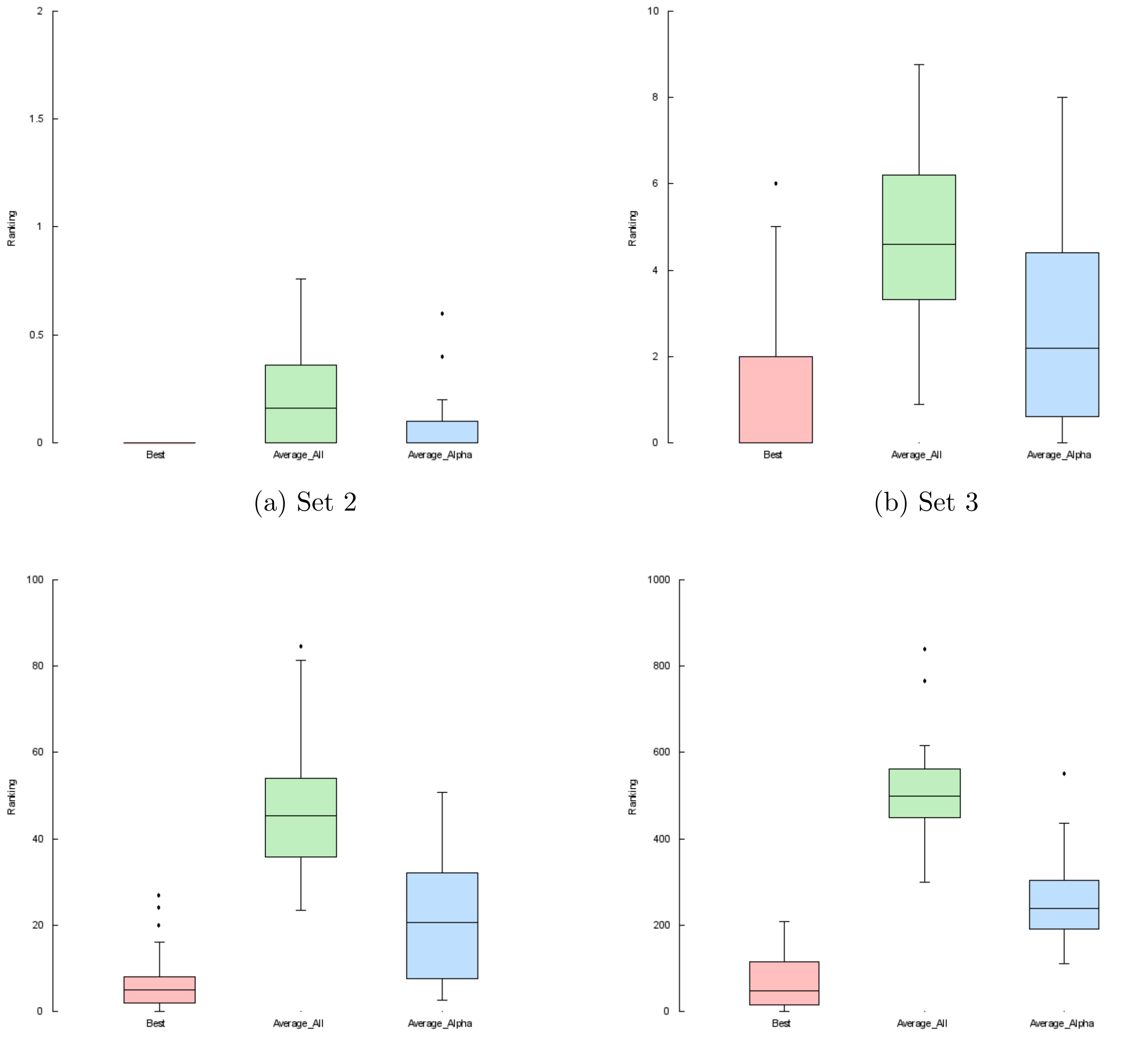


(a) Set5 2

(b) Set5 2 +2

Figure 21: Set5 2

- 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 ⇒ 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 ⇒ 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 ⇒ 1A2A3A4 (interaction of the intermediate intermediates before aggregation). 68.33% of the runs contain the intermediate pathway 1A2A3⇒ 5A1N2A3A4 (interaction needed in the hypothetical state where both packed stranded b-sheets are



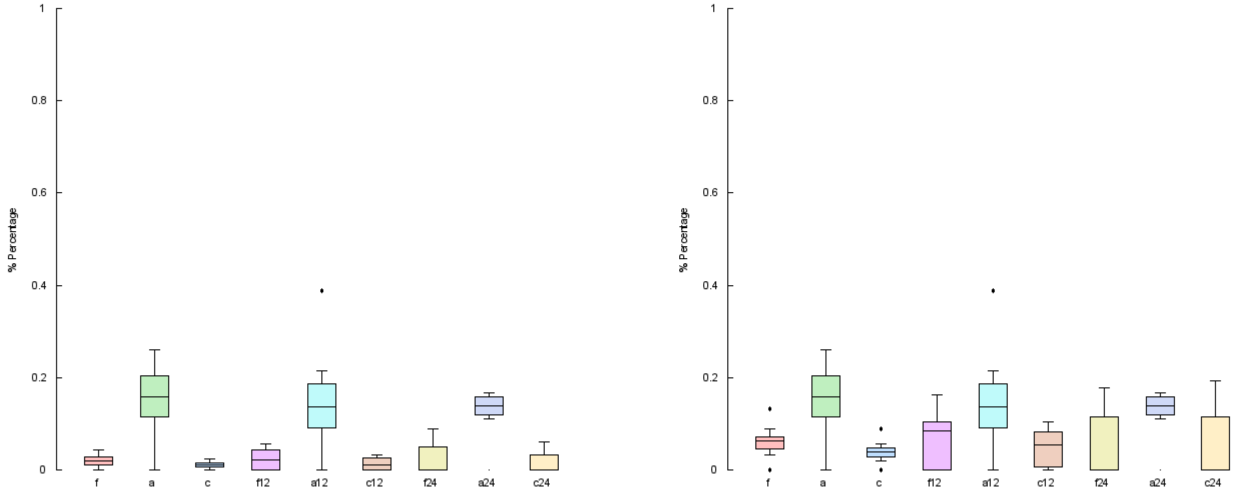
(a) Set 2

(b) Set 3

(c) Set 4

(d) Set 5

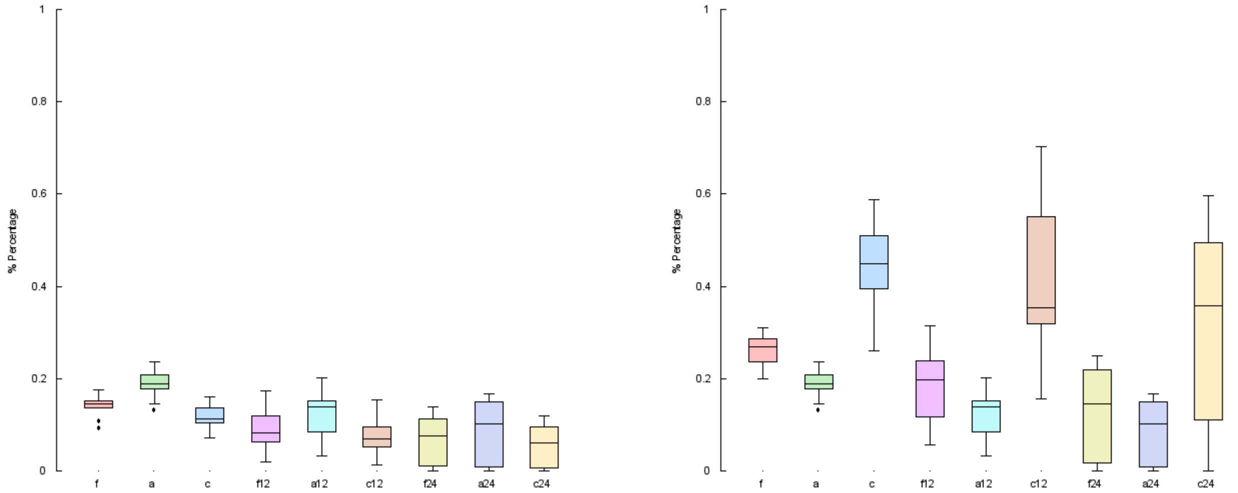
Figure 22: Set 916 - Ranking



(a) Set2 1

(b) Set2 1 +2

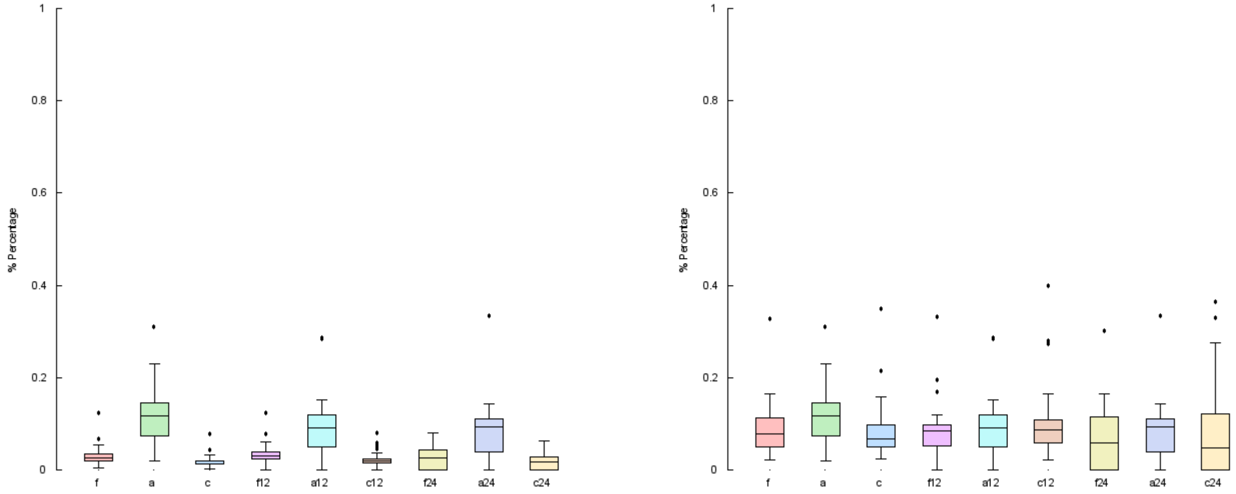
Figure 23: EvFold Set2 1



(a) Set2 2

(b) Set2 2 +2

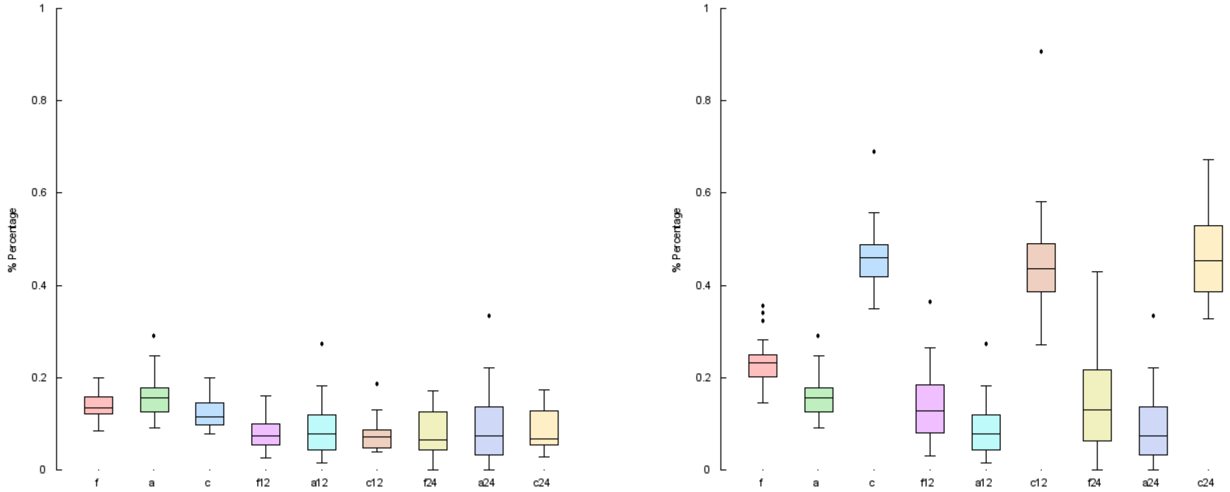
Figure 24: EvFold Set2 2



(a) Set3 1

(b) Set3 1 +2

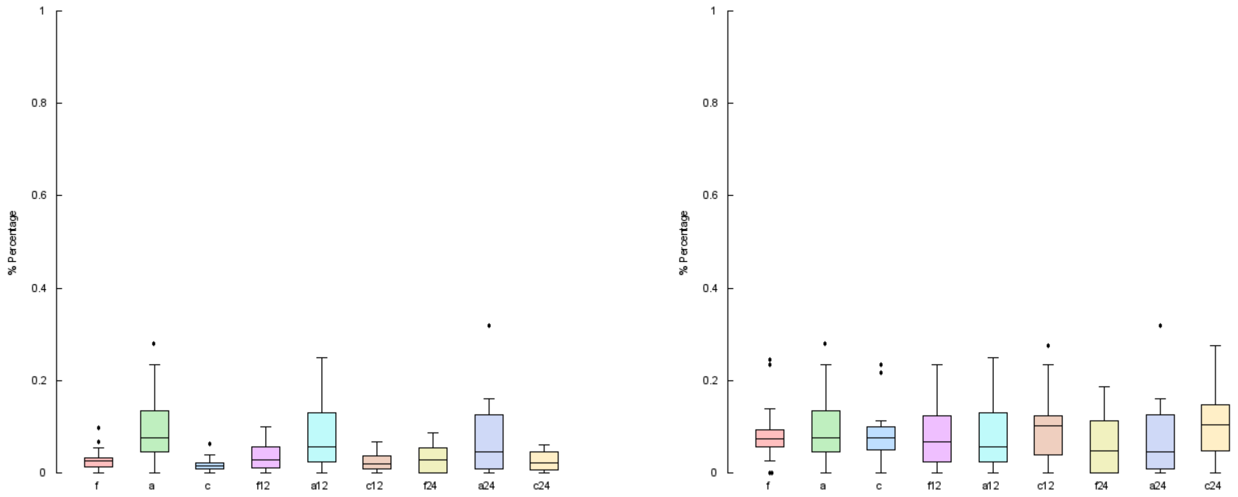
Figure 25: EvFold Set3 1



(a) Set3 2

(b) Set3 2 +2

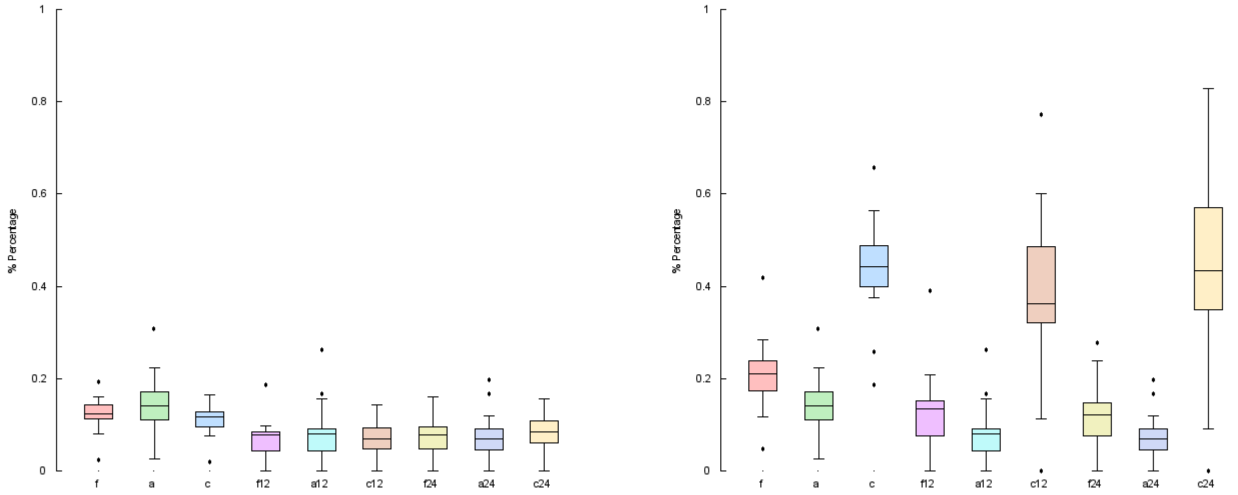
Figure 26: EvFold Set3 2



(a) Set4 1

(b) Set4 1 +2

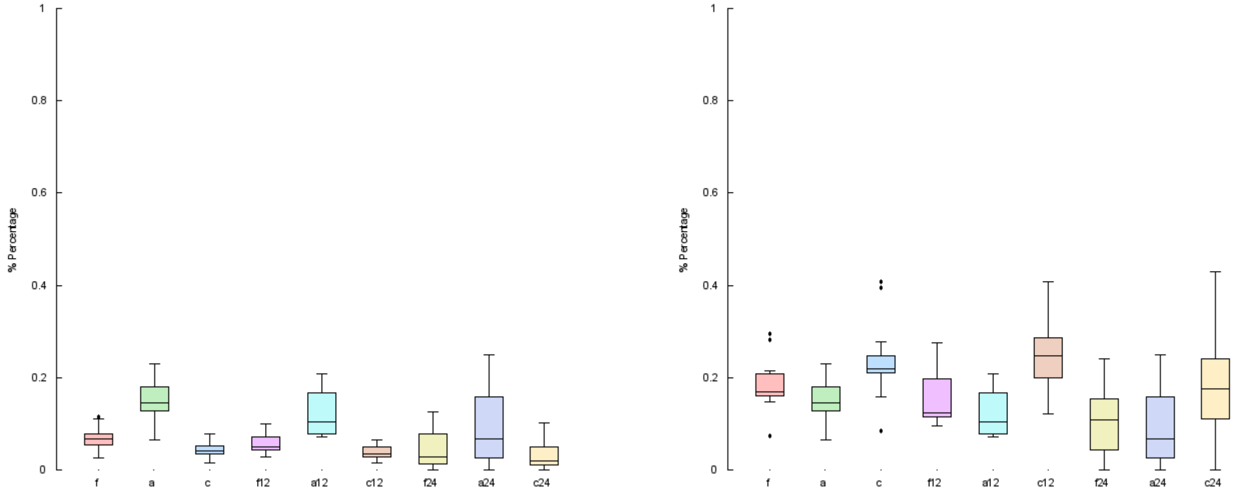
Figure 27: EvFold Set4 1



(a) Set4 2

(b) Set4 2 +2

Figure 28: EvFold Set4 2



(a) Set5 1

(b) Set5 1 +2

Figure 29: EvFold Set5 1

folded independently). 83% of the runs contain the intermediate pathway 1A2A3A4 ⇒ 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 ⇒1A2A3⇒1A2A3A4⇒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.

6.4.2 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.

6.4.3 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.

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] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

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

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

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.