Report eFold

Dr. Jérôme Waldispühl and David Becerra

February 3, 2015

Contents

1 Abstract

The protein-folding (PF) problem is interested in determining a protein tertiary structure from its amino acid sequence trying to understand the path that leads the folding process. Understanding the rules that govern the folding of proteins is one of the goals of biophysical studies that are still far from being achieved. High-resolution protein folding dynamics predictions are prohibitive expensive and they are typically produced through custom designed supercomputers and time-consuming molecular dynamics (MD) simulations. For the first time in the literature, we propose a novel methodology (called efold) that unifies the ensemble modeling and the evolutionary based sequence information framework to introduce an efficient (i.e., how well it performs computationally speaking) and effective (i.e., how good its solutions are) ab-initio protein folding method that reflect the ability of proteins to adopt different conformational states in vivo. The proposed method is tested on a benchmark of 125 proteins obtaining excellent results in terms of contact, strand, topology and pathway prediction. Efold represents a plausible advance in the PF state of the art through the modeling of the dynamics of protein folding, instead of focusing solely on the native conformation.

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 [24] 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.

OJO =¿ AGREGAR UN PARRAFO CON EL SIMIL DEL PUENTE Y LA RELACION UNFOLDED =¿ PATHWAY =¿ STRUCTURE. DICIENDO QUE EL DE NOSOTROS ES OTRO PARADIGMA, DAR RAZONES HISTORICAS.

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 $\beta1-\beta4$ sheet (I_3). The work [34, 35] reported an early formation of the second hairpin ($\beta3- \ turn-\beta4$) 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 $\beta1-\beta4$ 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 is 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 iso-structural 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 guaranteed 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 contain 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 experiments 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 ($\beta1$, $\beta2$ and $\beta3$), followed by the C terminal helix ($\alpha2$), the N terminal helix ($\alpha1$). The initial formation of the 30-51 disulfide can be in agreement with the early formation of the $\beta1\beta2\beta3$ sheet and its association with $\alpha2$. With the joining of $\alpha1$ 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 $\beta1$, $\beta2$, $\beta3$ strands, and the Sm2 sequence motif corresponds to the $\beta4$ and $\beta5$ 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 an energy landscape and a folding funnel model replace the idea of a single folding pathway.

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-grained 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.

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.

eFold

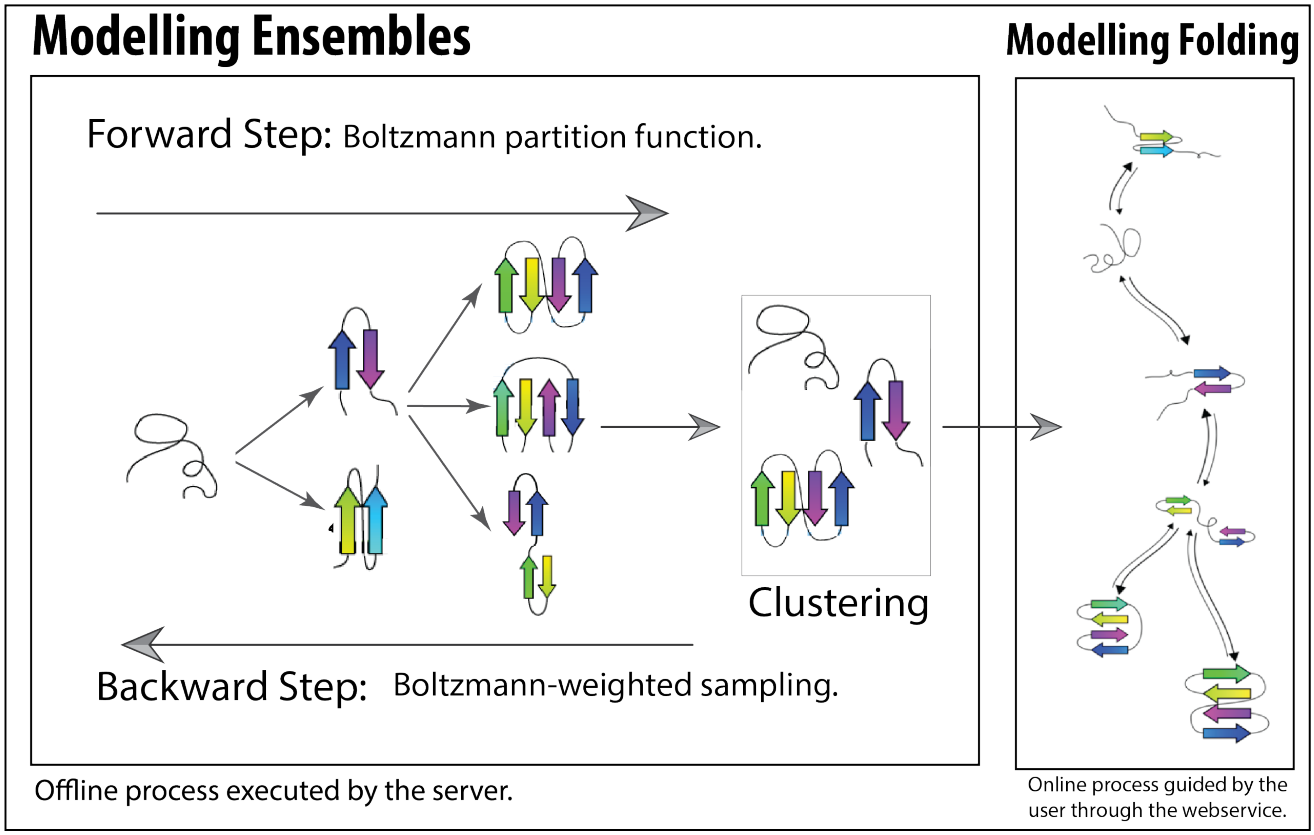

Figure 1: **efold:** is the proposed algorithm for predicting protein folding pathways and topologies using ensemble modelling and genomic variation. The algorithm is divided in two main phases, the modelling of ensembles and the modelling of the predicted folding dynamics. The first phase is computed 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.

Computation of the Partition Function:

Conceptually, each protein structure was described by a coarse-grained 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.

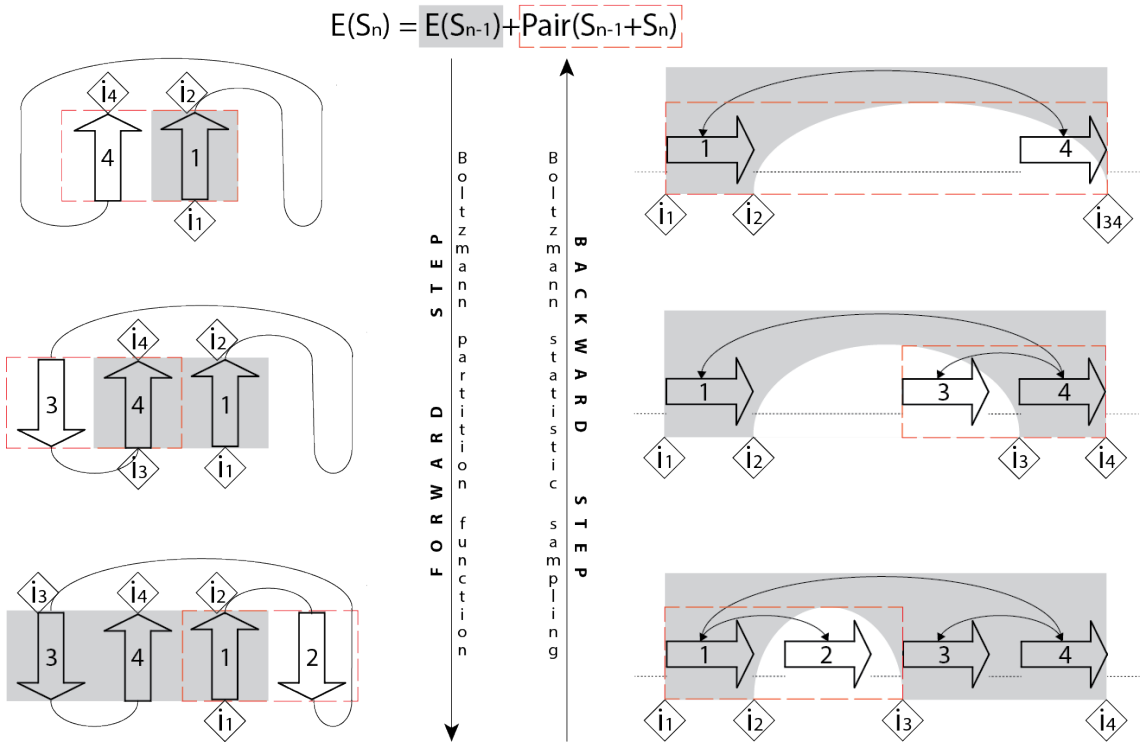

Figure 2: `efold` **Topologies:** *3A4P1A2* is the protein topology for the proteinG

The used evolutionary contact prediction method [16], called `EVfold`, 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 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 secondary structures is restrictive. Then, during the backward step, we compute a statistically representative sample of secondary structures. Additionally, clusters of these secondary structures 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 tries 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 a 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 this section, an experimental framework to report the results obtained using the proposed approach is presented. Specifically, an analysis of the results obtained over the set of benchmark proteins is provided. It is important to stress that when comparing our results with experimental and computational experiments, the numeration used for the strands follows the numeration proposed in our methodology.

In order to understand the performance of the proposed method, two 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 prediction of residue-residue contacts has been proved useful in reconstructing protein backbones by providing information to determine accurate 3D protein structures. Generally, the predictors of residue-residue proximity in folded structures (such as `EVfold`) are based on the existence of interdependent changes in groups of variable amino acids belonging to a protein family of homologues. Even if `efold` is not an algorithm developed to predict residue-residue contacts, we evaluated the prediction capabilities of `efold` to recognize contacts involved in

secondary structures. Then, we tested the proposed algorithm using the complete protein benchmark and compared the performance of `efold` with the predictions performed by `EVfold` and by our previous algorithm `tfolder`. The results for the contacts prediction are discussed in the Section 5.1.

As previously stated, `efold` is an ensemble algorithm aimed at protein structure and pathways prediction. Then, we are interested in study its capability to predict contacts involved in secondary structures. The prediction of those contacts is studied following the same methodology than normal 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. The results for the strand prediction are discussed in the Section 5.1

`efold` represents the proteins by a coarse-grained residue-level representation using the set of residue/residue contacts that form hydrogen bonds between β-strand backbones. `efold` performs the generation of all the admissible β-sheet topologies and the computation of the Boltzmann partition function over those topologies. Then, `efold` ranks those topologies based on their energy states and it clusters the top low-energy states to predict the folding dynamics. The experiment reported in section 5.2 quantify the ability of `efold` to correctly rank the admissible β-sheet topologies. Particularly, for each protein in the benchmark, we are interested in quantify the position of the topology reported by the corresponding PDB file in the rank computed by `efold`.

To study the efficacy of our technique for predicting protein folding pathways, we studied the folding landscape of proteinG, Ubiquitin and SH3 domains, proteins for which their pathways have been elucidated through many experimental studies and/or MD simulations (see Material section for more details). A graph for a specific 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. The results of this pathway prediction procedure are discussed in section .

Understanding the rules that govern the folding of proteins is one of the goals of biophysical studies that is still far from being achieved. Given the the diversity of protein structures of the native and denatured states, and the differences in sequences and amino-acid compositions, the extraction of general rules from the protein folding process is difficult by studying the folding of individual proteins. Two different strategies have been shown suitable in biophysical studies to extract general rules from the protein folding process. The first approach works by comparing

the mechanisms of proteins sharing the same overall fold but different sequence, i.e., members of the same protein family. The second strategy study proteins with a high degree of sequence identity, but different 3D structure. The results of our experiments based on the first technique are reported in section . On the other hand, results based on the first strategy are reported in section .

Measure	Approach	$x > 0$		$x \geq 12$		$x \geq 24$	
		± 0	± 2	± 0	± 2	± 0	± 2
Precision	Ours	**20.75**	**71.76**	**17.65**	**75**	**33.33**	**94.12**
	tFolder	13.3	52.1	10.6	54.1	14.0	58.3
Recall	Ours	**100**	**100**	**100**	**100**	25	**100**
	tFolder	56.3	97.9	53.8	61.5	**37.5**	87.5
F-measure	Ours	20.75	**68**	**35.48**	**87.1**	**61.11**	**100**
	tFolder	**21.5**	**68**	18.1	69.2	20.3	70

Table 1: The performance of the proposed approach for contact prediction is evaluated based on the precision, recall and F-measure of experimentally observed contacts. The performance metrics are reported for contacts which are 0, 12 and 24 residues apart. The metrics are also studied when predicted contacts are within ± 2 residues of an observed contact. The best results obtained are shown in bold.

5.1 Contact and Strand Prediction

The first experiment quantify the ability of `efold` to predict protein topologies through the prediction of residue-residue contacts. 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 assess the effectiveness of the method. These metrics are evaluated for all types of contacts (short, medium, and long-range). Particularly, an evaluation is performed when predicted contacts are within exact or ± 2 residues of an observed contact, and are more than 0, 12 and 24 residues apart.

Figure 3 (column *Contact prediction*) reports, through box plots, the best results of the experiments for the complete set of proteins (row *Complete*) and the set divided by the number of strands in the tested proteins. It is important to notice that the proposed method predicts residue contacts with an excellent precision for ± 2 for all contact separations in the complete benchmark. The precision for exact prediction (i.e., columns ± 0) is also high and it averages around 50% for the short and medium contacts and around 33% for long-range contacts. This result is significant given that critical protein folding steps can involve both short range and

long-range β-sheet contacts and that the precision assessed state-of-the-art algorithms in contact prediction for proteins without homology-based templates averages around 20% [48]. The recall obtained by `efold` average around 10% and 30% for contacts ± 0 and ± 2, respectively. This results is low as expected given that `efold` does not aim the prediction of all residue contacts, but those involved in secondary structures.

Comparing the predicted residue contacts between `efold` and our previous model `tfolder` for the Protein G, it is worth noticed in Table 1 that `efold` shows a better predicted accuracy than the `tfolder` algorithm. Furthermore, the proposed method performed better than tFolder in all the ranges except for the exact observed contacts studied 24 residues apart. The proposed approach not only kept sensitive to the distance of contact separation, but it increased the precision and sensitivity of the ensemble method.

`efold` is also compared with `EVfold` to measure at which extent `efold` improve the residue contact predictions used as input in our method. Particularly, the 500 best ranked evolutionary inferred contacts (i.e., EICs) computed by `EVfold` are used as input in `efold` (see section Methodology for more details). Figure 4 (column *Contact prediction*) reports, through box plots, the results of the experiments for the complete set of proteins. The recall of `EVfold` for ± 2 averages around 40% for all the range of contacts. Then, `EVfold` gets a better coverage than `efold` given that, unlike `efold`, `EVfold` does not focus only in contacts involved in secondary structures. Regarding the precision of the `EVfold` results, its low performance can be explained by the dependence of `EVfold` on the depth of the target alignments (See supplementary Table 1 for a complete list of the size of MSAs). Figure 4 shows that the methodology implemented by `efold` improved the initial contact predictions performed by `EVfold`.

Figure 3 (column *Strand prediction*) reports, through box plots, the best results of the experiments for contact prediction of residue-residue contacts involved in β-sheet structures. The results are spliced in rows showing the complete set of proteins (row *Complete*) and the set divided by the number of strands in the tested proteins. It is important to notice that the F-measure values for ± 2 for all contact separations in the complete benchmark is higher than 60%. This result corresponds to a very good performance of the method in terms of its accuracy and sensitivity. Furthermore, it confirms `efold` as a very good predictor of contacts involved in secondary structures. Regarding the exact prediction (i.e., columns ± 0), the precision decreases for long range contacts. Particularly, it goes from a precision around 30% for the complete set of contacts to a precision around 20% in long range contacts. The best and worst performance of `efold` can be recognized in proteins with three and two strands, respectively. It is important to stress that in `efold` there is not a big difference between the precision and sensitivity values for strand predictions, given that `efold` focuses on predicting contacts involved only in secondary structures.

Figure 4 (column *Strand prediction*) reports, through box plots, the results of the experi-

ments performed by `EVfold` for the complete set of proteins. It is important to notice that these results are more homogeneous than the results obtained in the contact prediction experiment. Moreover, there is not a big difference in the behaviour of the precision and recall measures, as noted in the contact prediction column. There is not a big difference either between the contacts ± 0 and ± 2 for all the contact ranges. The average of all the evaluation measures for an exact prediction (± 0), for the strand prediction experiment, fall in the same range (i.e., below a 10%) than its contact prediction counterpart. Then, it gives the insight that most of the contacts predicted by `EVfold` are involved in secondary structures. On the other hand, there is a clear decrease of the recall of the strand prediction with respect to the contact prediction, suggesting that for (± 2) predictions, `EVfold` reports a greater quantity of contacts not involved in secondary structure than the exact predictions (± 0). The prediction values for `EVfold` are much lower than the values of `efold`, showing that the methodology implemented by `efold` improved the initial contact predictions performed by `EVfold` regarding contact and strand predictions.

5.2 Protein Topologies Prediction

Figure 3 (column *Rank*) reports, through box plots, the position occupied by the PDB topologies in the ranking computed by `efold`. Particularly, for each protein composed by L strands, the position with respect to the top percentage when considering all the topologies with L strands (column *All*) and the attainable topologies given a common parent with $L-1$ strands (column P) are computed. It is important to stress that the likelihood of a specific topology to be chosen by `efold` to create clusters of low-energy states (used to predict the folding dynamics) increases as the topology move to the head of the ranking-list.

Figure 3 (column *Rank*) shows that, in average, the proposed approach ranks the target topologies (i.e., the topologies reported in the PDB) in the top 2% for the categories *All* and P. The best performance is computed for proteins of two and five strands. On the other hand, proteins of three strands compute the worst performances. `efold` performs better when the ranking is computed for all the the admissible β-sheet topologies (column *All*), than when compared with a subset (column P). Figure 3 (column *Rank*) shows excellent discrimination power of `efold` to separate conformational transitions that are critical to folding from those transitions that could simply result from minor structural fluctuations. In other words, `efold` allows the sampling of accurate conformations and it also score accurately those models more favourably from other decoys.

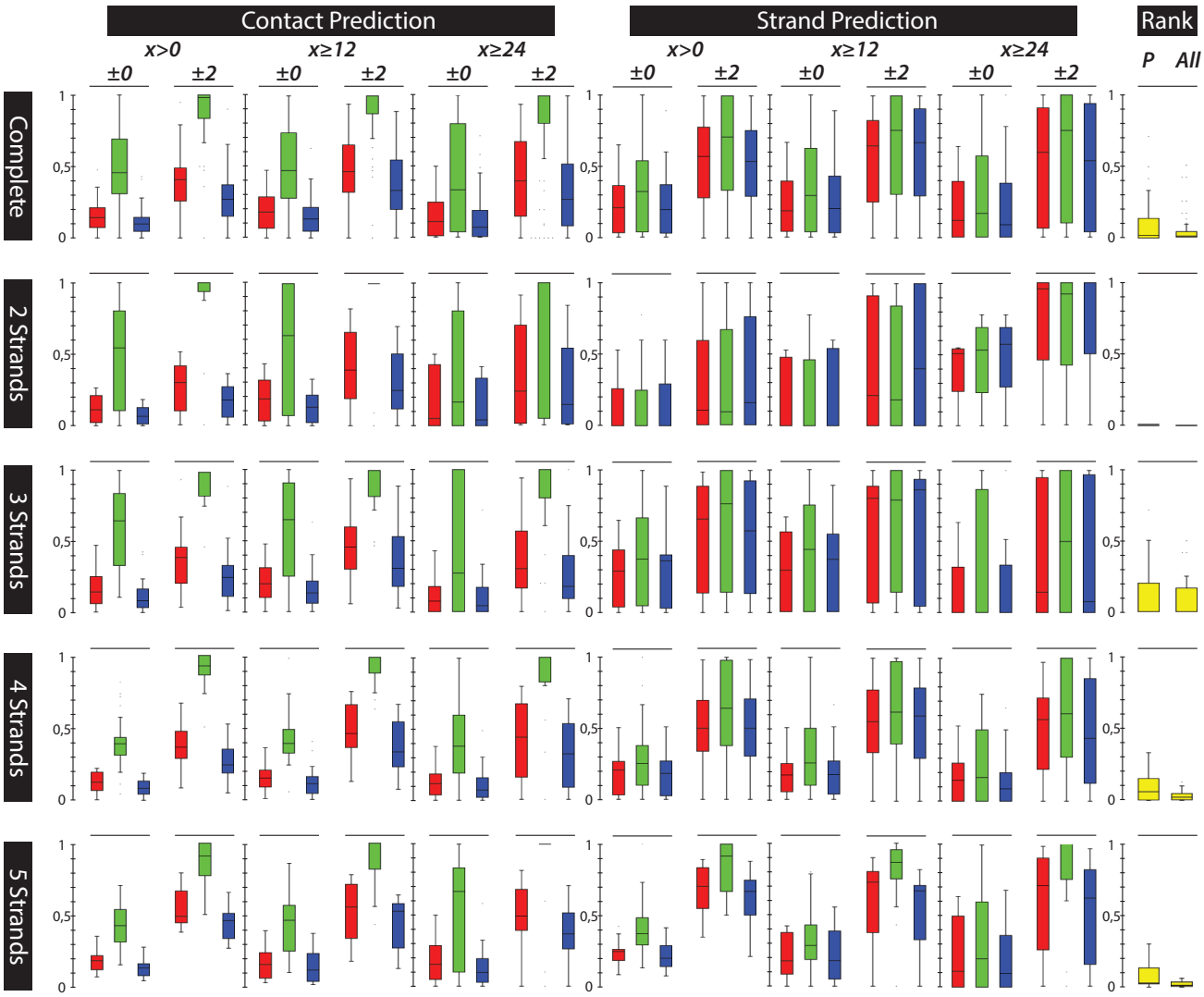

Figure 3: The performance of the proposed approach for contact prediction is evaluated based on the precision(green), recall(blue) and F-measure(red) of experimentally observed contacts. The performance metrics are reported for contacts which are 0, 12 and 24 residues apart. The metrics are also studied when predicted contacts are within ±2 residues of an observed contact.

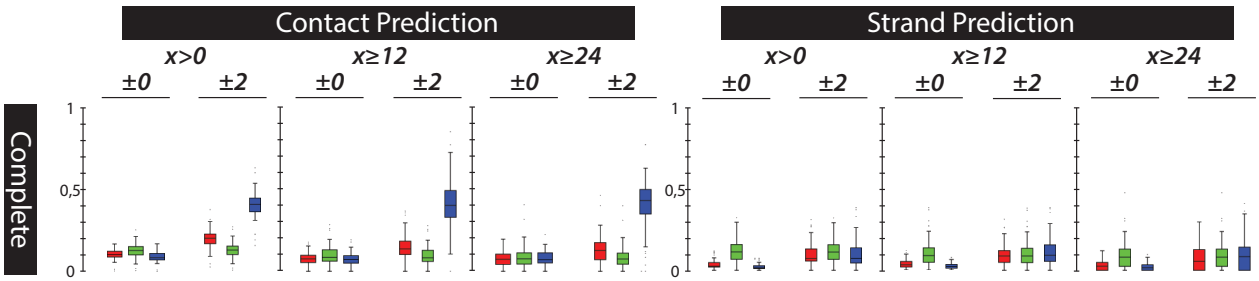

Figure 4: The performance of `EVfold` for contact prediction is evaluated based on the precision(green), recall(blue) and F-measure(red) of experimentally observed contacts. The performance metrics are reported for contacts which are 0, 12 and 24 residues apart. The metrics are also studied when predicted contacts are within ±2 residues of an observed contact.

5.3 Protein Pathway Prediction

To measure at what extent the predictions of `EVfold` influenced the `efold`'s predictions, runs with different influence of `EVfold` parameters where computed. Specifically, the parameter γ (see equation TBD) was set from a value of 0 (i.e., The `EVfold` predictions has not inference in the computation of the state's energy) to a value of 1 (i.e., The computation of the state's energy is based completely on the `EVfold` predictions) through adding 0.1 to the parameter γ. Five independent runs were computed for each experiment. Then, a total of 50 simulations were computed for each of the following experiments.

5.3.1 Protein G

Figure 5 reports a graph that represents the predicted folding transitions for protein G. The folding landscape and pathway of the peptide were reconstructed following the proposed methodology (see Section Methods). Inspection of this figure reveals that the folding intermediates are consisted with previous literature reports. Specifically, it is consistent with the work reported by ([34, 35]), with respect to the 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]) (See figure TBD). Our results also shows an early formation of the first hairpin ($\beta 1 - \ turn - \beta 2$)) and folding pathways were observed passing through the intermediate $\beta 1 - \ \beta 4$ sheet. This agrees with studies [49, 32] were it has been shown that proteinG folds through three pathways, all of which pass through an intermediate, to a single transition state. The three intermediates feature a near-native helix along with one go the three states determined by our simulations. Four folding pathways were observed in the simulations. Particularly, 80% of folded runs passed through a $helix-hairpin1$ complex (red pathway in Fig-

ure 5), 16% of the runs have a $helix - hairpin2$ complex and the 64% a $\beta1 - \beta4$ sheet complex. These pathways agrees with those found in [49]. Given the complexity of the chain topology, folding and intermediate states, we can expect a statistical distribution of topologically allowed pathways for assembling the tertiary fold as the ones found by our simulations.

Comparing the predicted folding pathways between `efold` and our previous model `tfolder` (See Supplementary Figure TBD), it is worth noticed that both models were able to predict the reported pathway for the tested protein; however, the probability of the experimental folding pathway is higher for `efold` than `tfolder`. The previous fact means that `efold` was more likely to correctly predict the observed folding pathway than its counterpart. Additionally, it is clear that `efold` was able to narrow the number of generated and predicted templates. There are two main factors that could help `efold` to got this performance with respect to `tfolder`. *i*) The penalization term used in `efold`, which penalizes the protein conformations that superimpose a $\beta - strand$ where an $\alpha - helix$ structure was predicted, allowed the model to narrow the search around conformational transitions that are critical to folding. *ii*) The hierarchical assembly folding model used in `efold` could represent a step-wise mechanism to narrow the conformation search through the correct prediction of the initial stages of protein folding process.

5.4 Ubiquitin

Figure 6 reports a graph representing the predicted folding transitions for ubiquitin protein. Inspection of this figure reveals that our simulations are in agreement with a view of ubiquitin folding suggested from previous experimental results and in-silico simulations. Particularly, our simulations give a special importance to a topology encompassing the $\beta1$ and $\beta2$ strand. This topology has been described to participate in a polarized and well-defined transition state ensemble [50, 51, 52]. Regions of the local $\beta1$-$\beta2$ hairpin populate native geometries, then this secondary structure can be stabilized by tertiary interactions between the α-helix and the $\beta1$-$\beta2$ sheet.

Computational analysis of experimental ϕ-values suggested that strands $\beta3$ and $\beta5$ might be formed adopting a native-like topology in the TSE [53]. Furthermore, ψ analysis suggested that ubiquitin folds through a much more organized TS ensemble with a common nucleus consisting of a partially formed four-stranded sheet network ($\beta1 - \beta2 - \beta5 - \beta3$). Then, in the ubiquitin folding pathway, regions of the local $\beta1$-$\beta2$ hairpin populate native geometries. Next, strand $\beta5$ joins the nascent hairpin-helix nucleus. Finally, the more distal strand $\beta3$ is joined to the core structure citekrantz2004discerning. Our simulations show an important role of the topology $\beta3 - \beta5$ when combining with the topology $\beta1$-$\beta2$. Particularly, it can be noted that the topology ($\beta2 - \beta1 - \beta5 - \beta3$) is the most populated edge in the graph. This topology receives the flow containing most of the in-coming paths. Furthermore,

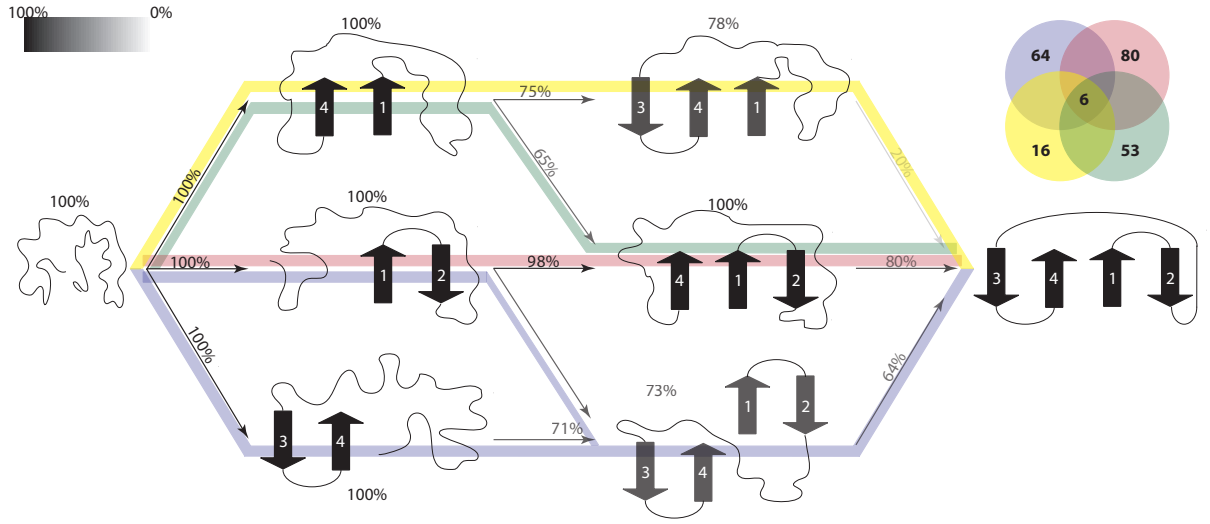

Figure 5: Pathways 1EM7: The figure represents the predicted transitions from a random coil to the native state of proteinG as a path in a graph of varyingly folded protein conformation states. The nodes in the graph represent energetically accessible conformation states which have been previously generated by the proposed Boltzmann-weighted ensemble sampling method. The percentage at top of each vertex (and its corresponding transparency) count for the number of times that this topology is predicted as a transition over the total number of runs. The vertices in the graph represent the transition between two topologies. The topologies connected by an edge are compatible topologies with structural similarity. The percentage at top of each edge (and its corresponding transparency) count for the number of times that this transition is found over the total number of runs. The predicted folding pathways are presented as paths in the graph. The percentage of presence of each path is presented through a Venn diagram at the right top corner of the figure.

this topology represents also the edge with more connections with the edge representing the native topology (i.e., ($\beta 4 - \beta 3 - \beta 5 - \beta 1 - \beta 2$)). The different paths coming into the topology $\beta 2 - \beta 1 - \beta 5 - \beta 3$ can represent the heterogeneity predicted for the transition state structure. The routes that stem from the core nucleus represent the heterogeneity of the transition state. The TS ensemble can contains subpopulations with additional structure formation.Particularly, this topology contains all the obligatory elements, but the nucleus of the TS structure can spread in different directions, adding more $\beta 1 - \beta 2$ or more $\beta 3 - \beta 5$ structures.

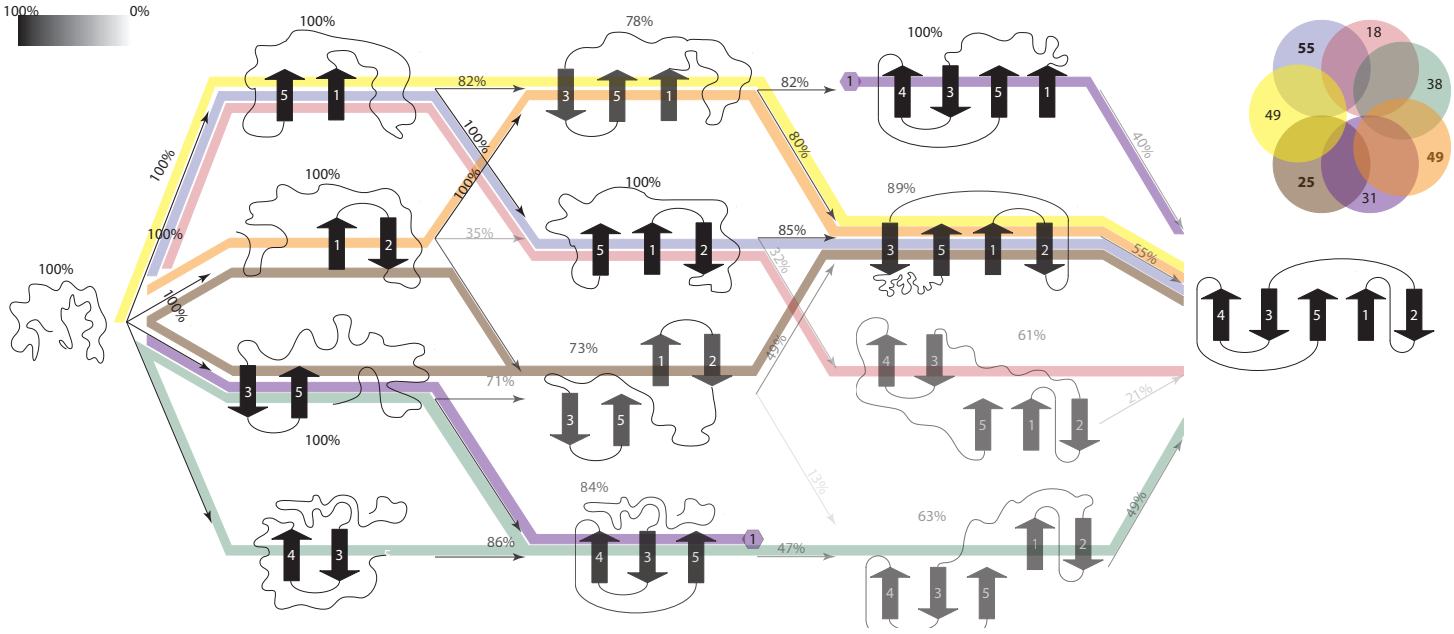

Figure 6: Pathways 1UBQ: The figure represents the predicted transitions from a random coil to the native state of ubiquitin protein as a path in a graph of varyingly folded protein conformation states. The nodes in the graph represent energetically accessible conformation states which have been previously generated by the proposed Boltzmann-weighted ensemble sampling method. The percentage at top of each vertex (and its corresponding transparency) count for the number of times that this topology is predicted as a transition over the total number of runs. The vertices in the graph represent the transition between two topologies. The topologies connected by an edge are compatible topologies with structural similarity. The percentage at top of each edge (and its corresponding transparency) count for the number of times that this transition is found over the total number of runs. The predicted folding pathways are presented as paths in the graph. The percentage of presence of each path is presented through a Venn diagram at the right top corner of the figure.

5.5 SH3 domain

Figure 7 reports a graph that represents the predicted folding transitions for proteins having the SH3 domain. Specifically, the proteins 1OOT, 1I0C, 1NEG and 2HDA were used to simulate the transitions from a random coil to the native state of proteins having this domain. Through the analysis of this figure, it is clear that the pathway crossing the four stranded $\beta-$sheet $\beta1-\beta2-\beta3-\beta4$ (yellow path in Figure 7) was present in 80% of the simulations. Then, the studied SH3 domains agree in the formation of the $\beta1-\beta2$, $\beta2-\beta3$, and $\beta3-\beta4$ motifs. Previous findings [54] suggested that this topology constitutes a metastable folding intermediate. This intermediate has been shown to be highly aggregation prone, as it exposes Strand $\beta1$. Then, the formation of native strand $\beta5$ (i.e., last folding step in the path) is critical in preventing aggregation during folding, where the native state is protected from aggregation, whereas the intermediate is highly aggregation prone.

Three of the predicted pathways cross through the $strand2$ $strand3$ $strand4$ topology. Experimental results indicate that the second, third, and the fourth β-strands are the most ordered regions of the TSE. This structure is common to all the domains simulated in our experiments and it represents the central (and hydrophobic) sheet $\beta2$ $\beta3$ $\beta4$. From that topology, the paths branch in three different pathways (i.e., the red, purple and yellow paths in Figure 7). These path counts for the most probable paths and end the folding building the second sheet (a less structured topology) with two terminal strands ($strand1 and strand5$).

It is important to stress that this protein does not contain α-helix motifs, showing that the evolutionary inference contact is able to correctly model β-strands motifs in the absent of other secondary structures. Then, the use of evolutionary information is applied not only as a constraint factor, but as a complementary module to the ensemble modelling procedure.

Figure TBD shows how the probability of observing any of the reachable states changes over time. Given the expected low probability of high size permutations, these values have been normalized for each size of pairing interactions. Interestingly, there are many interactions of three pairings (i.e., four β-strands), which agrees with previous findings about that four stranded β-sheet constitutes a metastable folding intermediate ([54]). NOOOTTEEEE: THIS PARAGRAPH IS INCOMPLETE.

5.6 ProteinG mutants

An novel engineering approach allowed the obtention of set of proteins with high sequence identity but different structure and function [55, 56, 57]. Particularly, the sequences of two domains from streptococcal protein G were subjected to an iterative design of heteromorphic pairs. Then, two different wild-type protein domains, called G_A and G_B showing an increasing degree of sequence identity (starting from 1% to 95%) have been created. G_A displays a

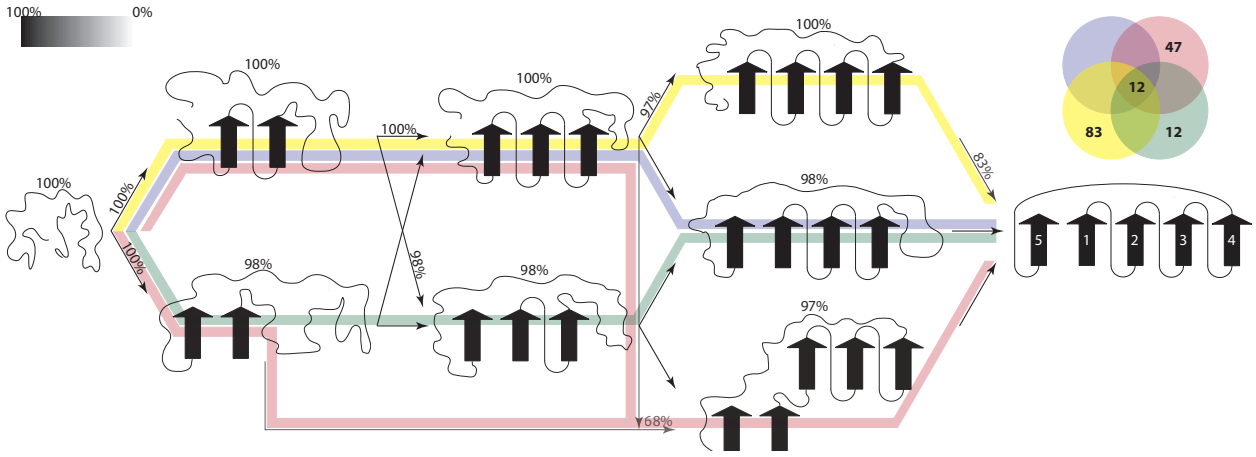

Figure 7: Pathways proteins 1OOT, 1I0C, 1NEG, 2HDA: The figure represents the predicted transitions from a random coil to the native state of proteins, containing the SH3 domain, as a path in a graph of varyingly folded protein conformation states. The nodes in the graph represent energetically accessible conformation states which have been previously generated by the proposed Boltzmann-weighted ensemble sampling method. The percentage at top of each vertex (and its corresponding transparency) count for the number of times that this topology is predicted as a transition over the total number of runs. The vertices in the graph represent the transition between two topologies. The topologies connected by an edge are compatible topologies with structural similarity. The percentage at top of each edge (and its corresponding transparency) count for the number of times that this transition is found over the total number of runs. The predicted folding pathways are presented as paths in the graph. The percentage of presence of each path is presented through a Venn diagram at the right top corner of the figure.

three-helix bundle fold, and G_B displays a $\alpha + \beta$ proteinG fold. Table 2 shows the percentage identity between the different variants of the wild-type proteins and proteinG. These set of proteins represent a great opportunity to elucidate relationships and dependency between `efold`, sequence information and folding mechanisms.

Protein	1EM7	G_{B95}	G_{A95}	G_{B88}	G_{A88}	G_{B77}	G_{A77}	G_{B30}	G_{A30}	G_{B1}	G_{A1}
1EM7	100	63	58	65	54	72	50	83	17	88	13
G_{B95}		100	95	93	92	90	88	74	54	67	45
G_{A95}			100	92	97	84	93	68	59	61	50
G_{B88}				100	88	93	84	77	50	70	42
G_{A88}					100	81	97	65	63	58	54
G_{B77}						100	77	84	43	77	34
G_{A77}							100	61	67	54	58
G_{B30}								100	31	93	24
G_{A30}									100	24	92
G_{B1}										100	17
G_{A1}											100

Table 2: Sequence identity between the different variants identified as G_{A30}, G_{A77}, G_{A88}, and G_{A95} for the GA fold, and G_{B30}, G_{B77}, G_{B88}, and G_{B95} for the GB fold, and the wild type proteins G_{A1} and G_{B1} and the proteinG $1EM7$

Figure 9 reports a graph that represents the predicted folding transitions for the variants and wild type proteins for the GB folds. Special interest is given to the GB folds given than `efold` is an algorithm for modelling the folding process of large β-sheet proteins, such as the $\alpha + \beta$ ubiquitin-like fold reported by the GB folds. From figure 9 it is clear that four main set of pathways can be obtained from the simulations. Specifically, the first group is constituted by the variants G_{B88} and G_{B95}, the variants G_{B77}, G_{B33} and G_{B1} composed the second, third and fourth groups, respectively. Comparing those sets, it is important to stress that the blue path is the most probable path in all the sets. These pathways correspond to the early formation of the second hairpin ($\beta 3 - \ turn - \beta 4$) and the first hairpin ($\beta 1 - \ turn - \beta 2$) followed by the formation of the $\beta 1 - \ \beta 4$ sheet. The red pathway is also highly present in our simulations. This pathway favours the early formation of the first hairpin followed by the formation of the sheet and second hairpin. These behaviours agree with the folding pathways attributed to a ProteinG like fold. Beside the common folding paths, our simulations exhibit distinct folding routes for the protein G variants. Particularly, the variants G_{B95} and G_B exhibit an path based on an early formation of the second hairpin followed by a $\beta 1 - \ \beta 4$ sheet formation (orange path in Figure 9) that is not present in the other variants. Additionally. the wild type G_{B95} does not show the path involving an early formation of the $\beta 1 - \ \beta 4$ sheet. The previous simulations agree with the evidence that proteinG variants exhibit distinct folding routes, where the main

difference between them is a different order of formation of the β-strands [58].

This experiment is also useful to analyze the sensitivity of `efold` to changes in the amino-acid sequence. In particular, it is clear that `efold` is sensitive enough to preserve the prediction of pathways for proteins that diverge in amino-acid sequence, but that contains similar folding pathways. In a similar way, `efold` was able to diverge the prediction of folding pathways that present a high sequence identity, but different folding pathways. Particularly, when comparing the ranking (as explained in the Section 5.2) between the variants of GA folds versus the variants of GB folds, the predictions of the GB folds, as a ProteinG fold-like, outperform the predictions of the GA counterparts. In other words, `efold` was able to predict with high accuracy the GB folds as containing a ProteinG fold-like topology; meanwhile `efold` predicted the GA folds as topologies with a low probability to belong to a ProteinG topology. It is important to stress that this divergent predictions were obtained on proteins with a high sequence identity, but different 3D structures.

TOCA HACER EL ANALISIS DE NUCLEI, ESTE PARRAFO SE PUEDE FINALIZAR EL PARRAFO DICIENDO QUE DEPENDE NO DEL NUMERO DE AMINOACIDOS QUE CAMBIEMOS, SINO CUALES SON LOS QUE CAMBIAMOS.

5.7 Pfam families

Experimental studies suggest that the native topology of a protein plays a key role in determining its folding pathway. These studies usually compare proteins that differ in sequence but share the same overall fold to identify relationships between sequence information and folding mechanism. As a result, the folding mechanisms of proteins belonging to a same protein family have been reported as conserved [59, 60]. In this section, we analyzed the results of comparing folding processes of different members of four protein families. Particularly, the Pfam families PF00014, PF00018, PF00240, and PF01423 are studied.

5.7.1 PF00014

Figure 10 reports a graph that represents the predicted folding transitions for proteins belonging to the Pfam family PF00014. Specifically, the proteins 1D0D, 1BUN, 1BIK and 5PTI were used to simulate the transitions from a random coil to the native state of proteins having this domain. Through the analysis of this figure, it is clear that there are two pathways that are present in all our simulations. The final topologies are represented by $1A2N3A4$ and $2A1A3$ arrangements. These topologies are attributed to the proteins 1BIK and 5PTI, respectively. The topology $1A2$ is traversed for all our simulations and it is present in all our simulations.

The folding pathways of disulfide proteins vary substantially [61]. Particularly, it has been shown that with two structurally homologous kunitz-type protease inhibitors, bovine pancreatic

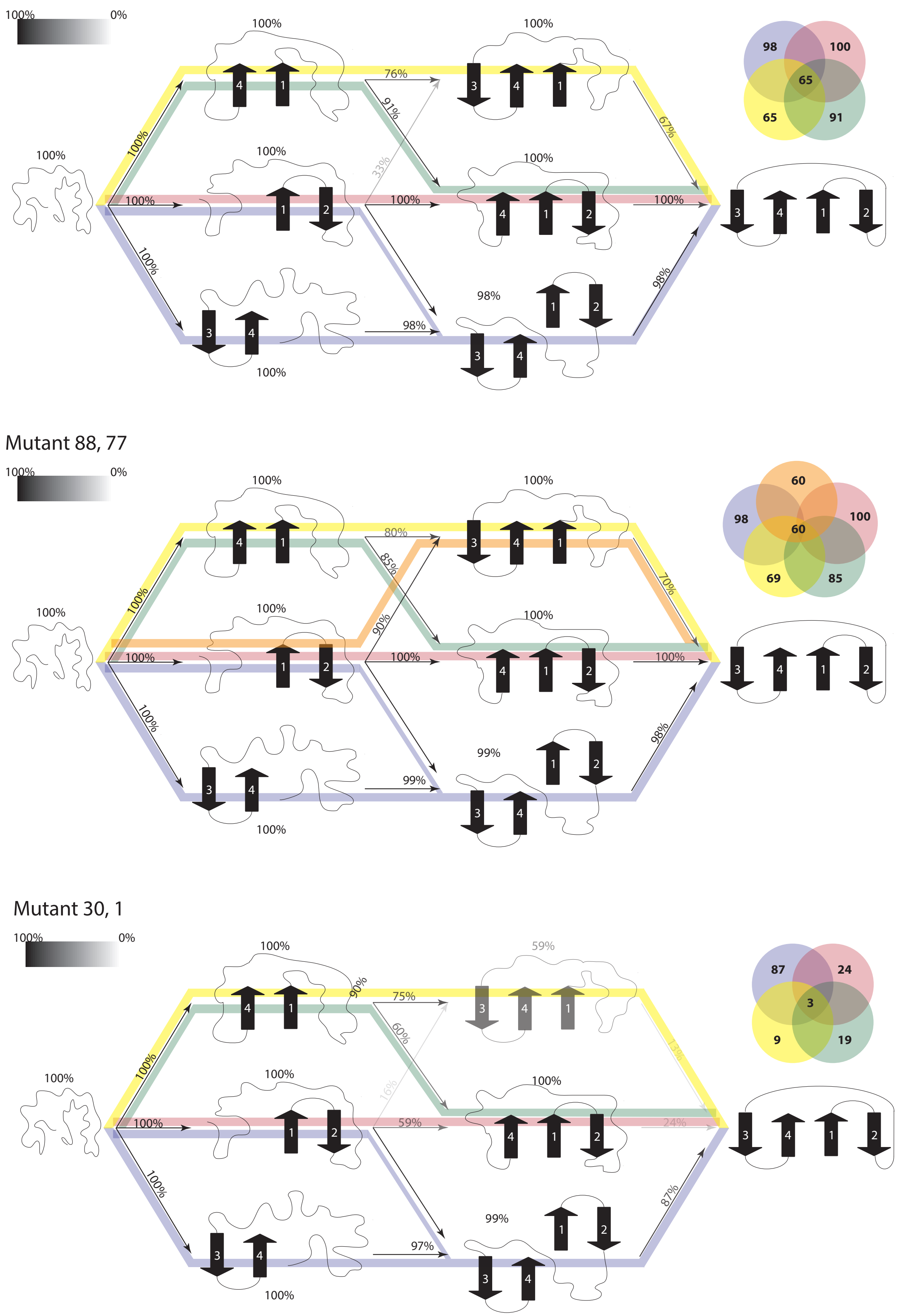

Figure 8: Pathways for the proteins G_{A30}, G_{A77}, G_{A88}, and G_{A95} for the GA fold: The nodes in the graph represent energetically accessible conformation states which have been previously generated by the proposed Boltzmann-weighted ensemble sampling method. The percentage

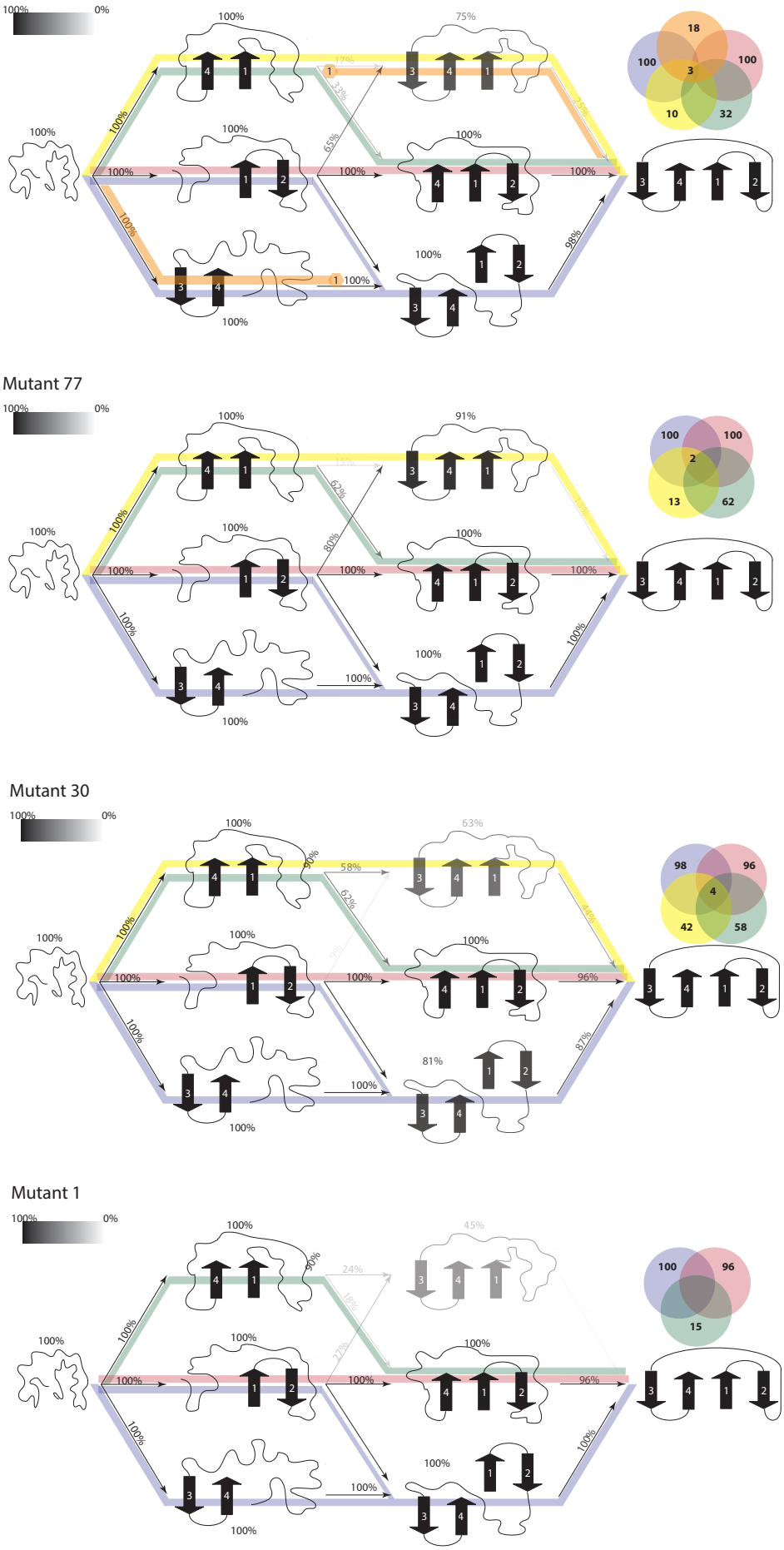

Figure 9: Pathways GB: Pathways for the proteins G_{B30}, G_{B77}, G_{B88}, and G_{B95} for the GB fold:

trypsin inhibitor and tick anticoagulant peptide, there is a heterogeneity of folding intermediates and folding kinetics [44]. The simulated proteins represent three different kunitz-type protease inhibitors. Specifically, two proteins (1D0D, 5PTI,) represent bovine pancreatic trypsin inhibitors (BPTI), 1BUN represent a serine protease inhibitor homolog beta-bungarotoxin B2 chain and 1BIK represents a AMBP protein. The proteins BPTI simulated in our experiments show a unique folding pathway with an early formation its secondary structures. This result agrees with experimental results that suggest for BPTI, the stable subdomain structures dictate the formation of native-like intermediates and limit the heterogeneity of folding intermediates [43].

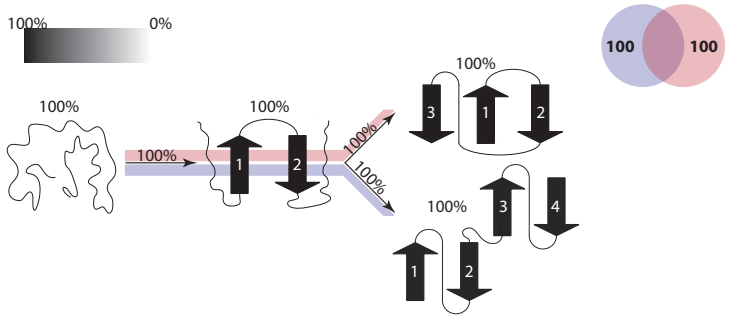

Figure 10: Pathways proteins 1D0D, 1BUN, 1BIK, 5PTI: The figure represents the predicted transitions from a random coil to the native state of proteins, belonging to the Pfam family PF00014, as a path in a graph of varyingly folded protein conformation states. The nodes in the graph represent energetically accessible conformation states which have been previously generated by the proposed Boltzmann-weighted ensemble sampling method. The percentage at top of each vertex (and its corresponding transparency) count for the number of times that this topology is predicted as a transition over the total number of runs. The vertices in the graph represent the transition between two topologies. The topologies connected by an edge are compatible topologies with structural similarity. The percentage at top of each edge (and its corresponding transparency) count for the number of times that this transition is found over the total number of runs. The predicted folding pathways are presented as paths in the graph. The percentage of presence of each path is presented through a Venn diagram at the right top corner of the figure.

5.7.2 PF00018

The results have been already analyzed in the section SH3 domain.

5.7.3 PF00240

Figure 11 reports a graph that represents the predicted folding transitions for proteins belonging to the Pfam family PF00240. Specifically, the proteins 1CMX, 1EUV and 1UBQ were used to simulate the transitions from a random coil to the native state of proteins having this domain. Through the analysis of this figure, it is clear that the topology $\beta 2 - \beta 1 - \beta 4 - \beta 3$ is the conserved topology for the family. This topology corresponds to the organized TS ensemble that consist of a four-stranded sheet network (i.e,. $\beta 2 - \beta 1 - \beta 5 - \beta 3$) in the folding pathway of the ubiquitin protein. The different paths that converge to the conserved structure can represent the different proportions of the four properly aligned strands that compose the β-sheet network. The two most populated paths (green (71%) and red (60%)) cross through the topology $\beta 2 - \beta 1 - \beta 4$, showing concordance with the ubiquitin folding suggested from experiments. The strand $\beta 4$ is present in all the paths that convert to the final topology. This strand is the central strand and a critical structural component in the ubiquitin topology.

The topology $\beta 2 - \beta 1 - \beta 4 - \beta 3$ is topographically related to the structure of ProteinG in that the order, positions and stretches of secondary structures are identical. This fact is important given that it has been shown that this fold is present in the ubiquitin family an in other proteins with biologically distinct functions, such as the (Ig)-binding protein G [62, 63]. Later, this common fold was termed β-grasp and it has been suggested to be a multi-functional scaffold in diverse biological contexts [64]. Studies has suggested that that proteins belonging to a same superfamily (i.e., with identical folds, but highly diverged sequences) retain identity at sequence positions that participate in the folding nucleus [65]. Monte Carlo simulations have identified nucleus positions that are conserved among structures with homologous folds for the ProteinG. Particularly, in [49], the identified nucleus identified residues in hairpin 1 (Y3 and L5), the helix (F30), and hairpin 2 (W43, Y45, and F52). All of these residues show low sequence entropy over aligned sequences in the ubiquitin superfamily [65]. With respect to our simulations, those same residues are reported to be involved in most of the predicted pathways, highlighting the importance that `efold` confers to these residues. Particularly, the aligned sequences in the ubiquitin superfamily (with respect to the protein G) reports the residues L5:K6, F30:V26, W43:L43, and Y45:F45 as having a frequency of 99.52%, 82.54%, 85.37% and 42.45% of presence in the reported pathways, respectively. 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.

5.7.4 PF01423

Figure 12 reports a graph that represents the predicted folding transitions for proteins belonging to the Pfam family PF01423. Specifically, the proteins 1KQ1, 1HK9 and 1H64 were used to

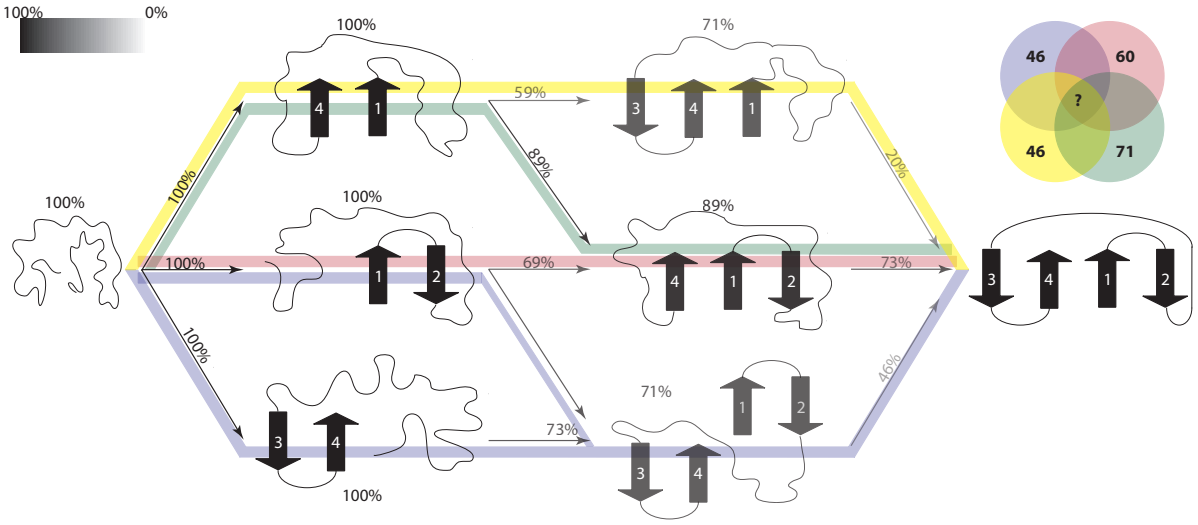

Figure 11: Pathways proteins 1CMX, 1EUV, 1UBQ: The figure represents the predicted transitions from a random coil to the native state of proteins, belonging to the Pfam family PF00240, as a path in a graph of varyingly folded protein conformation states. The nodes in the graph represent energetically accessible conformation states which have been previously generated by the proposed Boltzmann-weighted ensemble sampling method. The percentage at top of each vertex (and its corresponding transparency) count for the number of times that this topology is predicted as a transition over the total number of runs. The vertices in the graph represent the transition between two topologies. The topologies connected by an edge are compatible topologies with structural similarity. The percentage at top of each edge (and its corresponding transparency) count for the number of times that this transition is found over the total number of runs. The predicted folding pathways are presented as paths in the graph. The percentage of presence of each path is presented through a Venn diagram at the right top corner of the figure.

simulate the transitions from a random coil to the native state of proteins having this domain. From this figure, this is clear that the two most probable paths (i.e., the yellow path with a percentage of 77% and the blue path with a presence of 27%) correspond to an early formation of the $\beta 1$, $\beta 2$ and $\beta 3$ strands, followed by the formation of the $\beta 4$ and $\beta 5$ strands. These two sets of strands correspond with two sequence motifs (32 and 14 amino acids long, respectively) that have been identified between various LSm homologs.

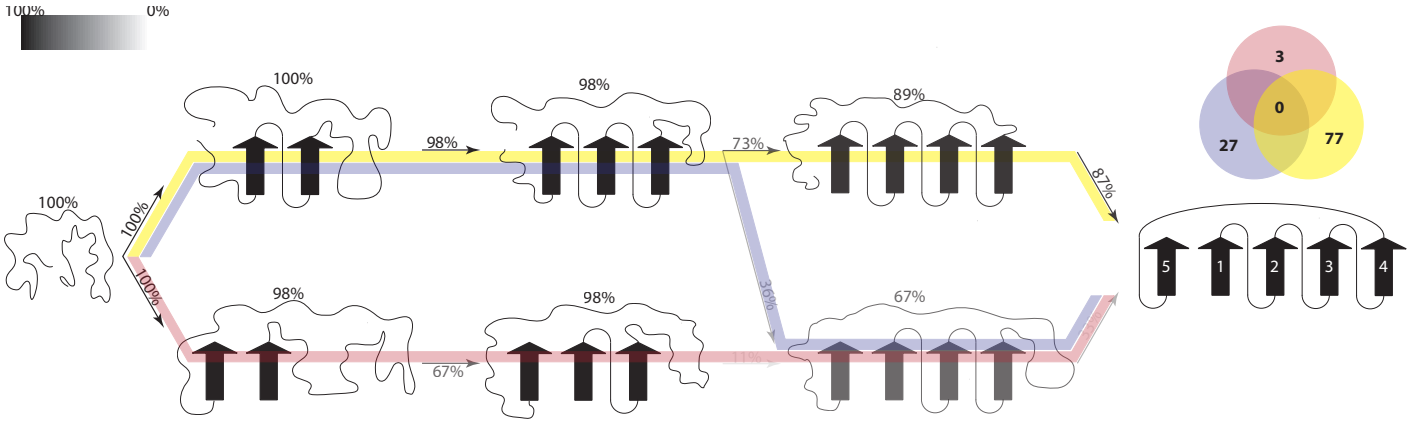

Figure 12: Pathways proteins 1KQ1, 1HK9, 1H64: The figure represents the predicted transitions from a random coil to the native state of proteins, belonging to the Pfam family PF01423, as a path in a graph of varyingly folded protein conformation states. The nodes in the graph represent energetically accessible conformation states which have been previously generated by the proposed Boltzmann-weighted ensemble sampling method. The percentage at top of each vertex (and its corresponding transparency) count for the number of times that this topology is predicted as a transition over the total number of runs. The vertices in the graph represent the transition between two topologies. The topologies connected by an edge are compatible topologies with structural similarity. The percentage at top of each edge (and its corresponding transparency) count for the number of times that this transition is found over the total number of runs. The predicted folding pathways are presented as paths in the graph. The percentage of presence of each path is presented through a Venn diagram at the right top corner of the figure.

This ability is studied on the different

folding intermediates nuclei conserved residues

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

6 Results

7 Conclusions and Discussions

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, protein folding methods aim the development of reliable prediction of protein 3D structures and structured folding pathways. An enormous challenge for protein folding prediction methods has been to predict 3D native structures and folding pathways for the broad range of proteins. This broad range is composed by thousands of different folds, thousands of different structural families and an unknown number of different folding mechanisms. Additionally, 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, however, 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.

In this paper we propose a new and effective coarse grained methodology for the pathway prediction problem using few computational resources. Specifically, we propose a new methodology to combine the ensemble modelling and the evolutionary based sequence information for folding pathway prediction. Particularly, the residue contact information is integrated into a Boltzmann sampling process to circumvent the limitations of potential energy scoring schemes and to narrow the conformational search space (the two most important bottlenecks in protein folding prediction). The proposed method expands the scope of previous ensembles prediction techniques. The proposed method differs from previous works in the following features. *i*) This work shows a clear improvement in performance (i.e. speed and accuracy). *ii*) A new energy model is implemented based on joint probabilities mimicking an hierarchy folding process and the ability of proteins to adopt different conformational states in vivo. *iii*) The proposed method exploits the evolutionary record in protein families adding information about evolutionary constrained interactions in the protein folding prediction. $iv-$) The method needs as input the protein sequence alone and it does not require any a priori knowledge of the native protein structure. $v-$) The method is able to model pure β, pure α and α/β interactions.

`efold` is tested using a considerable corpus of proteins with low identity, which includes the main protein fold families α/β, $\alpha+\beta$ and all-β. From the experiments, it can be stated that the proposed model is a good protein structure and pathway predictor.

There is not a clear consensus about what accuracy, coverage, and distribution of contacts

along the sequence are needed to improve the prediction of protein structures and/or pathways, however, in general the incorporation of contact information into protein folding programs leads to improvement of the results. Particularly, the correct prediction of long-range contacts must be predicted correctly to allow an accurate folding pathways to be reconstructed. Long-range contacts narrow the search space of possible conformation imposing strong constraints on the 3D structure. `efold` represents an ensemble predictor that is conceptually different to the state-of-the-art algorithms in contact prediction, however, it produce results comparable results with those algorithms. Furthermore, `efold` reports very good results, when compared with standard evaluation measures, for contact and residue prediction.

There is a growing body of evidence that indicates that proteins exhibit simultaneously a variety of folding pathways. Some paths will be more populated than others. Then, it is important to recognize the statistical dimensions of the protein folding process and to consider protein ensembles that could mimic the ability of proteins to adopt different conformational states in vivo. `efold` is an ensemble algorithm that follows this research line. Particularly, it predicts a statistical distribution of topologically allowed pathways through the use of Boltzmann probability function and the simulation of population dynamics to statistically characterized the protein ensembles. The good balance of `efold` regarding its effectiveness and efficiency, and its nature to predict structures from sequences alone, allows him to be used on large corpus of data, and eventually contribute decreasing the current gap between protein sequence and structure information.

Despite their importance, there are little experimental knowledge of protein-folding energy landscapes. Furthermore, there is not a good understanding of the folding routes or transition states for arbitrary protein sequences. Substantial improvements have been observed for protein folding methods. Particularly, the best predictions in CASP have been shown in average accurate enough to interpret biological mechanisms, to guide biochemical studies, or to initiate a drug discovery programs [66]. In spite of these improvements, there are great challenges to achieve in terms of the determination of a folding mechanism, of making ab-initio predictions consistent enough to decrease the current dependency on knowledge of existing structure, and of studying folding diseases, drug affinities, membrane proteins and disorder proteins, to name some. `efold` is a coarse-grain algorithm that does not provide specific solutions to any of those challenges. However, it is a method that with a low cost of computer resources, allows the collection of statistics over many protein trajectories, sampled over varying conditions and various models giving insights on the structures and pathways of proteins to draw general conclusion on those challenges.

8 Report December 3, 2014

8.1 Conclusions of the experimental framework

1. The proposed method improves the prediction of residue contacts.

   - :) The proposed method improves the prediction of residue contacts with respect to our previous approach (compared with protein G).
   - :) The proposed method improves the prediction of L_{500} for EvFold.
   - :) The proposed method predicts residue contacts with an average greater that 50% for all quality measures for ± 2 for all contact separations in the complete benchmark..
   - :) Exact prediction tends to be higher than 0.2 for all the contact separations.
   - :) The precision of strand prediction is high in the benchmark. Specially for ± 2.
   - :) An "homogeneous" behaviour can be noted in the data set regardless the diversity of the benchmark. Furthermore, there is a good balance between contact and strand prediction.
   - :— Our method is not a contact or secondary structure prediction method. However, our method is flexible enough to study TSs and nuclei residues in different proteins.
   - :(The variance for our results is very high, it can be explained given the diversity of our study

2. The proposed method improves the prediction of protein topologies.

   - :) The proposed method does not have the constrains imposed to our previous approach.
   - :) In average, the proposed approach ranks the target cluster in the top 2%.
   - :— It is hard to correlate the rank results with folding dynamic results, however, we have the data and we can select some proteins to do that.
   - :(The method predicts a lot of topologies. This prediction grows exponentially with the number of strands. It will be important to generate a filter during the running time.

3. The proposed method has a good prediction of protein pathways.

   - :) The proposed method correlates well the in-silico data vs experimental data (3 full experiments). However this correlation is constrained by the level of detail given by our method and the lack of an helix analysis. Inspection of the results reveals that the folding intermediates are consisted with previous literature reports.

Specifically, it is consistent with the work reported by ([34, 35]), with respect to the 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]). Our results also show the nucleation of the β-sheet residues between $\beta 1$ and $\beta 4$ as a next folding event (See permutation 3 4 1 2 with pairings *ANTI NONE ANTI*). This folding event can be preceded by the folding interaction between the α-helix and the second hairpin as also reported in ([31]). Interestingly, there are many interactions of three pairings (i.e., four β-strands), which agrees with previous findings about that four stranded β-sheets constitutes a metastable folding intermediate ([54]). This fact is very important because it gives a possible explanation about how the exposition of strand $\beta 1$ and the four β-strand complex can lead the amyloid aggregation process.

- :) The method is flexible enough to agree with different experimental results, even if those experiments are sometimes contradictory.
- :— The method validate some reported experiments and explore some pathways do not reported, however, the method can not make any consistent claim about new biological discoveries.
- :— The method is consistent in the prediction of TS. We were able also to study nuclei residues.
- :(First folding options do not agree with reported pathways.

4. The proposed method is able to discriminate and differentiate protein pathways even in high similar protein sequences.

   - :) The rank for proteins belonging to the "B" mutant group is better than the its "A" counterpart.
   - :) The pathways of the "B" mutant group keeps similarities with the "wildtype" group, however, there are also differences.
   - :) There are examples of similar sequences can keep very different pathways, also there are examples of similar sequences with similar pathways.
   - :) There are many combination to get conclusions.
   - :) The TS and nuclei residues are conserved. The bad news is that they are also conserved for the mutant A.
   - :(We do not have experimental data for the mutants.
   - :(We do not have a way to model the helices.

5. The proposed method is able to identify recurrent states of proteins belonging to a same PFAM family.
    - :) Pathways make sense with the structural features of the family.
    - :) Family PF00240 contains 1UBQ and family PF00018 contains SH3.
    - :— Problems validating the recurrent states. The folding pathways of disulphide proteins (Familly PF00014) vary substantially (BPTI vs TAP). Our study is based on BPTI and AMBP, there is no information about them.
6. The proposed method is able to identify recurrent states of proteins belonging to different PFAM families.
    - :) We were able to compare the nucleus residues of proteinG with ubiquitin family.
    - :) We were able to compare the pathways of proteinG with the ubiquitin family.
    - :— Our study is based on the PFAM paradigm, what about other organization models? SCOP to name one?.

TO DO LIST.....

1. Write down a paragraph for each of the conclusions listed previously.
2. Write down the conclusions.
3. Write down the abstract.
4. Filter the sections and paragraphs in the paper.
5. Choose a journal.
6. Write down the editor letter.
7. To clean the code.
8. To set the web service.

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.

[48] Bohdan Monastyrskyy, Daniel D'Andrea, Krzysztof Fidelis, Anna Tramontano, and Andriy Kryshtafovych. Evaluation of residue–residue contact prediction in casp10. Proteins: Structure, Function, and Bioinformatics, 82(S2):138–153, 2014.

[49] Jun Shimada and Eugene I Shakhnovich. The ensemble folding kinetics of protein g from an all-atom monte carlo simulation. Proceedings of the National Academy of Sciences, 99(17):11175–11180, 2002.

[50] Stefano Piana, Kresten Lindorff-Larsen, and David E Shaw. Atomic-level description of ubiquitin folding. Proceedings of the National Academy of Sciences, 110(15):5915–5920, 2013.

[51] Tobin R Sosnick, Robin S Dothager, and Bryan A Krantz. Differences in the folding transition state of ubiquitin indicated by φ and ψ analyses. Proceedings of the National Academy of Sciences of the United States of America, 101(50):17377–17382, 2004.

[52] Heather M Went and Sophie E Jackson. Ubiquitin folds through a highly polarized transition state. Protein Engineering Design and Selection, 18(5):229–237, 2005.

[53] Péter Várnai, Christopher M Dobson, and Michele Vendruscolo. Determination of the transition state ensemble for the folding of ubiquitin from a combination of φ and ψ analyses. Journal of molecular biology, 377(2):575–588, 2008.

[54] Philipp Neudecker, Paul Robustelli, Andrea Cavalli, Patrick Walsh, Patrik Lundström, Arash Zarrine-Afsar, Simon Sharpe, Michele Vendruscolo, and Lewis E Kay. Structure of an intermediate state in protein folding and aggregation. Science, 336(6079):362–366, 2012.

[55] Rajanish Giri, Angela Morrone, Carlo Travaglini-Allocatelli, Per Jemth, Maurizio Brunori, and Stefano Gianni. Folding pathways of proteins with increasing degree of sequence identities but different structure and function. Proceedings of the National Academy of Sciences, 109(44):17772–17776, 2012.

[56] Patrick A Alexander, Yanan He, Yihong Chen, John Orban, and Philip N Bryan. The design and characterization of two proteins with 88% sequence identity but different structure and function. Proceedings of the National Academy of Sciences, 104(29):11963–11968, 2007.

[57] Yanan He, Yihong Chen, Patrick Alexander, Philip N Bryan, and John Orban. Nmr structures of two designed proteins with high sequence identity but different fold and function. Proceedings of the National Academy of Sciences, 105(38):14412–14417, 2008.

[58] Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, and David E Shaw. How fast-folding proteins fold. Science, 334(6055):517–520, 2011.

[59] Carlo Travaglini-Allocatelli, Ylva Ivarsson, Per Jemth, and Stefano Gianni. Folding and stability of globular proteins and implications for function. Current opinion in structural biology, 19(1):3–7, 2009.

[60] Arash Zarrine-Afsar, Stefan M Larson, and Alan R Davidson. The family feud: do proteins with similar structures fold via the same pathway? Current opinion in structural biology, 15(1):42–49, 2005.

[61] Joan L Arolas, Francesc X Aviles, Jui-Yoa Chang, and Salvador Ventura. Folding of small disulfide-rich proteins: clarifying the puzzle. Trends in biochemical sciences, 31(5):292–301, 2006.

[62] Per J Kraulis. Similarity of protein g and ubiquitin. Science, 254(5031):581–582, 1991.

[63] John P Overington. Comparison of three-dimensional structures of homologous proteins. Current Opinion in Structural Biology, 2(3):394–401, 1992.

[64] A Maxwell Burroughs, S Balaji, Lakshminarayan M Iyer, and L Aravind. Small but versatile: the extraordinary functional and structural diversity of the beta-grasp fold. Biol Direct, 2(3):18, 2007.

[65] Stephen W Michnick and Eugene Shakhnovich. A strategy for detecting the conservation of folding-nucleus residues in protein superfamilies. Folding and Design, 3(4):239–251, 1998.

[66] Ken A Dill and Justin L MacCallum. The protein-folding problem, 50 years on. Science, 338(6110):1042–1046, 2012.