efold: An algorithm to Predict Protein Folding Pathways using Ensemble Modelling and Genomic Variation

David Becerra[1] Jérôme Waldispühl[1*]

[1]School of Computer Science and McGill Centre for Bioinformatics, McGill University, Montreal, QC, Canada

[*]To whom correspondence should be addressed; E-mail: jeromew@cs.mcgill.ca

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.

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*).

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). Both problems have been widely acknowledged as open problems, however PSP 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 has 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).

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 modeling process. 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 modeling and evolutionary information content-based methods belong to a newer and promising group of approaches that aims to offer a better trade-off between efficiency and accuracy for predicting structures and folding pathways.

Ensemble modeling methods employ a coarse-grained structural model that enables us to efficiently compute a complete conformational landscape and apply statistical mechanics techniques. 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 (*3, 4*). Although, those approaches can not directly be mapped to proteins, they have been an excellent starting point to model more accurate and complex scenarios. With respect to the protein structure scheme, we have already introduced a structural ensemble predictor for transmembrane β-barrel (TMB) proteins (*5*) continuing earlier work on molecular structure modelling (*6, 7*). We also introduced a method for modelling the folding process of large β-sheet proteins using sequence data alone (*8*).

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 (*9*). 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 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 that allows their systematic application to 3D structure prediction studies.

In this work, we introduce `efold`, a new protein folding pathway prediction framework that combines ensemble-modeling 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. `efold` also expands our previous protein folding prediction frameworks in several directions while keeping its low CPU-time requirements. First, `efold` models α-helices and multiple β-sheets. Next, `efold` algorithm applies memoization techniques and computes the conformational landscape of all β-sheet topologies 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 proposed method is tested on a benchmark of 125 proteins obtaining excellent results in terms of contact, strand, topology and pathway prediction.

Results

In order to understand the performance of the proposed method, experiments were performed to study the modeling ensemble (protein structure prediction) and modeling folding phases (protein pathway prediction). `efold` runs having as input only the amino-acid sequence and a set of parameters. Based on the experimental framework and user experiences with our previ-

ous techniques, we have fixed the limits of our algorithm 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.

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`. A pair of residues is 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. The results for the contact and strand prediction experiments are discussed in the Section 0.1.

`efold` performs the generation of all the admissible β-sheet topologies and the computation of the Boltzmann partition function over all the attainable topologies. Next, `efold` ranks those topologies based on their energy states and it clusters the top low-energy states to predict the folding dynamics. Then, in our experiments, we are interested in quantify the position of the topology reported by the corresponding PDB file in the rank computed by `efold` for each protein in the benchmark. Consequently, the experiments reported in Section 0.2 quantify the ability of `efold` to correctly rank the admissible β-sheet topologies.

To study the efficacy of our technique for predicting protein-folding pathways, we studied the folding landscape of proteins for which their pathways have been elucidated through many experimental studies and/or MD simulations. Graphs for specific folding pathways were constructed by considering all pairs of clusters computed during the modeling ensemble phase. If the minimum distance between two clusters was less than a transition threshold, we considered that there was exchange between the two states. The results of this pathway prediction procedure are discussed in Section 0.3.

Understanding the rules that govern the folding of proteins is one of the goals of biophysical studies that is still far from being achieved. 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 analyzes 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 0.4. On the other hand, results based on the second strategy are reported in Section 0.5.

0.1 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 8Å apart in the PDB file. The precision, sensitivity and F-measure 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.

The column (*Contact Prediction*) in figures 1 and S1 reports the best results of the experiments for the complete set of proteins and the set split by the number of strands, respectively. 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 by state-of-the-art algorithms for proteins without homology-based templates averages around 20% (*10*). 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.

Table 1 compares the results of `efold` and our previous model `tfolder` when predicting residue contacts for the Protein G. It is worth noticed 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 S2 (column *Contact prediction*) reports 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 Document 1 for a complete list of the size of MSAs). Figure S2 shows that the methodology implemented by `efold` improved the initial contact predictions performed by `EVfold`.

Figure 1 (column *Strand prediction*) reports the best results of the experiments for contact prediction of residue-residue contacts involved in β-sheet structures. The results spliced by the number of strands in the tested proteins is shown in Figure S1. 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 S2 (column *Strand prediction*) reports the results of the experiments 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.

0.2 Protein Topologies Prediction

Figure 1 (column *Rank*) reports 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.

Figures 1 and S1 (column *Rank*) show 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 1 (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.

0.3 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 6) 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.

0.3.1 Protein G

Figure 2(A) 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 (*11, 12*), 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 (*13*). Our results also show 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 (*13, 14*) were it has been shown that Protein G 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 of 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 Figure 2), 16% of the runs have a $helix - hairpin2$ complex and the 64% a $\beta 1 - \beta 4$ sheet complex. These pathways agrees with those found in (*14*). 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.

Figure 2(B) depicts the folding dynamics of Protein G. This figure allow a detailed understanding of the simulated folding process through the characterization of the protein conformations that emerge along the folding pathway. Particularly, the topology $\beta 3\mathbf{A}\beta 4\mathbf{P}\beta 1\mathbf{A}\beta 2$ has

the highest probability of observing this topology at the end of the protein folding process. This topology corresponds with the reported topology in PDB for the Protein G. The topology $\beta 3\mathbf{A}\beta 4$ presents the highest peak (around the time -1) and it finishes the simulation with the second highest probability. This topology correspond to the second hairpin and it has been reported to play a fundamental role during the folding. The picture emerging from our dynamic simulations is in agreement with previous dynamics experiments and simulations of Protein G folding. Specifically, the two most probable first folding events are the formation of the first and second β-hairpins (happening around the time -1), then at time -0.5 the most probable topology corresponds to $\beta 3\mathbf{A}\beta 4 \ \ \beta 1\mathbf{A}\beta 2$, which considers the complete formation of the β-hairpins. Next, the nucleation of the β-sheets residues between $\beta 1$ and $\beta 4$ occurs at time 0. Finally, from time 0 to 4 the protein obtains the correct fold topology (i.e., $\beta 3\mathbf{A}\beta 4\mathbf{P}\beta 1\mathbf{A}\beta 2$) and few fluctuation are observed for the last part of the simulation.

0.3.2 Ubiquitin

Figure 3(A) 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 $\beta 1$ and $\beta 2$ strand. This topology has been described to participate in a polarized and well-defined transition state ensemble (*15–17*). Regions of the local $\beta 1\mathbf{A}\beta 2$ hairpin populate native geometries, then this secondary structure can be stabilized by tertiary interactions between the α-helix and the $\beta 1\mathbf{A}\beta 2$ sheet.

Computational analysis of experimental ϕ-values suggested that strands $\beta 3$ and $\beta 5$ might be formed adopting a native-like topology in the TSE (*18*). 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 ($\beta 1\mathbf{A}\beta 2\mathbf{P}\beta 5\mathbf{A}\beta 3$). Then, in the ubiquitin folding pathway, regions of the local $\beta 1 - \beta 2$ hairpin populate native geometries. Next, strand $\beta 5$ joins the nascent hairpin-helix nucleus. Finally, the more distal strand $\beta 3$ is joined to the core structure (*19*). Our simulations show an important role of the topology $\beta 3\mathbf{A}\beta 5$ when combining with the topology $\beta 1\mathbf{A}\beta 2$. Particularly, it can be noted that the topology ($\beta 2\mathbf{A}\beta 1\mathbf{P}\beta 5\mathbf{A}\beta 3$) is the most populated edge in the graph. This topology receives the flow containing most of the in-coming paths. Furthermore, this topology represents also the edge with more connections with the edge representing the native topology (i.e., ($\beta 4\mathbf{A}\beta 3\mathbf{A}\beta 5\mathbf{P}\beta 1\mathbf{A}\beta 2$)). The different paths coming into the topology $\beta 2\mathbf{A}\beta 1\mathbf{P}\beta 5\mathbf{A}\beta 3$ can represent the heterogeneity predicted for the transition state structure, where the routes that stem from the core nucleus can 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 2(B) depicts the folding dynamics of Protein G. It is important to note that the simulation predict the correct fold topology with the highest probability. The topology ($\beta 2\mathbf{A}\beta 1\mathbf{P}\beta 5\mathbf{A}\beta 3$) reports the second highest probability and it stresses the importance given by our simulations to this four-stranded sheet network during the folding process. Our simulations of the transition paths of ubiquitin are also consistent with experimental and in-silico results. Particularly, our coarse level simulation reveals the formation of the sheet $\beta 1\mathbf{A}\beta 2$ and the $\beta 1\mathbf{P}\beta 5$ as the first folding steps. Next, the topologies $\beta 2\mathbf{A}\beta 1\mathbf{P}\beta 5$ and $\beta 3\mathbf{A}\beta 5\mathbf{P}\beta 1$. Both topologies have in common the central strand $\beta 5$, emphasizing the ability of the algorithm to identify this critical structural component through the modelling of non-local contacts. Once all the elements are folded, the four-stranded sheet network ($\beta 2\mathbf{A}\beta 1\mathbf{P}\beta 5\mathbf{A}\beta 3$) start to form around the time 0. Finally, at time 0.5 the native topology ($\beta 4\mathbf{A}\beta 3\mathbf{A}\beta 5\mathbf{P}\beta 1\mathbf{A}\beta 2$) start to be the most probable state until the end

of the simulation.

0.3.3 SH3 domain - PF00018

Figure 4(B) 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 $\beta 1\mathbf{A}\beta 2\mathbf{A}\beta 3\mathbf{A}\beta 4$ (yellow path in Figure 4(B)) was present in 83% of the simulations. Then, the studied SH3 domains agree in the formation of the $\beta 1 - \beta 2$, $\beta 2 - \beta 3$, and $\beta 3 - \beta 4$ motifs. Previous findings (*20*) suggested that this topology constitutes a metastable folding intermediate. This intermediate has been shown to be highly aggregation prone, as it exposes strand $\beta 1$. Then, the formation of native strand $\beta 5$ (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.

Two of the predicted pathways cross through the $\beta 2 - \beta 3 - \beta 4$ 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 $\beta 2 - \beta 3 - \beta 4$. From that topology, the paths branch in two different pathways (i.e., the red and yellow paths in Figure 4(B)). These path counts for the most probable paths and end the folding building the second sheet (a less structured topology) with two terminal strands ($\beta 1 - \beta 5$). 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.

0.4 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 (*21, 22*). 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.

0.4.1 PF00014

Figure 4(A) 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 $\beta 1\mathbf{A}\beta 2 \ \ \beta 3\mathbf{A}\beta 4$ and $\beta 2\mathbf{A}\beta 1\mathbf{A}\beta 3$ arrangements. These topologies are attributed to the proteins 1BIK and 5PTI, respectively. The topology $\beta 1\mathbf{A}\beta 2$ is traversed for all our simulations and it is present in all our simulations.

The folding pathways of disulfide proteins vary substantially (*23*). Particularly, it has been shown that with two structurally homologous kunitz-type protease inhibitors, bovine pancreatic trypsin inhibitor and tick anticoagulant peptide, there is a heterogeneity of folding intermediates and folding kinetics (*24*). 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 (*25*).

0.4.2 PF00240

Figure 4(C) 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 $\beta 2\mathbf{A}\beta 1\mathbf{P}\beta 4\mathbf{A}\beta 3$ 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,. $\beta 2\mathbf{A}\beta 1\mathbf{P}\beta 5\mathbf{A}\beta 3$) 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 $\beta 2\mathbf{A}\beta 1\mathbf{P}\beta 4$, 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 $\beta 2\mathbf{A}\beta 1\mathbf{P}\beta 4\mathbf{A}\beta 3$ is topographically related to the structure of Protein G 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 (*26, 27*). Later, this common fold was termed β-grasp and it has been suggested to be a multi-functional scaffold in diverse biological contexts (*28*). Studies have 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 (*29*). Monte Carlo simulations have identified

nucleus positions that are conserved among structures with homologous folds for the Protein G and Ubiquitin. Particularly, in (*14*), identified nucleus residues in hairpin 1 (Y3 and L5), the helix (F30), and the hairpin 2 (W43, Y45, and F52). All of these residues show low sequence entropy over aligned sequences in the ubiquitin superfamily (*29*). 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.

0.4.3 PF01423

Figure 4(D) 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 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.

0.5 Sequence vs Structure

An novel engineering approach allowed the obtention of set of proteins with high sequence identity but different structure and function (*30–32*). 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 three-helix bundle fold, and G_B displays a $\alpha + \beta$ Protein G fold. Table S1 shows the percentage identity between the different variants of the wild-type proteins and Protein G. These set of proteins represent a great opportunity to elucidate relationships and dependency between `efold`, sequence information and folding mechanisms.

Figure S3 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 S3 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 Protein G 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_{B88} 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 S3) that is not present in the other variants. Additionally. the wild type G_{B1} does not show the path involving an early formation of the $\beta 1 - \beta 4$ sheet. The previous simulations agree with the evidence that Protein G variants exhibit distinct folding routes, where the main difference between them is a different order of formation of the β-strands (*33*).

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 0.2) between the variants of GA folds versus the variants of GB folds, the predictions of the GB folds (3%), as a Protein G fold-like, outperform the predictions of the GA counterparts (6%). In other words, `efold` was able to predict with a high accuracy the GB folds as containing a Protein G fold-like topology; meanwhile `efold` predicted the GA folds as topologies with a lower probability to belong to a Protein G topology. It is important to stress that this divergent predictions were obtained on proteins with a high sequence identity, but different 3D structures.

Discussion

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 (*34*). 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 (*35, 36*) 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.

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. In this work, `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.

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 (*37*). 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 detailed 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. Then, the ability of `efold` to formulate quick, coarse-grained predictions in a matter of minutes or hours, rather than days of atomistic-detail simulation, can be used to support the initial stages of more complex and detailed models.

Materials and Methods

0.6 Algorithm Design

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 (*38*). 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 (*39, 40*). 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.

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 5. The specific details of the methodology are shown in the hereinafter subsections.

0.7 Modelling the ensembles

0.7.1 The forward step.

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

Computation of the Partition Function:

Conceptually, each protein structure was described by a coarse-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 6). 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 Dy-

namics 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 (*6*). 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 6a). 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 6). 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.

`efold` has to compute the partition function for all the attainable protein topologies. Thus, its time complexity depends on the computation of the function for each topology. Figure S4 depicts the dependency of `efold`'s complexity on the two primary factors influencing this calculation; the length of sequence, and the number of strands in the topology (i.e., the depth of the recursion).

The used evolutionary contact prediction method (*9*), 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 a parameter γ set by the user (see Equation 6)). The last procedure builds a selective constraint which can intensify the signal of β-strand interactions during the modelling of pathway kinetic.

$$weight_i = \big(\gamma \times EvFold(S_i)\big) + \big((1-\gamma) \times P(S_i)\big) \tag{6}$$

0.7.2 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 con-

figurations 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 6 for an example).

0.8 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 (*41*). 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 7).

$$\Delta G_{ij} = E(\chi_i) - E(\chi_j) = \sum_{x \in \chi_i} E(x) - \sum_{x \in \chi_j} E(x) \tag{7}$$

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 8)). 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 9, where p_i is the probability of state i, X is the state space).

$$r_{ij} = r_0 \exp(-\Delta G_{ij}/2RT) \tag{8}$$

$$\frac{dp_i}{dt} = \sum_{j \in X} r_{ij} p_j(t) \tag{9}$$

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 10. Then, the distribution of conformations over folding time is estimated by solving this system.

$$p(t) = \exp(Rt)p(0) \tag{10}$$

0.9 Protein Benchmark

0.9.1 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 (*13, 42*).

The B1 domain of protein G, generally called GB1 or Protein G, 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 (*43*).

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.

0.9.2 `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.

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.

0.9.3 916 Benchmark

The BetaSheet916 dataset was extracted from the Protein Data Bank of May 2004 by Cheng (*44*). 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 $15 - 20\%$. 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.

0.10 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.

1 Figures and Tables

Measure	Approach	$x > 0$		$x \geq 12$		$x \geq 24$	
		± 0	± 2	± 0	± 2	± 0	± 2
Precision	`efold`	**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	`efold`	**100**	**100**	**100**	**100**	25	**100**
	`tfolder`	56.3	97.9	53.8	61.5	**37.5**	87.5
F-measure	`efold`	20.75	**68**	**35.48**	**87.1**	**61.11**	**100**
	`tfolder`	**21.5**	**68**	18.1	69.2	20.3	70

Table 1: **Contact prediction performance of `efold` and `tfolder` for the Protein G.** 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.

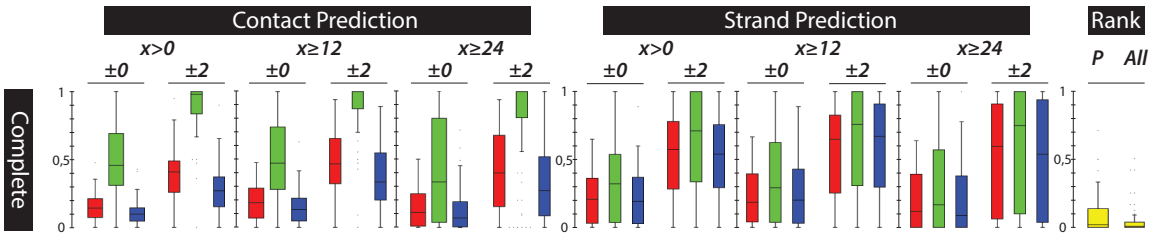

Figure 1: **Contact, Strand and Rank predictions performed by `efold` for the complete protein benchmark.** The performance is evaluated based on the precision(green), recall(blue) and F-measure(red) of experimentally observed contacts. The metrics are reported for contacts which are 0, 12 and 24 residues apart. The metrics are also studied when predicted contacts are within ± 2 residues of an observed contact. The rank is computed based on the position occupied by the correct protein topology with respect to the energies of the other topologies as computed by the Boltzmann partition function. A value close to zero means that `efold` is more likelihood to select the correct topology. The ranking is computed by considering all the topologies (column *All*) and the attainable topologies given a common parent (column *P*).

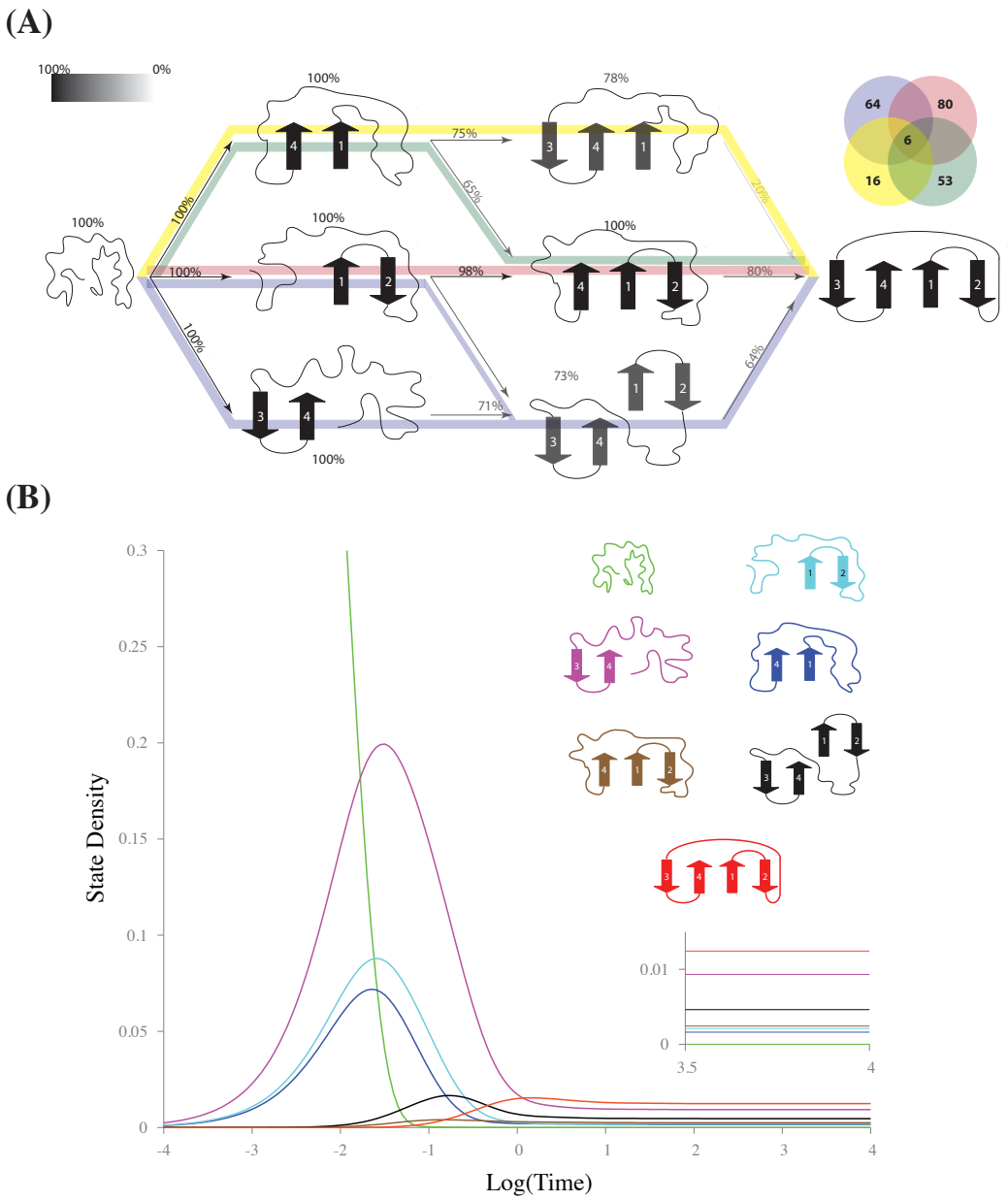

Figure 2: **Predicted transition from a random coil to the native state of Protein G. (A)** The transitions represented as a path in a graph of varyingly folded protein conformation states. The nodes represent energetically accessible conformation states. The vertices represent the transition between two topologies. The topologies connected by an edge are compatible topologies with structural similarity. The percentages at top of each vertex and edge (and its corresponding transparency) count for the number of times that this topology and transition, respectively, are 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. **(B)** The predicted folding dynamics of Protein G, which shows how the probability of observing any of the reachable topologies changes over time the protein folds. Each line is annotated through colors with the topology it represents.

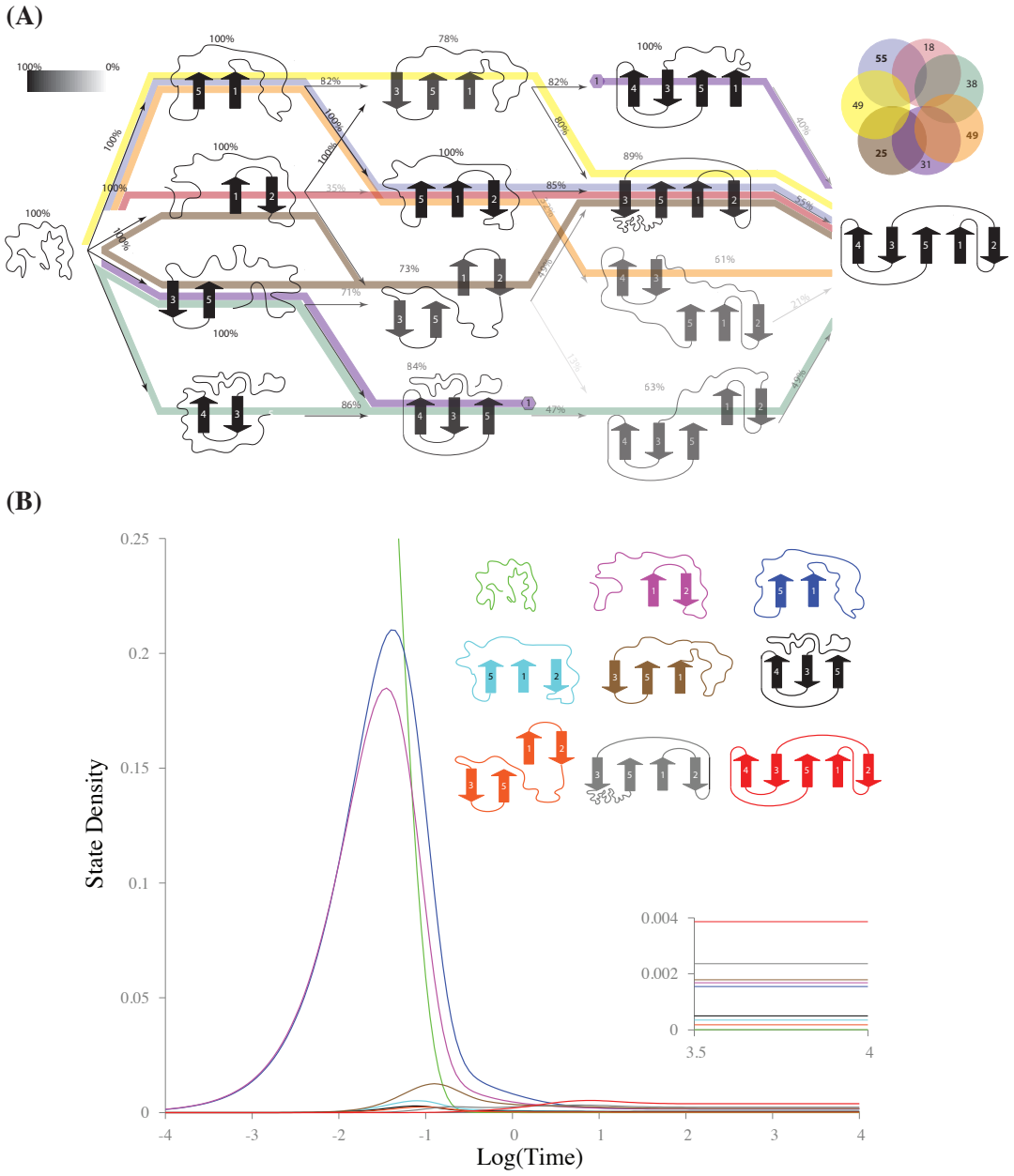

Figure 3: **Predicted transition from a random coil to the native state of Ubiquitin.** **(A)** The transitions represented as a path in a graph of varyingly folded protein conformation states. The nodes represent energetically accessible conformation states. The vertices represent the transition between two topologies. The topologies connected by an edge are compatible topologies with structural similarity. The percentages at top of each vertex and edge (and its corresponding transparency) count for the number of times that this topology and transition, respectively, are 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. **(B)** The predicted folding dynamics of Protein G, which shows how the probability of observing any of the reachable topologies changes over time the protein folds. Each line is annotated through colors with the topology it represents.

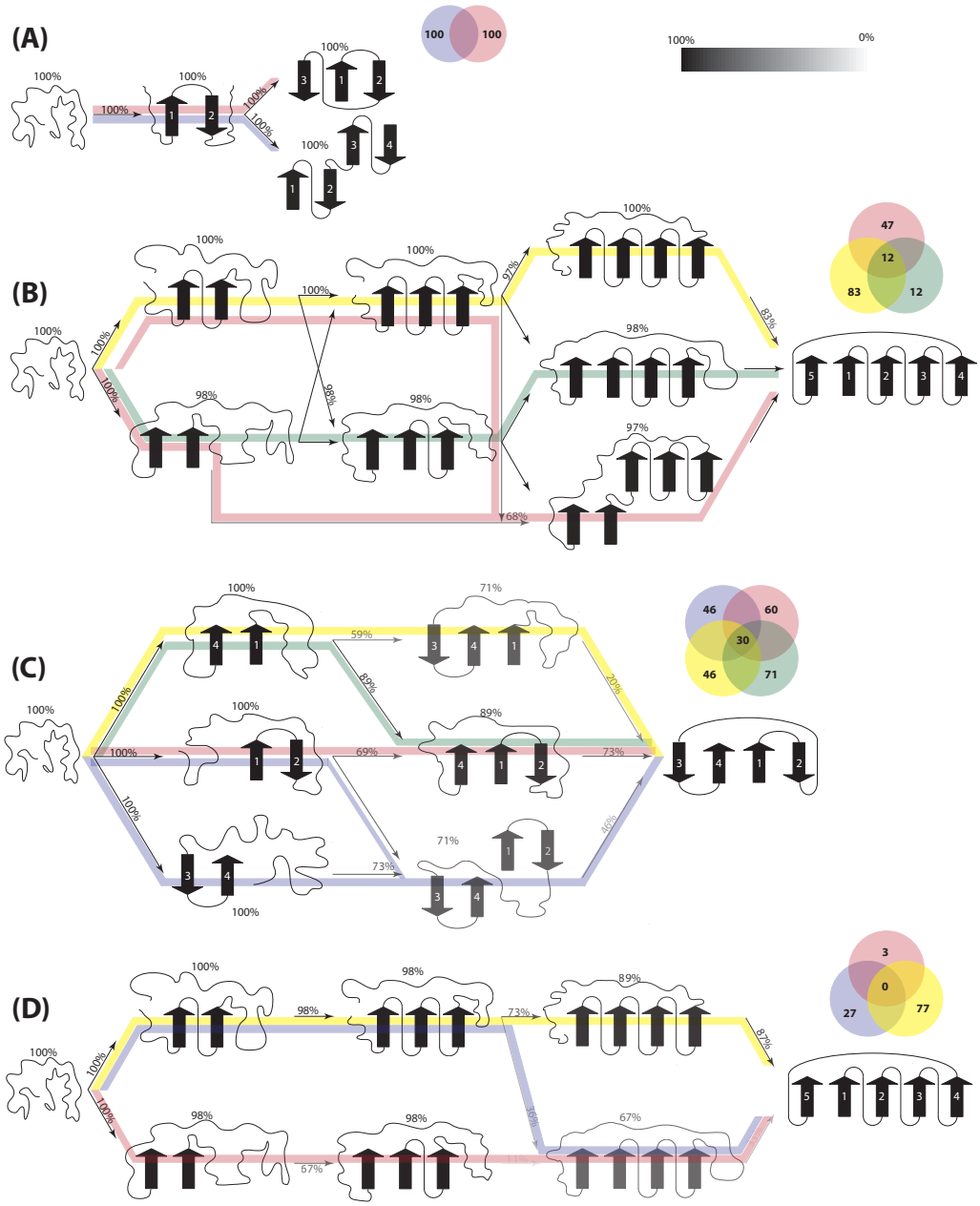

Figure 4: **Predicted transition from a random coil to the native state of the Pfam families (A) PF00014 (B) PF00018 (C)PF00240 (D)PF01423.** The are transitions represented as a path in a graph of varyingly folded protein conformation states. The nodes represent energetically accessible conformation states. The vertices represent the transition between two topologies. The topologies connected by an edge are compatible topologies with structural similarity. The percentages at top of each vertex and edge (and its corresponding transparency) count for the number of times that this topology and transition, respectively, are 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.

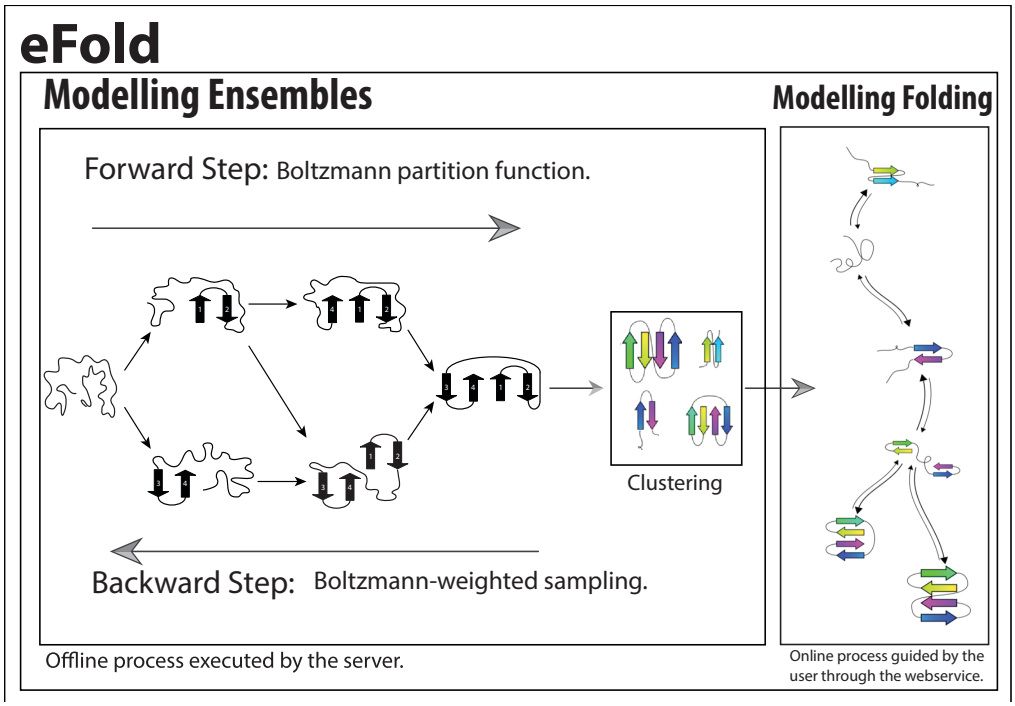

Figure 5: **`efold`: the proposed algorithm for predicting protein folding pathways and topologies using ensemble modelling and genomic variation.** The algorithm is divided in two main phases, the modelling of ensembles and the modelling of the predicted folding dynamics. The first phase is computed 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.

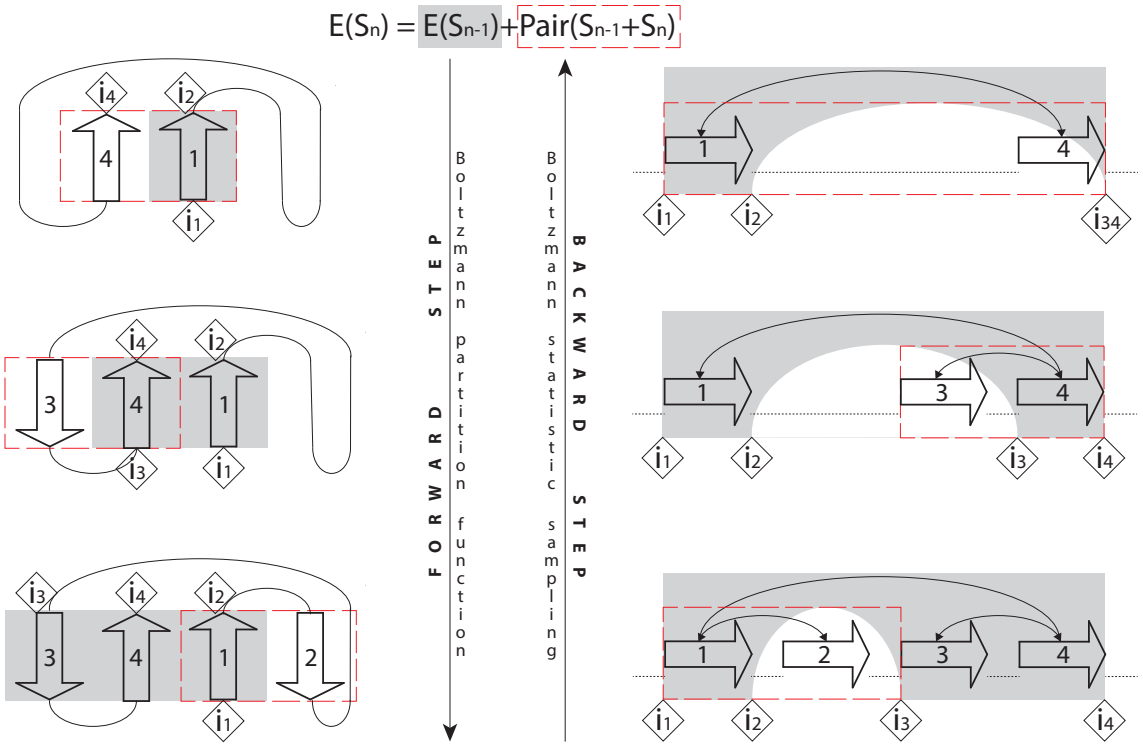

Figure 6: **Dynamic programming strategy encoded by `efold`**. An illustration of how `efold` encodes a permutable β-template representing the Protein G (i.e., $\beta 3\mathbf{A}\beta 4\mathbf{P}\beta 1\mathbf{A}\beta 2$) , how it recursively computes the Boltzmann partition function through an energy function composed by the sum of the contributions of the last two strands and the remaining structure (see left column of the figure), and how it performs a sampling procedure through the traceback of intermediate structures (see right column of the figure).

2 Supplementary Materials

Supplementary Material accompanies this paper at `http://http://csb.cs.mcgill.ca/efold/`.

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 S1: **Sequence identity between different variants of the wild-type Protein G.** The variants identified as G_{A1}, G_{A30}, G_{A77}, G_{A88}, and G_{A95} correspond to the GA fold, and G_{B1}, G_{B30}, G_{B77}, G_{B88}, and G_{B95} correspond to the GB fold. The Protein G is identified by its pdb code $1EM7$.

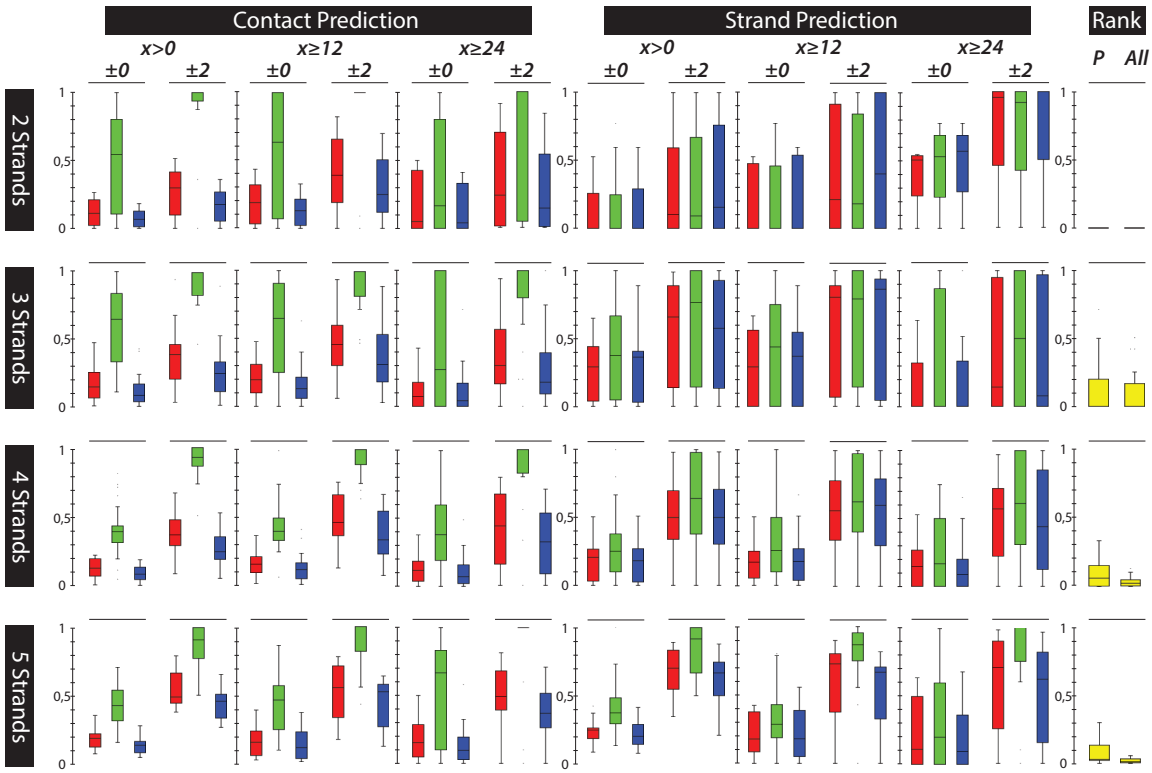

Figure S1: **Contact, Strand and Rank predictions performed by `efold` for the protein benchmark split by number of strands.** The performance is evaluated based on the precision(green), recall(blue) and F-measure(red) of experimentally observed contacts. The metrics are reported for contacts which are 0, 12 and 24 residues apart. The metrics are also studied when predicted contacts are within ± 2 residues of an observed contact. The rank is computed based on the position occupied by the correct protein topology with respect to the energies of the other topologies as computed by the Boltzmann partition function. A value close to zero means that `efold` is more likelihood to select the correct topology. The ranking is computed by considering all the topologies (column *All*) and the attainable topologies given a common parent (column *P*).

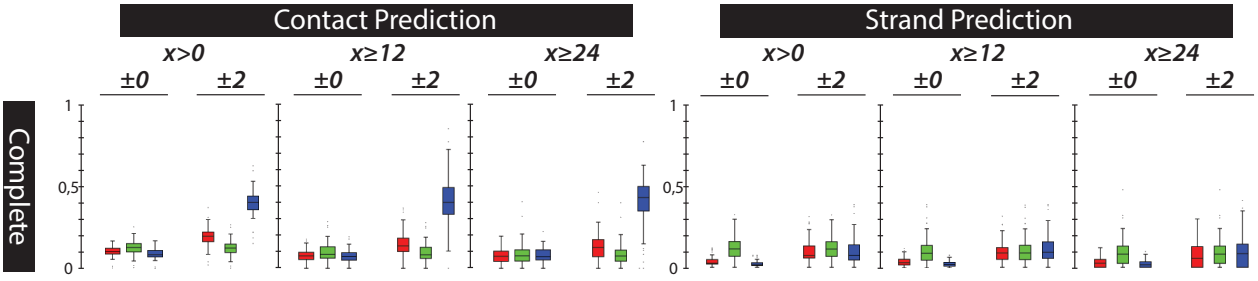

Figure S2: **Contact, Strand and Rank predictions performed by `EVfold` for the complete protein benchmark** The performance is evaluated based on the precision(green), recall(blue) and F-measure(red) of experimentally observed contacts. The metrics are reported for contacts which are 0, 12 and 24 residues apart. The metrics are also studied when predicted contacts are within ± 2 residues of an observed contact.

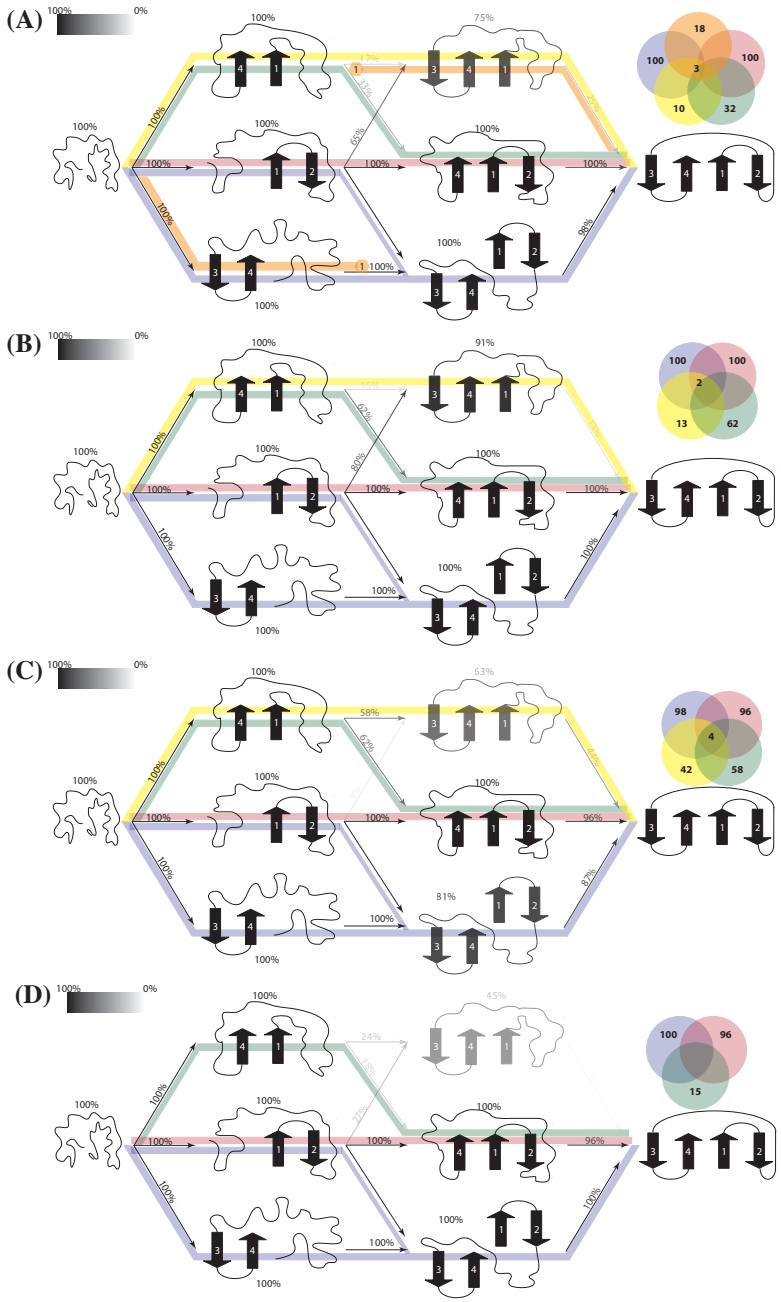

Figure S3: **Predicted folding transition of the variants corresponding to GB folds (A)** G_{B95} **and** G_{B88}**, (B)** G_{B77}**, (C)** G_{B30} **and (D)** G_{B1}**.** The transitions represented as a path in a graph of varyingly folded protein conformation states. The nodes represent energetically accessible conformation states. The vertices represent the transition between two topologies. The topologies connected by an edge are compatible topologies with structural similarity. The percentages at top of each vertex and edge (and its corresponding transparency) count for the number of times that this topology and transition, respectively, are 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.

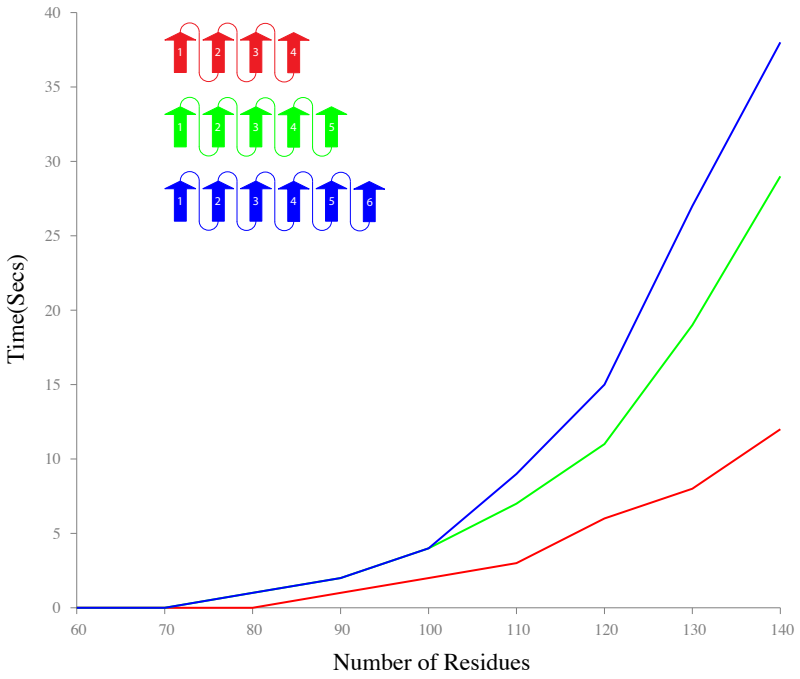

Figure S4: **Running time curve for the computation of the partition function by `efold`.** The time was computed by averaging over five independent runs, for sequences ranging from 40 to 140 residues in length, with 4, 5 or 6 strands. The experiment environment is a 2.8 GHz Intel Core i7 system, with 4GB of RAM, under the Mac OS X operating system.

References

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

2. D. Kihara, Y. Zhang, H. Lu, A. Kolinski, 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**, 5993 (2002).

3. Y. Ding, C. Lawrence, A statistical sampling algorithm for rna secondary structure prediction. *Nucleic acids research* **31**, 7280–7301 (2003).

4. J. McCaskill, The equilibrium partition function and base pair binding probabilities for rna secondary structure. *Biopolymers* **29**, 1105–1119 (2004).

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

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

7. J. Waldispühl, J. Steyaert, Modeling and predicting all-alpha transmembrane proteins including helix-helix pairing. *Theoretical computer science* **335**, 67-92 (2005).

8. S. Shenker, C. ODonnell, S. Devadas, B. Berger, J. Waldispühl, *Research in Computational Molecular Biology* (Springer, 2011), pp. 408–423.

9. D. Marks, *et al.*, Protein 3d structure computed from evolutionary sequence variation. *PloS one* **6**, e28766 (2011).

10. B. Monastyrskyy, D. D'Andrea, K. Fidelis, A. Tramontano, A. Kryshtafovych, Evaluation of residue–residue contact prediction in casp10. *Proteins: Structure, Function, and Bioinformatics* **82**, 138–153 (2014).

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

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

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

14. J. Shimada, E. I. Shakhnovich, The ensemble folding kinetics of protein g from an all-atom monte carlo simulation. *Proceedings of the National Academy of Sciences* **99**, 11175–11180 (2002).

15. S. Piana, K. Lindorff-Larsen, D. E. Shaw, Atomic-level description of ubiquitin folding. *Proceedings of the National Academy of Sciences* **110**, 5915–5920 (2013).

16. T. R. Sosnick, R. S. Dothager, B. 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**, 17377–17382 (2004).

17. H. M. Went, S. E. Jackson, Ubiquitin folds through a highly polarized transition state. *Protein Engineering Design and Selection* **18**, 229–237 (2005).

18. P. Várnai, C. M. Dobson, M. 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**, 575–588 (2008).

19. B. A. Krantz, R. S. Dothager, T. R. Sosnick, Discerning the structure and energy of multiple transition states in protein folding using ψ-analysis. *Journal of molecular biology* **337**, 463–475 (2004).

20. P. Neudecker, *et al.*, Structure of an intermediate state in protein folding and aggregation. *Science* **336**, 362–366 (2012).

21. C. Travaglini-Allocatelli, Y. Ivarsson, P. Jemth, S. Gianni, Folding and stability of globular proteins and implications for function. *Current opinion in structural biology* **19**, 3–7 (2009).

22. A. Zarrine-Afsar, S. M. Larson, A. R. Davidson, The family feud: do proteins with similar structures fold via the same pathway? *Current opinion in structural biology* **15**, 42–49 (2005).

23. J. L. Arolas, F. X. Aviles, J.-Y. Chang, S. Ventura, Folding of small disulfide-rich proteins: clarifying the puzzle. *Trends in biochemical sciences* **31**, 292–301 (2006).

24. J.-Y. Chang, Distinct folding pathways of two homologous disulfide proteins: bovine pancreatic trypsin inhibitor and tick anticoagulant peptide. *Antioxidants & redox signaling* **14**, 127–135 (2011).

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

26. P. J. Kraulis, Similarity of protein g and ubiquitin. *Science* **254**, 581–582 (1991).

27. J. P. Overington, Comparison of three-dimensional structures of homologous proteins. *Current Opinion in Structural Biology* **2**, 394–401 (1992).

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

29. S. W. Michnick, E. Shakhnovich, A strategy for detecting the conservation of folding-nucleus residues in protein superfamilies. *Folding and Design* **3**, 239–251 (1998).

30. R. Giri, *et al.*, Folding pathways of proteins with increasing degree of sequence identities but different structure and function. *Proceedings of the National Academy of Sciences* **109**, 17772–17776 (2012).

31. P. A. Alexander, Y. He, Y. Chen, J. Orban, P. 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**, 11963–11968 (2007).

32. Y. He, Y. Chen, P. Alexander, P. N. Bryan, J. Orban, Nmr structures of two designed proteins with high sequence identity but different fold and function. *Proceedings of the National Academy of Sciences* **105**, 14412–14417 (2008).

33. K. Lindorff-Larsen, S. Piana, R. O. Dror, D. E. Shaw, How fast-folding proteins fold. *Science* **334**, 517–520 (2011).

34. D. Liebler, *Introduction to proteomics: tools for the new biology* (Humana Pr Inc, 2002).

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

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

37. K. A. Dill, J. L. MacCallum, The protein-folding problem, 50 years on. *Science* **338**, 1042–1046 (2012).

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

39. D. Becerra, A. Sandoval, D. Restrepo-Montoya, L. Nino, *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on* (IEEE, 2010), pp. 137–141.

40. S. R. D. Torres, D. C. B. Romero, L. F. N. Vasquez, Y. J. P. Ardila, *Proceedings of the 9th annual conference on Genetic and evolutionary computation* (ACM, 2007), pp. 393–400.

41. M. Wolfinger, W. Svrcek-Seiler, C. Flamm, I. Hofacker, P. Stadler, Efficient computation of rna folding dynamics. *Journal of Physics A: Mathematical and General* **37**, 4731 (2004).

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

43. A. Gronenborn, *et al.*, A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein g. *Science(Washington)* **253**, 657–661 (1991).

44. J. Cheng, P. Baldi, Three-stage prediction of protein β-sheets by neural networks, alignments and graph algorithms. *Bioinformatics* **21**, i75–i84 (2005).

Acknowledgements: The authors thank TBD for helpful conversations.

Funding: DB is supported by Colciencia's Francisco Jose de Caldas scholarship

Author Contributions: TBD

Competing Interests: The authors declare that they have no competing financial interests.

Data and materials availability: The `efold` web service, additional data and materials are available online at `http://csb.cs.mcgill.ca/efold/`.