Novel Perspectives on Protein Structure Prediction

Bonnie Berger, Jérôme Waldispühl

Abstract Our understanding of the protein structure prediction problem is evolving. Recent experimental insights into the protein folding mechanism suggest that many polypeptides may adopt multiple conformations. Consequently, modeling and prediction of an *ensemble* of configurations is more relevant than the classical approach that aims to compute a single structure for a given sequence. In this chapter, we review recent algorithmic advances which enable the application of statistical mechanics techniques to predicting these structural *ensembles*. These techniques overcome the limitations of costly folding simulations and allow a rigorous model of the conformational landscape. To illustrate the strength and versatility of this approach, we present applications of these algorithms to various typical protein structure problems ranging from predicting residue contacts to experimental X-ray crystallography measures.

1 Introduction

The prediction of a protein's tertiary structure from its primary structure is one of the most important problems in computational biology and biochemistry [24, 53] yet also one of the most difficult [7]. Classical approaches to predicting protein structure follow the traditional schema, which aims to associate a single structure to each sequence. While this view of the problem seems supported by the the way that data have been accumulated over years in databases, the reality of the phenomena as described by experimentalists can be significantly more complex [4]. For instance,

Bonnie Berger
Department of Mathematics & Computer Science and AI Lab, MIT, Cambridge, MA, USA, e-mail: bab@mit.edu

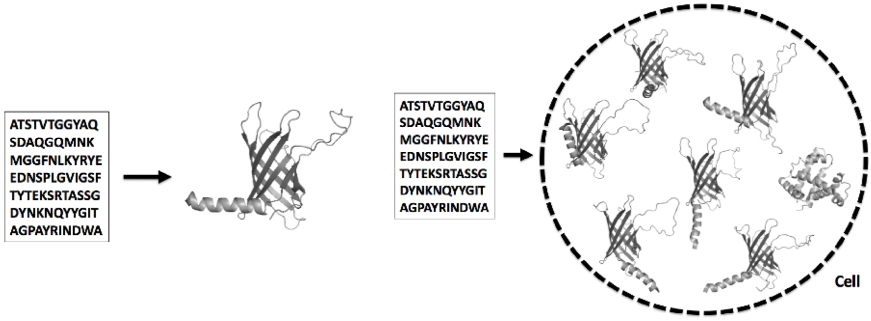
Jérôme Waldispühl
School of Computer Science, McGill University, Montreal, QC, Canada, e-mail: jeromew@cs.mcgill.ca

L.S. Heath and N. Ramakrishnan (eds.), *Problem Solving Handbook in Computational Biology and Bioinformatics*, DOI 10.1007/978-0-387-09760-2_9,

some proteins are intrinsically unstructured and characterized by lack of stable tertiary structure [27]. Other proteins such as prions have multiple stable, distinct, and functionally-related conformations [26, 44, 65]. There is also evidence that some proteins fold in multiple step processes using intermediate meta-stable structures in the folding landscape [61, 63]. Thus it is not unlikely for proteins to have alternate folds.

Beyond these examples, considering the protein structure prediction problem in a broader context can also radically change our perspective. Indeed, a cell contains many duplicates of the same protein sequence, which are all folding independently, potentially into similar but not necessarily identical structures. A molecule is never frozen forever in a rigid structure. In vivo, a polypeptide is perpetually adapting its structure, jumping from one stable conformation to another.

All these observations suggest that the protein structure prediction problem needs to be revisited. Computing a single conformation cannot reflect the diversity of the folds that a protein may adopt *in-vivo*. A complete view of the phenomena requires an embodiment of all these varying aspects of the same molecule in the same comprehensive model. We illustrate the differences between classical and modern approaches in Fig. 1. While the classical approach aims to assign a single structure to a given protein sequence (Fig. 1(a)), modern techniques aim to compute the *ensemble* of conformations that a polypeptide can adopt (Fig. 1(b)).



(a) Classical approach

(b) Ensemble approach

Fig. 1 Folding approaches. From a sequence, the classical method (a) aims to predict a single native structure, while the *ensemble* approach (b) aims to compute a picture of the complete ensemble of possible structures.

There has been substantial work on characterizing globular protein folding landscapes for lattice and non-lattice models – see Levitt and co-workers [55, 28, 29], Shakhnovich and co-workers [53, 40] and Dill and co-workers [16, 23, 52, 3, 37, 72]. In particular, it has been shown that the native state may be quite different from the predicted minimum energy conformation; indeed, Zhang and Skolnick [83] have shown that the native state is often closer to the centroid of the largest cluster of low energy conformations obtained by Monte Carlo sampling. Unfortunately, these studies could not complete a full description of the conformational landscape and were mostly restricted to computationally expensive folding simulations of single polypeptides.

In this chapter, we describe recent algorithmic advances that enable efficient computation of the complete *ensemble* of structures of a given polypeptide and prediction of stable conformations. These techniques aim to provide a realistic representation of the conformational landscape of a protein that could potentially be useful to study the folding dynamics of large polypeptides [69].

Seen at the cell level, the dynamic aspect of the system is indiscernible. The motion of individual molecules cannot be observed but the multiple conformational states remains visible. According to statistical mechanics principles, at equilibrium, the molecules achieving a particular fold are perpetually changing but the number of molecules in a specific state remains constant. Originally conceived for modeling the behavior of gas [13], the theory has been applied to other areas of computational biology, including the prediction of RNA secondary structure [50] and the study of transcription factor binding sites [6, 30, 54, 74].

We describe how statistical mechanical principles can be applied to modeling and predicting protein structures. As the theory is still progressing and a general discussion would be too long to conduct in this chapter, we will focus our discussion on the description of the first application of these techniques to a difficult but important class of proteins, namely the transmembrane β-barrel proteins [77, 75]. The techniques detailed in this chapter have potential to be extended to transmembrane α-helix bundles [79], certain β-sheet architectures [1, 11, 51, 21] and other structures that can be modeled using tree structures [16].

Transmembrane β-barrels (TMBs) constitute an important class of proteins typically found in the outer membrane of gram-negative bacteria, mitochondria and chloroplasts. These proteins display a wide variety of functions and are relevant to various aspects of cell metabolism. In particular, outer-membrane proteins (omps) are used in active ion transport, passive nutrient intake, membrane anchors, membrane-bound enzymes, and defense against membrane-attack proteins.

Since omps were discovered relatively recently and are difficult to crystallize, there are currently only about one hundred TMBs in the Protein Data Bank, and only 20 after the removal of homologous sequences. Some *in vitro* and *in vivo* mutation studies of omps [81, 46] have been performed, but, compared with the overwhelming amount of data on globular proteins, outer membrane proteins remain a biologically important but technically difficult area of research.

In this chapter, the ensemble of TMB structures of a given polypeptide is characterized by the Boltzmann partition function of the system. This achievement requires us (i) to provide a model of the structures to which we can apply dynamic programing principles for exploring the full conformational space, and (ii) to design an energy model which allow us to evaluate the stability of each conformation. From this quantity we show how to compute the Boltzmann pair probabilities $P(i,j)$ that residues i,j form an inter-β-strand contact, and rigorously sample conformations from the Boltzmann low energy ensemble. Additionally, we show how this partition function value can be used to estimate statistical mechanical parameters such as ensemble free energy, average internal energy, and heat capacity. Rigorously defined *stochastic contact maps*, sampling, and thermodynamic parameters give us insight into the folding landscape of outer membrane proteins — an insight that cannot be gained by methods solely dedicated to the prediction of native state conformations. This approach also provides a unified framework that allows us to simultaneously tackle a wide variety of structural prediction problems that were previously addressed by independent algorithms. This unified approach achieves a clear gain in accuracy, circumventing the problem of contradictory predictions encountered when interpreting the results of multiple, independent algorithms.

This chapter is organized as follows. In Section 2, we describe the combinatorial model used to represent the TMBs. Then, in Section 3, we introduce the energy model used to weight the structures, which consists of an extension of the state-of-the-art BETAWRAP energy model [11, 21] specialized for TMBs [75, 77].

In Section 4, we detail the algorithms used to compute the complete folding landscape. These algorithms run in polynomial time and space. These results have been obtained by taking advantage of the planarity imposed on a TMB by the cell membrane to derive a model that allows the computation of the partition function to be performed in polynomial time. A related approach was suggested by S. Istrail who proved that the partition function of an Ising model can be computed in polynomial time given a 2D lattice [42].

In Section 5, we illustrate the insight provided by these techniques by demonstrating its effectiveness on a variety of difficult protein prediction problems: (i) how to perform reliable residue contact predictions; (ii) how to provide a simple and intuitive representation of the folding landscape of a given polypeptide using *stochastic contact maps*; (iii) how X-ray crystal per-residue B-factors can be predicted with an accuracy rivaling that of leading specific B-factor prediction algorithms; and (iv) how Boltzmann-distributed structure sampling can be used to improve the accuracy of whole structure prediction over classical minimum folding energy approaches. In addressing this set of challenging structural prediction problems, we wish to underscore the strength and potential of this approach.

To conclude this chapter, we complete our review of recent protein structure ensemble analysis tools by addressing a related problem. Once a stable conformation has been identified, the question of how rigid or flexible the structure is remains. Thus, we present in Section 6 recent methods enabling efficient sampling of the local neighborhood of a given conformation.

2 Modeling transmembrane β-barrel structure

This section provides a simple and unambiguous representation of transmembrane protein structure that enables the design of dynamic programing equations for recursively enumerating all possible TMB structures. Originally, this modeling employed multi-tape context-free grammars [79, 75]; however, in this chapter we provide a more classical description using a graphical representation.

Transmembrane β-barrel (TMB) proteins are embedded in the outer membrane of Gram-negative bacteria, mitochondria and chloroplasts. The envelop of Gram-negative bacteria is built with two membranes (inner and outer) separated by a region called the periplasm. The composition of the bacterial outer membrane differs from that of the inner membrane by, among other things, the structure of its outer leaflet which include a complex lipopolysaccharide.

To accurately represent TMBs (in agreement with Schulz's summary [64]) three fundamental features of these structures are modeled: (i) the overall shape of the barrel (the number of TM β-strands and their relative arrangement); (ii) an exact description of the anti-parallel β-strand pairs which explicitly lists all residue contacts and their orientation (side-chains exposed toward the membrane or toward the lumen), as well as possible strand extensions; and (iii) the inclination of TM β-strands through the membrane plane. This decomposition of the structure into elementary units is illustrated in Figure 2.

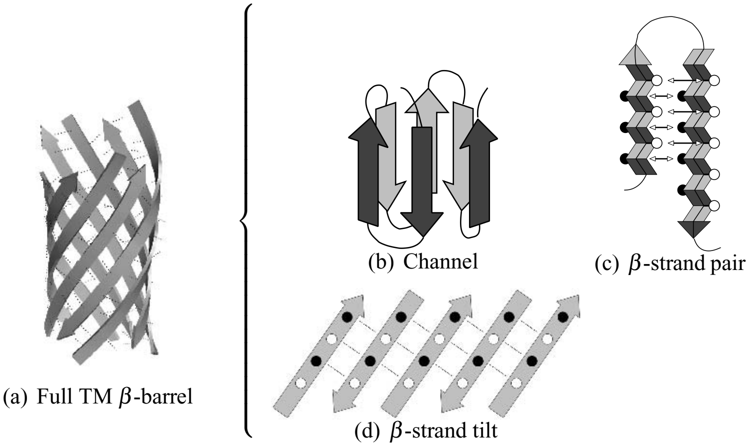


Fig. 2 Structure decomposition of transmembrane β-barrel. (a): The global structure of a transmembrane β-barrel. (b): Overall shape of the channel. (c): Anti-parallel β-strands. (d): Inclination of TM β-strands across the membrane plane.

The principle behind this modeling lies in a decomposition of the β-structure into individual blocks of β-strand pairs. In the case of TMBs, all these pairs are anti-parallel with the exception of the closing one in case the barrel has an odd number of strands. (Thus far only one TMB with an odd number of strands has been found via crystallization [5].) Consequently, the complete structure can be described as a sequence of individual β-strand pairs that will be used in Section 4 to design a dynamic programming algorithm for enumerating all the structures of the conformational landscape.

In TMBs, each strand is paired with two others. Graphically, our decomposition can be seen as follows: Instead of pairing each strand twice, we duplicate all strands and isolate each β-strand pair (see Fig. 3). Then, the barrel can be described as a sum of all its strand pairs.

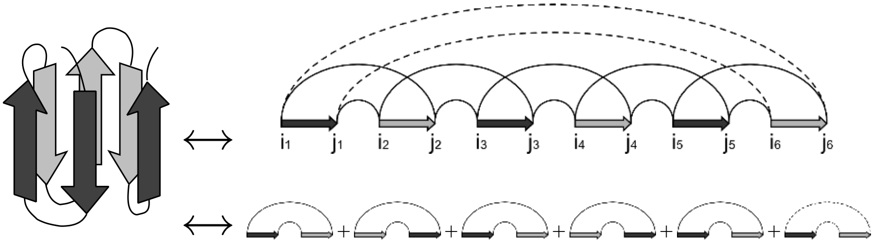


Fig. 3 Graphical decomposition of a transmembrane β-barrel. The strands are duplicated and each strand pair is isolated. The closing β-strand pair (with dashed lines) can be extracted and represented exactly as the others.

For TMBs, each strand is coupled with its two sequential neighbors (previous and next) and all pairings are anti-parallel with the exception of the closing strand pair that can be parallel if the barrel has an odd number of strands. With no restriction on generality, we will assume in this chapter that the TMBs have an even number of strands.

Formally, we define a β-strand pair (i.e. a block as seen in Fig. 3) with a 4-tuple $\binom{i_1,j_1}{i_2,j_2}$, where i_1 and j_1 (s.t. $i_1 < j_1$) are the indices of the left strand and i_2 and j_2 (s.t. $i_2 < j_2$) those of the right one (see Fig. 3). The left strand corresponds to the subsequence $[i_1, j_1]$, the right strand to $[i_2, j_2]$, and the loop connecting them corresponds to the subsequence $[j_1+1, i_2-1]$.

The length of the TM β-strands may vary. The number of residues in contact is $L_c = \min(j_1 - i_1 + 1, j_2 - i_2 + 1)$ and the length of the strand extension is $L_e = \left|(j_1 - i_1) - (j_2 - i_2)\right|$. To avoid invalid configurations, only one strand from each pair can be extended. In addition, for simplicity of description, we assume that the rightmost amino acid at index j_1 of the left strand is paired with the leftmost residue at index i_2 of the right strand. An example of a model freed from this constraint can be found in [78]. When an extension is done on the left strand, the right strand becomes shorter and the extension is called a *reduction* (Fig. 4(a)); when an extension occurs on the right strand, the latter is elongated and the operation, an *extension* (Fig. 4(b)).

The set $\mathscr{C}$ of residue-residue contacts involved in strand pairing can be defined as follows: $\mathscr{C} = \left\{(j_1 - k, i_2 + k) \,\middle|\, 0 \leq k < L_c\right\}$. The side-chain orientation alternates strictly around the strand backbone and can be labeled: *outwards*, that is facing toward the membrane, or *inwards*, that is facing toward the inside of the barrel, or channel (which can vary from entirely aqueous to mostly filled). Thus, we distinguish the subsets of residue contacts exposed to the same environment by $\mathscr{C}_0 = \left\{(j_1 - 2\cdot k, i_2 + 2\cdot k) \,\middle|\, 0 \leq k < \lfloor \frac{L_c}{2} \rfloor\right\}$ and $\mathscr{C}_1 = \left\{(j_1 - 1 - 2\cdot k, i_2 + 1 + 2\cdot k) \,\middle|\, 0 \leq k\right.$

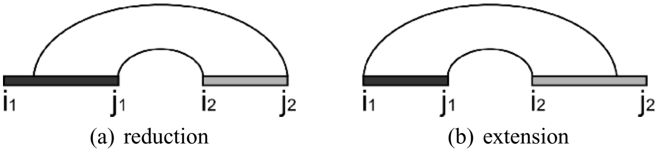


Fig. 4 (4(b)) Strand reduction: the left strand is elongated. (4(a)) Strand extension: the right strand is elongated.

$< \lfloor \frac{L_c}{2} \rfloor \}$. Assuming the location of the closest contact is known, we can also assign the nature of the milieu (i.e. membrane or channel).

Thus, we integrate these features in each block $\binom{i_1,j_1}{i_2,j_2}$ by annotating each residue appropriately. In practice, since residue labels strictly alternate, only the side-chain orientation of the first residue contact needs to be recorded. Figure 5 illustrates this modeling, although these details will be omitted when they are not crucial to the discussion.

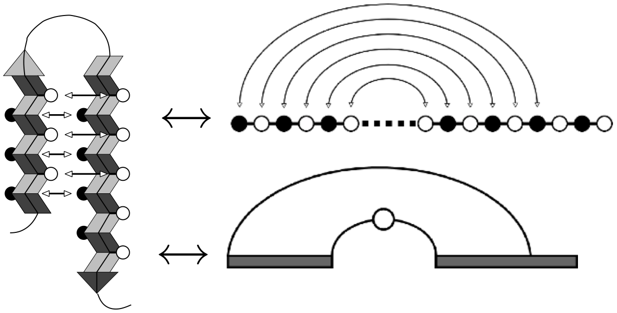


Fig. 5 Representation of a TM β-strand pair with extension on right strand. Residues are annotated by the side-chain orientation. Since the residue labels strictly alternate, only the first side-chain orientation needs to be indicated in a simplified representation (bottom).

The inclination of strands through the membrane is modeled using the *shear number*. This number represents the shift in the sequence of inter-strand residue contacts between consecutive β-strands, imposed by the inclination of these strands (cf. Fig. 6). This feature is implemented with the help of strand extension. Indeed, strictly alternating *reductions* and *extensions* to consecutive strand pairs allows us to obtain the desired configuration. Without loss of generality, and in conjunction with experimental observations [64], we assume that (i) the N-terminus is located on the periplasmic side and that (ii) the shear number is positive. It follows that the first loop (between the first and second TM strand) is on the extra-cellular side. Then we restrict *reductions* to occur around periplasmic loops and *extensions* around extra-cellular loops. Figure 6 illustrates how to proceed.

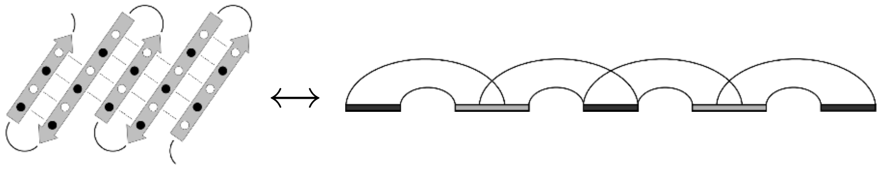


Fig. 6 Representation of strand inclination using shear number. *Reductions* and *extensions* alternate around periplasmic loops and extra-cellular loops in order to preserve the coherence of the orientation. The N-terminus of the protein sequence in the left diagram is at the right extremity.

It is noteworthy that, in principle, a similar representation could be used to include other classes of β-sheet protein domains as long as their structures follow similar topological rules. TMBs are well suited to this methodology since the cell membrane restricts the number of possible structural conformations that can arise, reducing the complexity of the representation. However, soluble β-barrel proteins can allow more flexibility in the barrel forming β-sheet and would thus require more sophisticated rules (such as consecutive strands that are out of sequence order) resulting in an increase in the computational complexity of the method.

3 Energy model

In this section, we describe a simple pseudo-energy model inherited from the state-of-the-art BETAWRAP energy model [11, 21] and specialized for TMBs [75]. In practice more refined versions have been designed [77]; however, all these rely on similar principles.

In the previous section, we defined a TMB as a sum of anti-parallel β-strand pairs, which are themselves defined as sequences of inter-strand residue contacts. These long-range interactions stabilize the β-strand pairs and thus the entire β-barrel. It follows that a reliable scoring function must explicitly integrate the stabilizing effect of these residue contacts.

We describe a simple model where the energy of the whole structure is the sum of the energies of each inter-strand residue contact found in the barrel. The challenge is thus to estimate reliable potentials for any pair of residues in contact. Unlike for RNAs, experimental measures to allow us to directly estimate the binding energies are not available for TMBs. Nevertheless, it is possible to estimate these potentials by computing residue contact statistics from known protein structures.

The statistical potentials for all possible amino acid pairs can be obtained by computing the probability of observing amino acid contacts in solved β-sheet structures with characteristics closely matching those found in TMBs. The classical approach used in [11, 21, 75, 77] takes a 50% non-redundant set of protein structures (`PDB50`) from the `PDB` [8], and uses `STRIDE` [31] to identify secondary structure features, solvent accessibility, and hydrogen bonds. Naturally, all solved structures of TMBs have to be removed as to not corrupt the testing.

In order to obtain the best possible estimate of these potentials, Berger and coworkers [11, 21] introduced a major conceptual advance with the BETAWRAP program. Instead of simply counting the occurrences of β-sheet amino acid pairings in all known proteins, the search is restricted to better match the environment associated with a given contact. Here, the barrel fold of TMBs is thought to consist of antiparallel, amphipathic β-sheets[1], with a hydrophobic environment in the outer membrane side of the barrel and a hydrophilic environment commonly existing within. Therefore, the anti-parallel bonded β-strands that exhibit an amphipathic pattern mimic relatively well the features of TM β-strand pairs, and thus can be used to count the frequency of pairs of residues. Alternating buried/exposed residues define amphipathicity. Usually, a buried residue is required to have less than 4% of the solvent accessible area as when that residue is in an extended G-X-G tripeptide [17], and an exposed residue is required to have an area greater than 15%.

The amino acid pair frequency counts are then used to estimate the probability $P(X,Y)$ of observing the amino acid pair (X,Y). Finer granularity information such as side-chain rotomers or atomic coordinates are not included in this model, but may be integrated into a more sophisticated model. Of note, Waldispühl *et al.* [77], introduce a variant of this model incorporating the notion of *stacking pairs* of adjacent pairs of residue contacts, which results in a significant gain of accuracy.

Once this amino acid pairs count is calculated, these frequency counts can be changed into statistical potentials. Let x,y be the indices of two amino acid that are in contact, and let $M \in \{0,1\}$ be a variable which represents the type of environment in which such a contact occurs (which side of the amphipathic sheet). Specifically, $M = 0$ ($M = 1$) when the side-chain orientation is toward the channel interior (membrane). Let $E(x,y,M)$ denote the energy of the contact between amino acids X and Y at positions x and y, with the environment M.

Pairwise frequencies are transformed into energy potentials using the standard procedure (taking the negative logarithm — see pp. 223–228 of [18] and [68] for details). Specifically, if $p_M(X,Y)$ is Boltzmann distributed, then $E(x,y,M) = -RT\log(p_M(X,Y)) - RT\log(Z_c)$, where $\log(Z_c)$ is a statistical re-centering constant that is chosen as a parameter. Further, although RT has no effect when computing the minimum folding energy structure [75], this is not the case when computing the partition function for β-barrel structures. For this reason, the current implementation in the program *partiFold* (http://partiFold.csail.mit.edu) [75, 76, 77] allows the user to stipulate an arbitrary Boltzmann constant. The folding pseudo-energy of the structure is the sum of all contact potentials. Formally, we have:

$$E = \sum_{(x,y)\in\mathscr{C}_0} E(x,y,0) + \sum_{(x,y)\in\mathscr{C}_1} E(x,y,1) \tag{1}$$

This model does not contain any energy contribution for periplasmic or extracellular loops, although such features can easily be computed and integrated with similar techniques.

[1] The amphipaticity defines a molecule which contains both polar (hydrophilic) and non-polar (hydrophobic) domains.

4 Algorithms

4.1 Computing the partition function

Since a TMB structure can be represented as a sequence of anti-parallel TM β-strand pairs, given any four indices i_1, j_1, i_2, j_2 and the environment M of the closing TM β-strand pair contact (i.e. "membrane" or "channel"), we can compute the energy $E(i_1, j_1, i_2, j_2, M)$ for the anti-parallel β-strand pairing of sequences $[i_1, j_1]$ with $[i_2, j_2]$. For all possible values of i_1, j_1, i_2, j_2 and M, we store the Boltzmann values $\exp\big(-E(i_1, j_1, i_2, j_2, M)/RT\big)$ in the array Q_{ap}. Since the length of TM strands, as well as those of strand extensions are bounded, the array can be filled in time $\mathscr{O}(n^2)$, where n is the sequence length.

$$Q_{ap}(i_1, j_1, i_2, j_2, M) = \prod_{k=0}^{L_c-1} \exp\left[-\frac{E(i_2-k, j_1+k, M+k \bmod 2)}{RT}\right] \tag{2}$$

Since the energy function is additive, we can decompose the energy of a TMB as the sum of the energies associated with each distinct anti-parallel TM β-strand pair. Let N be the number of TM β-strands of the TMB s and let i_k (resp. j_k) denote the index of the leftmost (resp. rightmost) residue of the k-th strand. In order to simplify the algorithms description, in the following we will omit the parameter M used to indicate the environment of the first contact of an anti-parallel TM β-strand pair. Therefore, the energy $E(s)$ of a given TMB structure s can be written as:

$$E(s) = E(i_N, j_N, i_1, j_1) + \sum_{k=1}^{N-1} E(i_k, j_k, i_{k+1}, j_{k+1}) \tag{3}$$

The Boltzmann partition function is defined as the sum $\sum_s e^{-\frac{E(s)}{RT}}$ taken over all the TMB structures s. To compute the partition function, we first introduce a dynamic table Q_{sheet} to store the partition function values for β-sheets built from concatenating anti-parallel TM β-strand pairs, i.e. TMB without closure. This table can be dynamically filled using the following recursion:

$$Q_{sheet}\begin{pmatrix} i_1, j_1 \\ i_k, j_k \end{pmatrix} = \sum_{(i_{k-1}, j_{k-1})} Q_{sheet}(i_1, j_1, i_{k-1}, j_{k-1}) \cdot Q_{ap}(i_{k-1}, j_{k-1}, i_k, j_k) \tag{4}$$

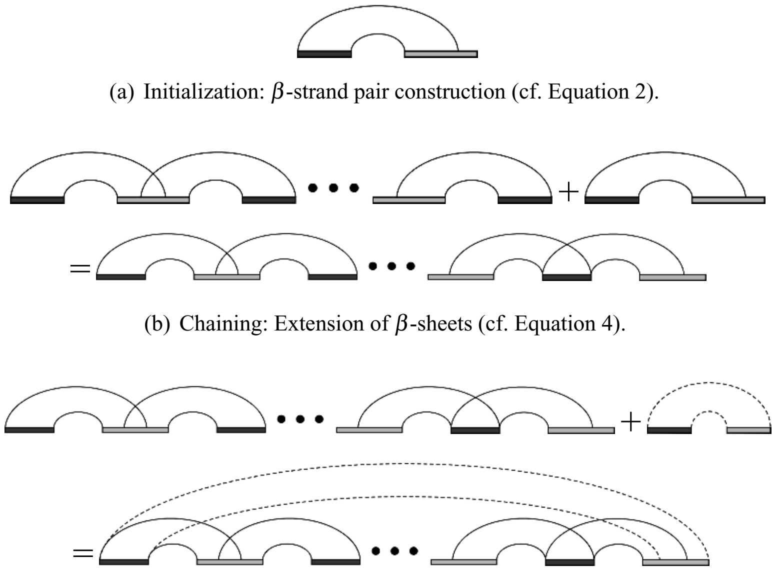
Once filled, we use this array to compute the partition function Q_{tmb} over all TMBs. Note that the index k can be used to control the number of strands in the barrels. This operation consists of adding the contributions of the anti-parallel β-strand pairs which close the extremities of the β-sheet. For this, we could use the values stored in the Boltzmann value array Q_{ap}; however, in practice, we use a special array which is better suited to the special rules for this last β-strand pair.[2]

[2] The rules for the closing pair, explicitly described in [75], mainly consist of relaxing some constraints, and allowing extensions on both sides of the strand.

$$Q_{tmb} = \sum_{(i_1,j_1)} \sum_{(i_N,j_N)} Q_{sheet}(i_1,j_1,i_N,j_N) \cdot Q_{ap}(i_N,j_N,i_1,j_1) \tag{5}$$

Note that in order to respect the pairwise orientation as well as strand inclination, the indices i_1, j_1 and i_N, j_N are swapped. Finally, it should be mentioned that in computing the partition function, the dynamic programming must ensure an exhaustive and non-overlapping count of all structures; in particular, the cases treated must be mutually exclusive, as is clearly the case in our algorithm.

We illustrate these equations and overview the complete procedure in Fig. 7. First, we initialize the dynamic arrays by computing all possible β-strand pairs (Fig. 7(a) and Equation 2). Then, we build the β-sheets by concatenating β-strand pairs (Fig. 7(b) and Equation 4). Finally, we close the TMBs by pairing the first and last β-strands (Fig. 7(c) and Equation 5)



(a) Initialization: β-strand pair construction (cf. Equation 2).

(b) Chaining: Extension of β-sheets (cf. Equation 4).

(c) Closing: add closing β-strand pair (cf. Equation 5).

Fig. 7 Graphical representations of the recursive rules used to enumerate the conformational landscape.

Using formulas from classical statistical mechanics, a number of important thermodynamic parameters can be computed immediately from the partition function. These parameters, including ensemble free energy, heat capacity, average internal energy, etc. (see [22]), lead to a better understanding of the folding landscape. For example, as shown in [19], the average internal energy of the structures $\langle E(s) \rangle$ can be computed by

$$\langle E(s) \rangle = RT^2 \cdot \frac{\partial}{\partial T} \log Q(s), \tag{6}$$

while the standard deviation can be computed with a similar formula. Such thermodynamic parameters provide information on the stability of folds for a given sequence.

4.2 Computing the residue contact probability

In this section, we address the problem of computing the Boltzmann pair probabilities from the dynamic tables filled when computing the partition function value Q_{tmb}. First, we need to characterize the anti-parallel β-strand pairs that contain a given contact.

Proposition 0.1. *Let x and y ($x < y$) be two residues of two distinct consecutive anti-parallel β-strands, and j_1 and i_2 (s.t. $i_1 \leq x \leq j_1 < i_2 \leq y \leq j_2$) the two residues at the extremities of the connecting loop. Then, residues (x,y) are brought into contact if and only if $i_2 + j_1 = x + y$.*

It follows from this proposition that (x,y) is a valid contact if and only if the anti-parallel β-strands $\binom{i_1,j_1}{i_2,j_2}$ verify $x+y = j_1 + i_2$ and $i_1 \leq x \leq j_1 < i_2 \leq y \leq j_2$.

To evaluate the residue pair probability $p(x,y)$, we must compute the partition function value over all TMBs $Q(x,y)$ which contain this contact. Such TMBs can be decomposed into two, three, or four parts, depending on the strand pair where the contact occurs (i.e. in the the closing strand pair, the first and last pair of the sheet or in an intermediate one). All these cases are illustrated in Figure 8.

Let $\binom{i,j}{i',j'}$ be an index of a block modeling an anti-parallel TM β-strand pair. Then, we define $Q^{close}\binom{i,j}{i',j'}$, $Q^{first}\binom{i,j}{i',j'}$, $Q^{last}\binom{i,j}{i',j'}$ and $Q^{inter}\binom{i,j}{i',j'}$ to be the partition functions over all TMB structures which contain this anti-parallel TM β-strand pair as, respectively, the pair closing the barrel (Figure 8(a)), the first pair of the TM β-sheet (Figure 8(b)), the last pair of the TM β-sheet (Figure 8(c)) or any other intermediate pair (Figure 8(d)). Formally:

$$Q^{close}\begin{pmatrix} i_1,j_1 \\ i_N,j_N \end{pmatrix} = Q_{\text{sheet}}\begin{pmatrix} i_1,j_1 \\ i_N,j_N \end{pmatrix} \cdot Q_{\text{ap}}\begin{pmatrix} i_N,j_N \\ i_1,j_1 \end{pmatrix} \tag{7}$$

$$Q^{first}\begin{pmatrix} i_1,j_1 \\ i_2,j_2 \end{pmatrix} = \sum_{(i_{N-1},j_{N-1})} Q_{\text{ap}}\begin{pmatrix} i_1,j_1 \\ i_2,j_2 \end{pmatrix} \cdot Q_{\text{sheet}}\begin{pmatrix} i_2,j_2 \\ i_N,j_N \end{pmatrix} \cdot Q_{\text{ap}}\begin{pmatrix} i_N,j_N \\ i_1,j_1 \end{pmatrix} \tag{8}$$

$$Q^{last}\begin{pmatrix} i_{N-1},j_{N-1} \\ i_N,j_N \end{pmatrix} = \sum_{(i_1,j_i)} Q_{\text{sheet}}\begin{pmatrix} i_1,j_1 \\ i_{N-1},j_{N-1} \end{pmatrix} \cdot Q_{\text{ap}}\begin{pmatrix} i_{N-1},j_{N-1} \\ i_N,j_N \end{pmatrix} \cdot Q_{\text{ap}}\begin{pmatrix} i_N,j_N \\ i_1,j_1 \end{pmatrix} \tag{9}$$

$$Q^{inter}\begin{pmatrix} i_k,j_k \\ i_{k+1},j_{k+1} \end{pmatrix} = \sum_{\substack{(i_1,j_1) \\ (i_N,j_N)}} Q_{\text{sheet}}\begin{pmatrix} i_1,j_1 \\ i_k,j_k \end{pmatrix} \cdot Q_{\text{ap}}\begin{pmatrix} i_k,j_k \\ i_{k+1},j_{k+1} \end{pmatrix} \cdot Q_{\text{sheet}}\begin{pmatrix} i_{k+1},j_{k+1} \\ i_N,j_N \end{pmatrix} \cdot Q_{\text{ap}}\begin{pmatrix} i_N,j_N \\ i_1,j_1 \end{pmatrix} \tag{10}$$

Finally, using these functions, the partition function $Q(x,y) = \sum_S e^{-\frac{E(S)}{RT}}$, where the sum is over all TMBs that contain the residue contact (x,y), is computed as follows:

$$Q(x,y) = \sum_{\substack{(i,j)\\(i',j')}}^{x+y=j+i'} \left(Q^{close}\begin{pmatrix} i,j \\ i',j' \end{pmatrix} + Q^{first}\begin{pmatrix} i,j \\ i',j' \end{pmatrix} + Q^{last}\begin{pmatrix} i,j \\ i',j' \end{pmatrix} + Q^{inter}\begin{pmatrix} i,j \\ i',j' \end{pmatrix} \right) \tag{11}$$

Finally, the Boltzmann probability $p(x,y)$ of a contact between the residues at indices x and y can be obtained by computing the value $p(x,y) = \frac{Q(x,y)}{Q_{tmb}}$. However, we note that an extra field counting the number of strands in Q^{sheet} is required to ensure that the minimal number of strands in a TMB is not violated.

Assuming the length of TM β-strands and loops, as well as the shear number values, are bounded, the time complexity is $\mathcal{O}(n^3)$, where n is the length of the input sequence. When the maximal length of a loop is in $\mathcal{O}(n)$, this complexity should approach $\mathcal{O}(n^4)$. Similarly, the space complexity can be bounded by $\mathcal{O}(n^2)$.

4.3 Improved computation of the contact probabilities

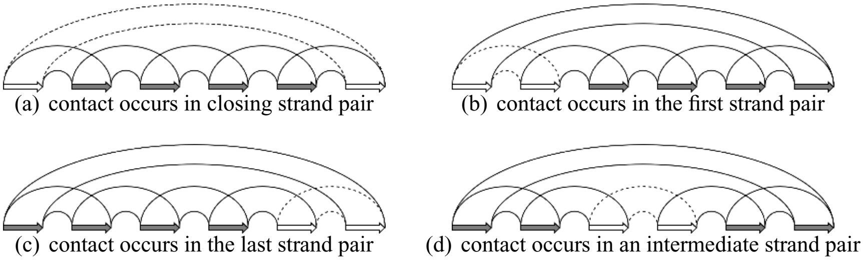
The formidable time requirement for a brute force algorithm to compute Equation 10 prevents any immediate efficient application. Indeed, naively applying this equation to the $\mathcal{O}(n^2)$ possible residue pairs results in an overall time complexity of $\mathcal{O}(n^5)$. In this section, we show how a simple strategy using additional dynamic tables, has been used to reduce the time complexity by a factor of $\mathcal{O}(n^2)$.

Two basic observations lead to a natural improvement over a brute force algorithm. First, when the TM β-strand pair that contains the residue contact is not involved, the product of the partition function of two sub-structures is realized over all possible configurations (i.e. $Q_u\binom{i,j}{i',j'} \cdot Q_v\binom{i',j'}{i'',j''}$ is computed over all possible pairs of indices (i',j')). In equation 10, the pairs of indices (i_k, j_k) and (i_{k+1}, j_{k+1}) are used for different residue contacts since the pair (i_N, j_N) varies. Thus we can precompute the values of $Q_{\text{sheet}}\binom{i_{k+1},j_{k+1}}{i_N,j_N} \cdot Q_{\text{ap}}\binom{i_N,j_N}{i_1,j_1}$ over all possible (i_N, j_N) and store them in a dynamic table for later retrieval. Given (i_1, j_1) and (i_{k+1}, j_{k+1}), let Q_{tail} be the array storing the values $\sum_{(i_N,j_N)} Q_{\text{sheet}}\binom{i_{k+1},j_{k+1}}{i_N,j_N} \cdot Q_{\text{ap}}\binom{i_N,j_N}{i_1,j_1}$. This table can be filled in time $\mathcal{O}(n^3)$. Then, in place of equation 10, we now have equation 12.

$$Q^{inter}\begin{pmatrix} i_k, j_k \\ i_{k+1}, j_{k+1} \end{pmatrix} = \sum_{(i_1,i_2)} Q_{\text{sheet}}\begin{pmatrix} i_1, j_1 \\ i_k, j_k \end{pmatrix} \cdot Q_{\text{ap}}\begin{pmatrix} i_k, j_k \\ i_{k+1}, j_{k+1} \end{pmatrix} \cdot Q_{\text{tail}}\begin{pmatrix} i_{k+1}, j_{k+1} \\ i_1, j_1 \end{pmatrix} \tag{12}$$

This improvement cannot be applied to Equations 8 and 9, since there is no redundancy in those cases. The time complexity for computing all possible contact probabilities $p(i,j)$ is now $\mathscr{O}(n^4)$. However, further observation allows us to save an additional factor of $\mathscr{O}(n)$ in the time complexity: when a TMB structure is considered in one of the equations 7, 8, 9 or 10, the TM β-strand pair which contains the contact (x,y) also involves many other contacts. Hence, instead of using these equations to compute the values $Q(x,y)$ and $p(x,y)$ separately, we consider each possible β-strand pair and immediately add its contribution to the partition function. From these improvements, we now have an algorithm to compute all the contact probabilities of a TMB, which runs in time $\mathscr{O}(n^3)$.

Although not explicitly mentioned thus far, we should emphasize that we can also compute the contact probability $p_M(x,y)$ for a specific environment M — i.e. membrane or channel (see Section 3 for an explanation of environment). To do so, we simply need to duplicate the dynamic tables in order to take into account the side-chain orientation for extremal TM β-strand pairs.


(a) contact occurs in closing strand pair

(b) contact occurs in the first strand pair

(c) contact occurs in the last strand pair

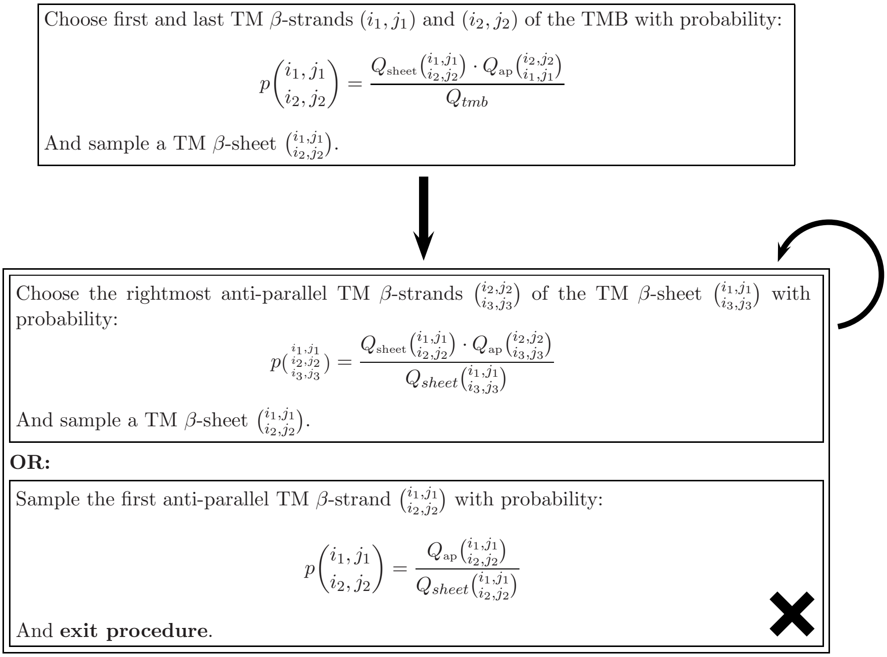
(d) contact occurs in an intermediate strand pair

Fig. 8 Decompositions of a transmembrane β-barrel, which allows us to isolate the antiparallel TM β-strand pair that contains the residue contact. The block that corresponds to this strand pair is indicated with white β-strands connected with dashed lines. The blocks in gray represent TM β-sheets (i.e. a sequence of anti-parallel TM β-strands).

4.4 Rigorous sampling of transmembrane β-barrels

In this section, we describe a rigorous sampling algorithm for TMBs. Given an amino acid sequence ω, it randomly generates, according to the distribution of structures in the Boltzmann ensemble, low energy TMB structures for ω. By sampling, we expect to be able to efficiently estimate non-trivial features concerning the ensemble of potential TMB folds, with the long-term goal of potentially contributing to drug design engineering.

The sampling algorithm uses the dynamic table filled during the computation of the partition function. It essentially proceeds in two steps illustrated in Figure 9. First, the "closing" anti-parallel strand pair is sampled according to the weight of all TMBs that contain it over all possible TMBs. Then, we sample each anti-parallel strand pair of the TM β-sheet from left to right (or alternatively from right to left) until the last one, according to the weight of that structure over all possible TM β-sheets. The full procedure is depicted in Figure 9. The correctness of the algorithm is ensured by construction of the dynamic table in Equations 4 and 5.



Choose first and last TM β-strands (i_1, j_1) and (i_2, j_2) of the TMB with probability:

$$p\binom{i_1, j_1}{i_2, j_2} = \frac{Q_{\text{sheet}}\binom{i_1,j_1}{i_2,j_2} \cdot Q_{\text{ap}}\binom{i_2,j_2}{i_1,j_1}}{Q_{tmb}}$$

And sample a TM β-sheet $\binom{i_1,j_1}{i_2,j_2}$.

Choose the rightmost anti-parallel TM β-strands $\binom{i_2,j_2}{i_3,j_3}$ of the TM β-sheet $\binom{i_1,j_1}{i_3,j_3}$ with probability:

$$p\left(\substack{i_1,j_1\\ i_2,j_2\\ i_3,j_3}\right) = \frac{Q_{\text{sheet}}\binom{i_1,j_1}{i_2,j_2} \cdot Q_{\text{ap}}\binom{i_2,j_2}{i_3,j_3}}{Q_{sheet}\binom{i_1,j_1}{i_3,j_3}}$$

And sample a TM β-sheet $\binom{i_1,j_1}{i_2,j_2}$.

OR:

Sample the first anti-parallel TM β-strand $\binom{i_1,j_1}{i_2,j_2}$ with probability:

$$p\binom{i_1, j_1}{i_2, j_2} = \frac{Q_{\text{ap}}\binom{i_1,j_1}{i_2,j_2}}{Q_{sheet}\binom{i_1,j_1}{i_2,j_2}}$$

And **exit procedure**.

Fig. 9 Sampling procedure: The first and last TM β-strands of the barrel are sampled (left box). Then the remaining TM β-sheet is sampled by iteratively sampling the rightmost anti-parallel β strand of the remaining sequence, until the first β-strand pair of the sheet is sampled.

5 Applications

The algorithms described in the previous section are implemented in the program *partiFold* [77]. The *partiFold* algorithms use the Boltzmann partition function to predict the ensemble of structural conformations a TMB may assume instead of predicting a single minimum energy structure. From this ensemble, experimentally testable TMB properties are computed that describe the folding landscape and suggest new hypotheses. In the following, we illustrate the flexibility of the approach and show how the method can be used for predicting individual contacts, investigating the conformational landscape and predicting Debye-Waller factors (a X-ray crystallography measure accounting for the thermal motion of the atom - a.k.a. B-factors). We finally apply whole structure sampling to demonstrate the benefits of ensemble modeling over single structure prediction and the possibilities for structural exploration provided by these techniques.

5.1 Residue contact prediction

Single contact prediction remains an important concern when reconstructing 3D models [33, 43, 57]. Several machine-learning methods have been developed for this task, among them PROFcon [57] and FOLDpro [15] (general predictors), BETApro [14] (specialized for β-structures) and TMBpro [59] (specialized for TMBs) are among the most reliable. However, it should be noted that, while some of them can provide stochastic contact map of β-strand interactions, the interaction probabilities are not related to a Boltzmann distribution of conformations, but rather based on sophisticated neural networks and graph algorithms that aim to predict a single structure. In addition, their energy models also do not appear to be common across all proteins, resulting in difficulties to interpret and compare results between different proteins.

But even conceptually, the ensemble approach is radically different from previous machine-learning methods. Indeed, while the latter first start by making individual and unrelated contact predictions and finish by reconstructing a whole structure, the ensemble method does not dissociate both aspects, since the set of TMB structures is computed first and the contact probabilities are subsequently evaluated from the folding energies of these structures.

To test the ensemble method, single contact predictions are made by selecting all pairwise contacts that have a probability greater than a given threshold p_t in the stochastic contact map, and compare those against the corresponding contacts found in X-ray crystal structures as annotated by STRIDE [31].

To evaluate the contact predictions, we classically rely on three standard measures: the sensitivity (or coverage), where

$$\text{sensitivity} = \frac{\text{number of correctly predicted contacts}}{\text{number of observed contacts}},$$

the positive predictive value (abbreviated PPV and also known as accuracy), where

$$\text{PPV} = \frac{\text{number of correctly predicted contacts}}{\text{number of predicted contacts}},$$

and the F-measure, where

$$\text{F-measure} = \frac{2 \cdot \text{sensitivity} \cdot \text{PPV}}{\text{Sensitivity} + \text{PPV}}.$$

To demonstrate how these metrics would apply to this type of contact prediction, we refer to Figure 10, which depicts the accuracy of contact prediction for the crystal structure of outer membrane protein X (abbreviated OmpX) [73]. The flatness of the curves further indicates a good separation between accurate, highly probable contacts, and background predictive noise. This type of result could suggest a good scaffold of likely contacts when constructing a 3D model of an unknown structure.

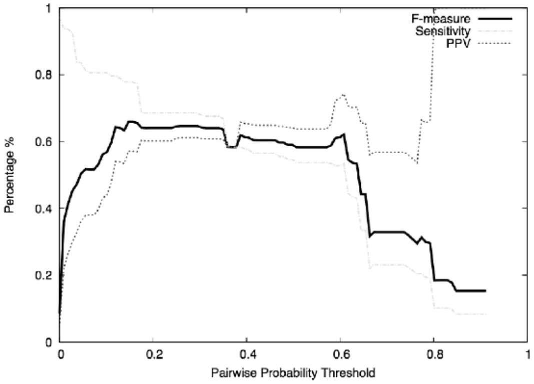


Fig. 10 Predicting residue contact probabilities in OmpX (1QJ8 [73]). The x-axis represents the threshold used to select the residue contact predictions. The graph shows the curves of the F-measure, sensitivity (or coverage) and positive predictive value (or accuracy, abbreviated PPV).

These techniques have proven to provide the state-of-the-art predictions for TMBs [77]. However, there is still room for improvement. For instance, current algorithms do not yet model bulges in β-sheets and suffer slightly in performance where bulges exist.

5.2 Representations of ensembles

The class of predictions enabled by these techniques embody whole-ensemble properties of a protein. The contact probabilities can be treated all together to represent and analyze different aspects of the folding properties of a polypeptide.

In Fig. 11, a single structure is chosen (in this case the X-ray structure of OmpX [73]), and displayed as an unrolled 2D representation of the β-barrel strands and their adjacent residue contacts. Using the stochastic contact map, residue contact pairs are then colored to indicate a high (black), a medium (dark gray) or a low (light gray) probability in the Boltzmann distributed ensemble. From this, substructures may be analyzed from their relative likelihood of pairing.

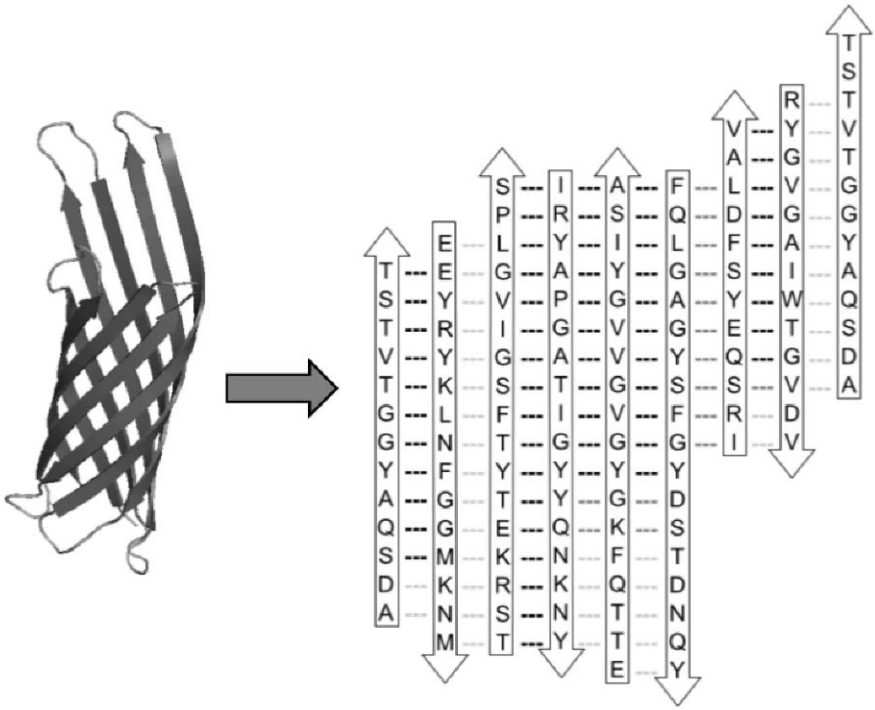


Fig. 11 Contact probabilities mapping to OmpX (1QJ8) X-ray crystal structure [73]. 2D representation (unrolled β-barrel) showing only those residues involved in β-strands (shown vertically and successively numbered) and their associated, in-register H-bonding partners. Computed contact probabilities indicated by color hue (highly probable in black, medium probability in dark gray and low probability in light gray). The leftmost β-strand is repeated on the right to allow the barrel to close.

In Fig. 12, the inter-strand residue contact probabilities are merged in the upper triangle of a single matrix called a *stochastic contact map*. This reflects the likelihood of two β-strand amino acids pairing in the (estimated) Boltzmann distribution of conformations, and not one single minimum folding energy structure. This graphical representation provides an intuitive way to depict the variety of structures that can be found in the conformational landscape. We can also compare with these maps the contacts of a given structure (in this case, the contact found in the X-ray structure are plotted in the lower triangle) to estimate its adequacy with the conformational landscape suggested by the high contact probabilities (gray regions of the stochastic contact maps).

While the mapping allows analysis of the likelihood of a given structure from the ensemble perspective, the stochastic contact maps enable us to investigate the folding landscape and estimate the variety of folds of a given polypeptide.

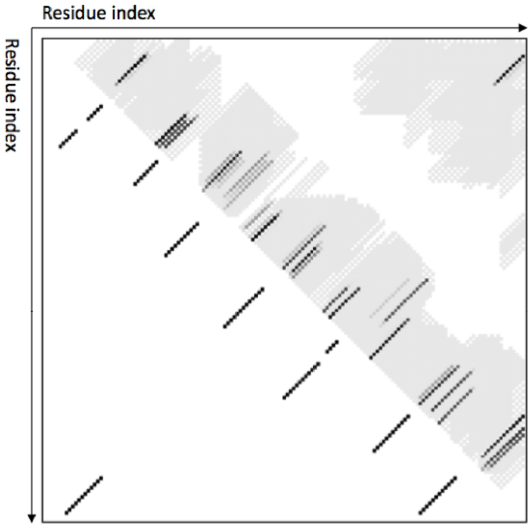


Fig. 12 Stochastic contact map for Neisserial Surface Protein A (NspA). Horizontal and vertical axes represent residue indices in sequence (indices 1 to 155 from left to right and top to bottom), and points on the map at location (i, j) in the upper triangle represent the probability of contact between residues i and j (where darker gray implies a higher probability). The X-ray crystal structure contacts of 1P4T [71] are shown in the lower triangle.

A striking example of the biological relevance of these techniques is shown in Fig. 13. The stochastic contact map of the Outer Membrane Enzyme PagP protein is computed. It contains the contacts found on the X-ray crystal structure 1THQ [2] (in black in the lower triangle), and those of the minimum folding energy structure (in gray in the lower triangle). Here, we note that (i) it is clear that the native conformation (black, lower triangle) differs radically from the minimum energy structure (gray, lower triangle), and (ii) the stochastic contact map reveals alternate β-strand pairs with high probabilities (in gray in the upper triangle).

These discrepancies may be explained through the lens of a recent experimental study. Indeed, Huysmans *et al.* [41] showed that the N-terminal α-helix found in the native structure is essential for the stability of the native β-barrel structure. If we constrain the corresponding α-helical regions of the contact map to not fold into a barrel (peach regions in Fig. 13), this prevents the protein from folding as the minimum free energy structure and thus allows it to adopt one of the other conformations suggested by the stochastic contact map (gray regions in Fig. 13), which coincides with the native structure (in black in the lower triangle). This example illustrates how the contact maps can suggest alternate folds.

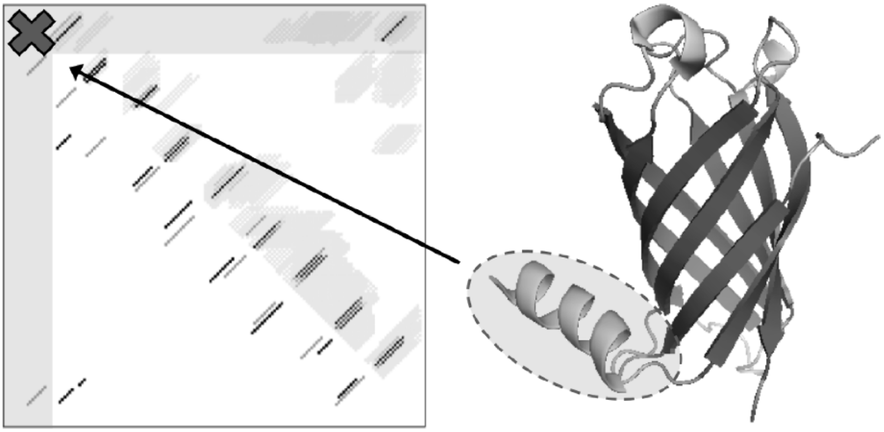


Fig. 13 Multiple conformations of PagP proteins. The stochastic contact map (gray regions in the upper triangle) is compared with the contacts found on the X-ray crystal structure 1THQ [2] (in black in the lower triangle), and those of the minimum folding energy structure (in gray in the lower triangle). If we constrain the corresponding α-helical regions of the contact map to not fold into a barrel (gray stripes), this prevents the protein from folding as the minimum free energy structure and thus allows it to adopt one of the other conformations suggested by the stochastic contact map (gray regions in the upper triangle), which coincides with the native structure (in black in the lower triangle).

5.3 Prediction of residue flexibility

We now show how the contact probabilities can be used to predict per-residue flexibility and entropy. To a first approximation, this flexibility correlates with the *Debye-Waller factor* (a.k.a. B-factor) found in X-ray crystal structures [60]. This demonstrates an important purpose for computing the Boltzmann partition function: to provide biologically-relevant grounds for the prediction of experimentally testable macroscopic and microscopic properties.

Predicting residue B-factors is important because it roughly approximates the local mobility of flexible regions, which might be associated with various biological processes, such as molecular recognition or catalytic activity [62]. In this context, flexible regions are strong candidates for loop regions connecting anti-parallel TM β-strands that extend either into the extracellular or intracellular milieu.

Classical B-factor predictors use machine learning approaches [62]. However, as is the case for contact predictions (cf. Section 5.1) these techniques do not provide a comprehensive framework facilitating the understanding of these results in a larger context. Indeed, previous methods were specifically designed to make only these predictions, while in ensemble approaches, B-factors are just one of the multiple characteristics that can be extracted from an ensemble model.

We define the *contact probability profile* of every amino acid index i in a TM β-barrel to be $P_c(i) = 2 - \sum_{j=1}^{n} p_{i,j}$, and compare this against the normalized B-factor.

Since a residue may be involved in two contacts in a β-sheet, the value of $P_c(i)$ can range between 0 and 2, where higher values indicate greater flexibility. Similarly, residues with a positive B-factor are considered flexible or disordered, while others are considered rigid. In Figure 14, we illustrate this method by comparing the curves of X-ray B-factors and contact profiles of OmpX (1QJ8) [73] and NspA (1P4T) [71] proteins.

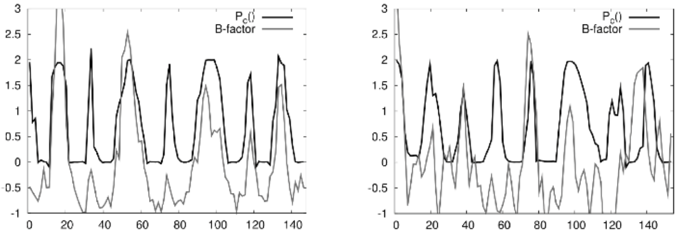


Fig. 14 Comparison of B-factors and contact probability profiles of OmpX (1QJ8) [73], left, and NspA (1P4T) [71], right, proteins.

Computing the cross-correlation coefficient between the P_c and B-factor of test proteins reveals that this method provides state-of-the-art predictions [77]. But the real purpose of this work actually goes much beyond that. The direct predictions of experimental measures are of fundamental importance. It enables biologists to directly compare computational predictions to experimental measures and avoid any misleading interpretations. These methods also can be efficiently used to tune the theoretical folding model to fit experimental data.

5.4 Whole structure prediction through Boltzmann sampling

Finally, we show how ensembles of structures can characterize protein structure better than the minimum folding energy (m.f.e.) structure. We perform stochastic conformational sampling (cf. Section 4.4) to map the landscape defined by the Boltzmann partition function. This also illustrates how the approach can be used to rigorously explore the space of all possible TMB structures. By clustering a large set of full TMB structure predictions, a small distinguishable collection of unique conformations are exposed.

Waldispühl *et al.* [77] sampled $1,000$ TMB structures and grouped them into 10 clusters according to hierarchical clustering. Similar to prior methods developed for RNAs [25], for each cluster one can designate a centroid representative conformation that is chosen as the structure with the minimum total distance to all other structures in the set. To facilitate this clustering, a metric named *contact distance* is introduced: $d_c(S_1,S_2) = |\mathscr{C}_1| + |\mathscr{C}_2| - 2 \cdot |\{\mathscr{C}_1 \cap \mathscr{C}_2\}|$, where $\mathscr{C}_1$ and $\mathscr{C}_2$ are the sets of contacts in S_1 and S_2 (which represents the minimum number of contacts to be removed and added to pass from S_1 to S_2 or vice versa). Other metrics could be defined but the latter seemed to provide the best results.

The results showed that the centroid of the largest cluster usually provides a better solution than the minimum folding energy structure [77]. It has also been found than in some cases a centroid of another cluster provides significantly better structure predictions. However, the identification of the "best" cluster, as well as the the robustness of the clusters to the distance used, remains to be investigated.

6 Sampling the local neighborhood of 3D structures

We conclude this chapter by addressing a different but related problem. Instead of sampling the global folding landscape of a protein sequence of unknown structure, we aim to sample the local neighborhood of a given 3D structure. In other words, we seek to estimate the stability of a structure and explore the variations of specific folds at a precision not achieved by *partiFold*.

Unlike the methods described in previous sections, we no longer restrict our conformational space to TMBs. In the following, we overview the principal aspect of this approach. In Section 6.1, we introduce a structural modeling approach that suits the problem well. In Section 6.2, we describe the sampling procedure. Finally, in Section 6.3, we give an application of this approach.

6.1 Structure modeling

Since we aim to sample in the local neighborhood of a given 3D structure, the size of the explored conformational landscape is drastically smaller than in previous sections. Thus, we can afford to use a more detailed description of the structure. Of the many different representations of protein structure, one that has gained popularity is that of the torsion angle representation. This representation makes use of the fact that bond lengths and bond angles show little variation across structures [34], and hence can be assumed to be fixed (cf. Fig. 15). The flexibility of protein molecules can thus almost entirely be described by rotation about covalent bonds.

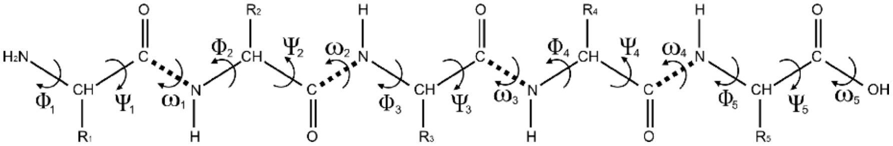


Fig. 15 Torsion angle representation of a polypeptide.

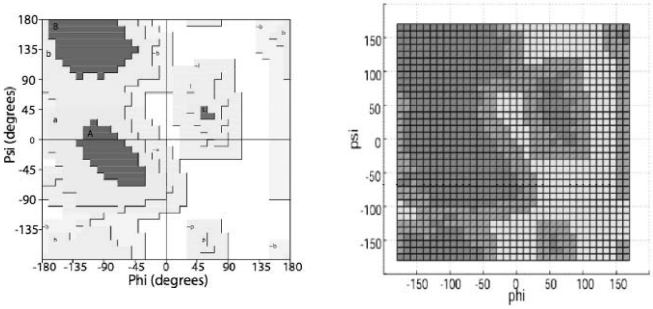
Such a simplified model has the advantage of not having to enforce regular geometry constraints. But, at the same time, non-bonded interactions are non-trivial to calculate in this reduced representation.

6.2 Sampling in the Torsion space

One of the biggest advantages of using this reduced model of a protein is the speed with which one is able to sample the conformation space. By discretizing the dihedral angle space (i.e. the Ramachandran plot [58], see Fig. 16(a)) and biasing solutions lying within specified regions, one can sample protein conformations in the neighborhood of a native structure (see Fig. 16(b)). By efficiently exploiting various algorithms developed in the Inverse Kinematics community, the algorithm ChainTweak [67] was shown to be capable of exploring a much larger conformational space than previous methods [10, 12, 35, 39, 47, 55, 70].

ChainTweak iteratively perturbs the base conformation using the torsion (a.k.a. dihedral angle) representation. A sliding window approach is used to successively move some atoms by 0-2 Å, while keeping all others fixed (see Fig. 17(a)). Inside the window, loop closure methods are used to generate such perturbations [20, 49, 80]. Moreover, residue specific Phi-Psi angle preferences, given by a Phi-Psi priority scheme (Fig. 16(b)) inherited from a Ramachandran plot [58] (Fig. 16(a)), can be used to choose a perturbation. The loop closure problem was informally discussed by Robert Diamond and M. Levitt and formally defined by Go and Scheraga [32]. The input to such a problem is the relative position of two fixed residues (anchors) at each end and the goal is to find different possible conformations for a polypeptide chain of length m joining the fixed ends.

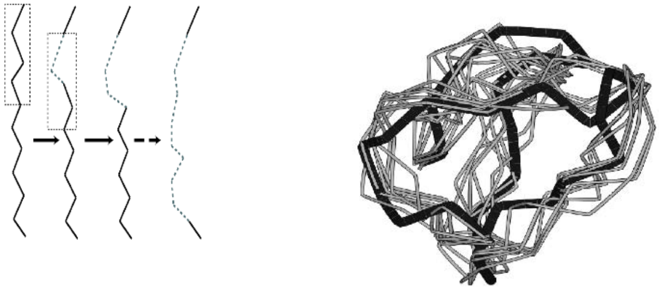
Rather than being closely tied to some search strategy (or an energy function), ChainTweak is a stand-alone method that can be used by researchers as a black-box, allowing them to focus on other parts of the search problem (e.g., energy function design [48]). Unlike classical molecular dynamic simulations, which are constrained by folding trajectories, it allows an unbiased sampling of the conformational landscape in the neighborhood of a given structure.



(a) Ramachandran plot (b) Phi-Psi priority scheme

Fig. 16 (a) Reference Ramachandran plot and (b) phi-psi priority scheme: (dark gray: highest priority)>(medium gray: medium priority)>(light gray: lowest priority) [67].

We show in Fig. 17(b) an application of this program and sample ten conformations from the neighborhood of a 32-residue protein structure (PDB:1CLV, chain I [56]). Using the LoopClsr [67] algorithm iteratively (Fig. 17(a)) on the backbone of a protein, we generated conformations in the neighborhood of this structure [1-4Å] within a few seconds. The size of the neighborhood explored and topology of the backbone can be constrained depending on the context of the simulations. ChainTweak is purely geometric in nature and does not inherently depend on any energy function. This makes it a useful standalone sampling algorithm, which can then be combined with existing energy functions. Such a methodology completely eliminates the dependence of sampling on the limitations of the energy function.



(a) Sampling procedure (b) Example output

Fig. 17 (a) Iteratively modifying the backbone of a protein. Sliding window formulation implemented in ChainTweak [67]. (b) Example output from ChainTweak. Ten conformations from the neighborhood of a 32-residue protein structure (PDB:1CLV, chain I [56]) were sampled and aligned with the original. The original structure is in black, the others are in gray [67].

6.3 Structure Determination

Structure prediction and determination are still significant bottlenecks to the goals of the Structural Genomics initiative [9]. Due to great advances in sequencing technologies and algorithms for analyzing sequence data, the gap between the number of genomes known and number of structures known is even increasing. In order to close this gap, significant advances need to be made in two areas: 1) accurately and efficiently determine structures from incomplete experimental data [9], and 2) develop accurate energy functions that can filter out native structures from a set of decoys. The first problem can be set up as an optimization problem, for which the sampling algorithm is critical for exploring diverse structures and thus maximizing the likelihood of the observed data. Furthermore, as we move away from a “static” picture of a protein to a more dynamic one, the ability to exhaustively sample all degrees of freedom becomes critical to our understanding of the structure. ChainTweak is ideally suited for such an analysis, as we have demonstrated by modeling the heterogeneity in crystal structures solved at medium to low resolutions [38]. More importantly from a biological standpoint, such an analysis potentially provides a mechanism for understanding the protein structure-function relationship [45].

7 Exercises

1. In Section 4, we add an energy term $\mathscr{L}(n)$ to the loop connecting the β-strand pairs, where n is the number of residues in the loop. Modify the recursive equations of Section 4 to account for this change.
2. Write a backtracking algorithm for computing the Boltzmann probability of closing β-strand pairs of length n (i.e. n inter-strand residue contacts).
3. We assume that the sampling algorithm in Section 4 returns the energy $E(S)$ of the sampled TMB structure S. Let $\mathscr{Z}$ be the partition function value. Then, the Boltzmann probability of a structure S is $P(S) = \frac{e^{-E(S)/RT}}{\mathscr{Z}}$. Write an iterative procedure for sampling TMBs until a ratio ρ $(0 < \rho \leq 1)$ of the folding landscape has been covered.
4. Write pseudo-code for the ChainTweak algorithm illustrated in Fig. 17(a).

8 Further reading

Many studies provide reliable techniques for sampling structures from sequence. The popular Rosetta [66] uses multiple sequence alignments to select small protein backbone fragments that are assembled together using a simulated annealing procedure. More recently, novel approaches have been proposed for overcoming the difficulty of designing a reliable energy function required to perform a simulated annealing procedure. Indeed, lattice-based techniques [82], HMMs [36], and more general Conditional Random Fields (CRFs) [84, 85] have been successfully applied for this purpose. However, it is worth noting that the methods cited here generate decoys but do not sample from a rigorously defined distribution of structures.

To conclude this chapter, we note that the program *partiFold* detailed in Section 4 has recently been extended to perform structural *ensemble* comparisons [78] and generate accurate sequence alignments of proteins with low sequence identity.

Acknowledgements The authors would like to thank Charles W. O'Donnell and Raghavendra Hosur for useful comments and suggestions.

References

1. Abe, N., Mamitsuka, H.: Predicting protein secondary structure using stochastic tree grammars. Machine Learning **29**(2-3), 275–301 (1997)
2. Ahn, V.E., Lo, E.I., Engel, C.K., Chen, L., Hwang, P.M., Kay, L.E., Bishop, R.E., Prive, G.G.: A hydrocarbon ruler measures palmitate in the enzymatic acylation of endotoxin. EMBO J **23**(15), 2931–2941 (2004 Aug 4)
3. Amato, N., Dill, K., Song, G.: 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)
4. Bartlett, A.I., Radford, S.E.: An expanding arsenal of experimental methods yields an explosion of insights into protein folding mechanisms. Nat Struct Mol Biol **16**(6), 582–588 (2009)
5. Bayrhuber, M., Meins, T., Habeck, M., Becker, S., Giller, K., Villinger, S., Vonrhein, C., Griesinger, C., Zweckstetter, M., Zeth, K.: Structure of the human voltage-dependent anion channel. Proc Natl Acad Sci U S A **105**(40), 15,370–15,375 (2008 Oct 7)
6. Berg, O.G., von Hippel, P.H.: Selection of DNA binding sites by regulatory proteins. statistical-mechanical theory and application to operators and promoters. J Mol Biol **193**(4), 723–750 (1987 Feb 20)
7. Berger, B., Leighton, T.: Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. J Comput Biol **5**(1), 27–40 (1998)
8. Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., Bourne, P.: The Protein Data Bank. Nucleic Acids Research **28**, 235–242 (2000)
9. Bourne, P., Weissig, H.: Structural Bioinformatics. Wiley-Liss (2003)
10. Bradley, P., Chivian, D., Meiler, J., Misura, K.M.S., Rohl, C.A., Schief, W.R., Wedemeyer, W.J., Schueler-Furman, O., Murphy, P., Schonbrun, J., Strauss, C.E.M., Baker, D.: Rosetta predictions in CASP5: Successes, failures, and prospects for complete automation. Proteins **53 Suppl 6**, 457–468 (2003)
11. Bradley, P., Cowen, L., Menke, M., King, J., Berger, B.: Betawrap: Successful prediction of parallel beta-helices from primary sequence reveals an association with many microbial pathogens. Proceedings of the National Academy of Sciences **98**(26), 14,819–14,824 (2001)
12. Cahill, M., Cahill, S., Cahill, K.: Proteins wriggle. Biophys J **82**(5), 2665–2670 (2002 May)
13. Chandler, D.: Introduction to Modern Statistical Mechanics. Oxford University Press (1987)
14. Cheng, J., Baldi, P.: Three-stage prediction of protein beta-sheets by neural networks, alignments and graph algorithms. Bioinformatics **21 Suppl 1**, i75–84 (2005 Jun)
15. Cheng, J., Baldi, P.: Improved residue contact prediction using support vector machines and a large feature set. BMC Bioinformatics **8**, 113 (2007)
16. Chiang, D., Joshi, A.K., Dill, K.: A grammatical theory for the conformational changes of simple helix bundles. Journal of Computational Biology **13**(1), 27–42 (2006)

17. Chotia, C.: The nature of the accessible and buried surfaces in proteins. J Mol. Biol. **105**(1), 1–14 (1975)
18. Clote, P., Backofen, R.: Computational Molecular Biology: An Introduction. John Wiley & Sons (2000). 279 pages
19. Clote, P., Waldispühl, J., Behzadi, B., Steyaert, J.M.: Energy landscape of *k*-point mutants of an RNA molecule. Bioinformatics **21**(22), 4140–4147 (2005)
20. Coutsias, E.A., Seok, C., Jacobson, M.P., Dill, K.A.: A kinematic view of loop closure. J Comput Chem **25**(4), 510–528 (2004)
21. Cowen, L., Bradley, P., Menke, M., King, J., Berger, B.: Predicting the beta-helix fold from protein sequence data. J of Computational Biology **9**, 261–276 (2002)
22. Dill, K., Bromberg, S.: Molecular Driving Forces. Garland Science, Taylor & Francis (2003). New York
23. Dill, K., Phillips, A., Rosen, J.: Protein structure and energy landscape dependence on sequence using a continuous energy function. J Comput Biol. **4**(3), 227–39 (1997)
24. Dill, K.A., Ozkan, S.B., Shell, M.S., Weikl, T.R.: The protein folding problem. Annu Rev Biophys **37**, 289–316 (2008)
25. Ding, Y., Lawrence, C.: A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res. **31(24)**, 7280–7301 (2003)
26. Dobson, C.M.: Protein folding and misfolding. Nature **426**(6968), 884–890 (2003)
27. Dyson, H.J., Wright, P.E.: Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol **6**(3), 197–208 (2005 Mar)
28. Fain, B., Levitt, M.: A novel method for sampling alpha-helical protein backbones. J Mol Biol. **305**(2), 191–201 (2001)
29. Fain, B., Levitt, M.: Funnel sculpting for in silico assembly of secondary structure elements of proteins. Proc. Natl. Acad. Sci. USA **100**(19), 10,700–5 (2003)
30. Foat, B.C., Morozov, A.V., Bussemaker, H.J.: Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. Bioinformatics **22**(14), e141–9 (2006 Jul 15)
31. Frishman, D., P., A.: Knowledge-based protein secondary structure assignment. Proteins **23**, 566–579 (1995)
32. Go, N., Scheraga, H.A.: Ring closure and local conformational deformations of chain molecules. Macromolecules **3**(2), 178–187 (1970)
33. Grana, O., Baker, D., MacCallum, R., Meiler, J., Punta, M., Rost B.and Tress, M., Valencia, A.: CASP6 assessment of contact prediction. Proteins **61**(7), 214–224 (2005)
34. Grosberg, A., Khokhlov, A.: Statistical Physics of Macromolecules. AIP Press (1994)
35. Guntert, P., Mumenthaler, C., Wuthrich, K.: Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol **273**(1), 283–298 (1997 Oct 17)
36. Hamelryck, T., Kent, J.T., Krogh, A.: Sampling realistic protein conformations using local structural bias. PLoS Comput Biol **2**(9), e131 (2006 Sep 22)
37. Hockenmaier, J., Joshi, A., Dill., K.: Routes are trees: The parsing perspective on protein folding. PROTEINS: Structure, Function, and Bioinformatics **66**, 1–15 (2007)
38. Hosur, R., Singh, R., Berger, B.: Personal communication
39. Huang, E.S., Subbiah, S., Tsai, J., Levitt, M.: Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations. J Mol Biol **257**(3), 716–725 (1996 Apr 5)
40. Hubner, I.A., Deeds, E.J., Shakhnovich, E.I.: Understanding ensemble protein folding at atomic detail. Proc Natl Acad Sci U S A **103**(47), 17,747–17,752 (2006 Nov 21)
41. Huysmans, G.H.M., Radford, S.E., Brockwell, D.J., Baldwin, S.A.: The N-terminal helix is a post-assembly clamp in the bacterial outer membrane protein PagP. J Mol Biol **373**(3), 529–540 (2007 Oct 26)
42. Istrail, I.: Statistical mechanics, three-dimensionality and NP-completeness: I. Universality of intractability of the partition functions of the Ising model across non-planar lattices. In: A. Press (ed.) Proceedings of the 32nd ACM Symposium on the Theory of Computing (STOC00), pp. 87–96 (2000)

43. Izarzugaza, J.M.G., Grana, O., Tress, M.L., Valencia, A., Clarke, N.D.: Assessment of intramolecular contact predictions for CASP7. Proteins **69 Suppl 8**, 152–158 (2007)
44. King, J., Haase-Pettingell, C., Gossard, D.: Protein folding and misfolding. American Scientist **90**(5), 445–453 (2002)
45. Knight, J.L., Zhou, Z., Gallicchio, E., Himmel, D.M., Friesner, R.A., Arnold, E., Levy, R.M.: Exploring structural variability in X-ray crystallographic models using protein local optimization by torsion-angle sampling. Acta Crystallogr D Biol Crystallogr **64**(Pt 4), 383–396 (2008 Apr)
46. Koebnik, R.: Membrane assembly of the *Escherichia coli* outer membrane protein OmpA: Exploring sequence constraints on transmembrane β-strands. J. Mol. Biol. **285**, 1801–1810 (1999)
47. Kolodny, R., Koehl, P., Guibas, L., Levitt, M.: Small libraries of protein fragments model native protein structures accurately. J Mol Biol **323**(2), 297–307 (2002 Oct 18)
48. Krishnamoorthy, B., Tropsha, A.: Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. Bioinformatics **19**(12), 1540–1548 (2003 Aug 12)
49. Manocha, D., Zhu, Y., Wright, W.: Conformational analysis of molecular chains using nanokinematics. Comput Appl Biosci **11**(1), 71–86 (1995)
50. McCaskill, J.: The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers **29**, 1105–1119 (1990)
51. McDonnell, A.V., Menke, M., Palmer, N., King, J., Cowen, L., Berger, B.: Fold recognition and accurate sequence-structure alignment of sequences directing beta-sheet proteins. Proteins **63**(4), 976–985 (2006 Jun 1)
52. Miller, D., Dill, K.: Ligand binding to proteins: the binding landscape model. Protein Sci. **6**(10), 2166–79 (1997)
53. Mirny, L., Shakhnovich, E.: Protein folding theory: from lattice to all-atom models. Annu Rev Biophys Biomol Struct. **30**, 361–96 (2001)
54. Morozov, A.V., Havranek, J.J., Baker, D., Siggia, E.D.: Protein-DNA binding specificity predictions with structural models. Nucleic Acids Res **33**(18), 5781–5798 (2005)
55. Park, B., Levitt, M.: Energy functions that discriminate X-ray and near native folds from well-constructed decoys. J Mol Biol **258**(2), 367–392 (1996 May 3)
56. Pereira, P.J., Lozanov, V., Patthy, A., Huber, R., Bode, W., Pongor, S., Strobl, S.: Specific inhibition of insect alpha-amylases: yellow meal worm alpha-amylase in complex with the amaranth alpha-amylase inhibitor at 2.0 A resolution. Structure **7**(9), 1079–1088 (1999 Sep 15)
57. Punta, B., Rost, B.: Profcon: novel prediction of long-range contacts. Bioinformatics **21**(13), 2960–2968 (2005)
58. Ramachandran, G., Sasisekharan, V.: Conformation of polypeptides and proteins. Adv. Protein. Chem. **23**, 283–437 (1968)
59. Randall, A., Cheng, J., Sweredoski, M., Baldi, P.: TMBpro: secondary structure, beta-contact and tertiary structure prediction of transmembrane beta-barrel proteins. Bioinformatics **24**(4), 513–520 (2008 Feb 15)
60. Rhodes, G.: Crystallography Made Crystal Clear, 2nd edn. Academic Press: San Diego (2000)
61. Rumbley, J., Hoang, L., Mayne, L., Englander, S.W.: An amino acid code for protein folding. Proc Natl Acad Sci U S A **98**(1), 105–112 (2001)
62. Schlessinger, A., Rost, B.: Protein flexibility and rigidity predicted from sequence. Proteins **61**(1), 115–126 (2005)
63. Schultz, C.: Illuminating folding intermediates. Nature Structural Biology **7**, 7–10 (2000)
64. Schulz, G.: β-barrel membrane proteins. Current Opinion in Structural Biology **10**, 443–447 (2000)
65. Shorter, J., Lindquist, S.: Prions as adaptive conduits of memory and inheritance. Nat Rev Genet **6**(6), 435–450 (2005 Jun)
66. Simons, K.T., Kooperberg, C., Huang, E., Baker, D.: Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. J Mol Biol **268**(1), 209–225 (1997 Apr 25)

67. Singh, R., Berger, B.: ChainTweak: Sampling from the neighbourhood of a protein conformation. Proceedings of the 10th Pacific Symposium on Biocomputation pp. 52–63 (2005)
68. Sippl, M.J.: Calculation of conformational ensembles from potentials of mean force. Journal of Molecular Biology **213**, 859–883 (1990)
69. Thomas, S., Song, G., Amato, N.M.: Protein folding by motion planning. Phys Biol **2**(4), S148–55 (2005 Nov)
70. Ulmschneider, J.P., Jorgensen, W.L.: Polypeptide folding using monte carlo sampling, concerted rotation, and continuum solvation. J Am Chem Soc **126**(6), 1849–1857 (2004 Feb 18)
71. Vandeputte-Rutten, L., Bos, M.P., Tommassen, J., Gros, P.: Crystal structure of neisserial surface protein A (NspA), a conserved outer membrane protein with vaccine potential. J Biol Chem **278**(27), 24,825–24,830 (2003 Jul 4)
72. Voelz, V., Dill, K.: Exploring zipping and assembly as a protein folding principle. Proteins: Structure Function and Bioinformatics **66**, 877–888 (2007)
73. Vogt, J., Schulz, G.E.: The structure of the outer membrane protein OmpX from *escherichia coli* reveals possible mechanisms of virulence. Structure **7**(10), 1301–1309 (1999 Oct 15)
74. Wagner, G.P., Otto, W., Lynch, V., Stadler, P.F.: A stochastic model for the evolution of transcription factor binding site abundance. J Theor Biol **247**(3), 544–553 (2007 Aug 7)
75. Waldispühl, J., Berger, B., Clote, P., Steyaert, J.M.: Predicting transmembrane β-barrels and inter-strand residue interactions from sequence. Proteins: Structure, Function and Bioinformatics **65**, 61–74 (2006). Doi:10.1002/prot.2146
76. Waldispühl, J., Berger, B., Clote, P., Steyaert, J.M.: transfold: A web server for perdicting the structure of transmembrane proteins. Nucleic Acids Research (Web Server Issue) **34**, W189–W193 (2006). Doi:10.1093/nar/glk205
77. Waldispühl, J., O'Donnell, C.W., Devadas, S., Clote, P., Berger, B.: Modeling ensembles of transmembrane beta-barrel proteins. Proteins **71**(3), 1097–1112 (2008 May 15)
78. Waldispühl, J., O'Donnell, C.W., Will, S., Devadas, S., Backofen, R., Berger, B.: Simultaneous alignment and folding of protein sequences. In: S. Batzoglou (ed.) Research in Computational Molecular Biology, *Lecture Notes in Computer Science*, vol. Volume 5541/2009, pp. 339–355. Springer Berlin / Heidelberg (2009)
79. Waldispühl, J., Steyaert, J.M.: Modeling and predicting all-alpha transmembrane proteins including helix-helix pairing. Theor. Comput. Sci. **335**(1), 67–92 (2005)
80. William J. Wedemeyer, H.A.S.: Exact analytical loop closure in proteins using polynomial equations. J Comput Chem **20**(8), 819–844 (1999)
81. Wimley, W.C., White, S.H.: Reversible unfolding of β-sheets in membranes: A calorimetric study. Journal of Molecular Biology **342**, 703–711 (2004)
82. Xia, Y., Huang, E.S., Levitt, M., Samudrala, R.: Ab initio construction of protein tertiary structures using a hierarchical approach. J Mol Biol **300**(1), 171–185 (2000 Jun 30)
83. Y., Z., J., S.: SPICKER: A clustering approach to identify near-native protein folds. Journal of Computational Chemistry **25**, 865–871 (2004)
84. Zhao, F., Li, S., Sterner, B.W., Xu, J.: Discriminative learning for protein conformation sampling. Proteins **73**(1), 228–240 (2008 Oct)
85. Zhao, F., Peng, J., DeBartolo, J., Freed, K.F., Sosnick, T.R., Xu, J.: A probabilistic graphical model for ab initio folding. In: S. Batzoglou (ed.) Research in Computational Molecular Biology, *Lecture Notes in Computer Science*, vol. Volume 5541/2009, pp. 59–73. Springer Berlin / Heidelberg (2009)