PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Modeling ensembles of transmembrane β-barrel proteins

Jérôme Waldispühl,[1,2]* Charles W. O'Donnell,[2]* Srinivas Devadas,[2] Peter Clote,[3] and Bonnie Berger[1,2]†

[1] Department of Mathematics, MIT, Cambridge, Massachusetts
[2] Computer Science and Artificial Intelligence Lab, MIT, Cambridge, Massachusetts
[3] Department of Biology, Boston College, Chestnut Hill, Massachusetts

ABSTRACT

Transmembrane β-barrel (TMB) proteins are embedded in the outer membrane of Gram-negative bacteria, mitochondria, and chloroplasts. Despite their importance, very few nonhomologous TMB structures have been determined by X-ray diffraction because of the experimental difficulty encountered in crystallizing transmembrane proteins. We introduce the program partiFold to investigate the folding landscape of TMBs. By computing the Boltzmann partition function, partiFold estimates inter-β-strand residue interaction probabilities, predicts contacts and per-residue X-ray crystal structure B-values, and samples conformations from the Boltzmann low energy ensemble. This broad range of predictive capabilities is achieved using a single, parameterizable grammatical model to describe potential β-barrel supersecondary structures, combined with a novel energy function of stacked amino acid pair statistical potentials. PartiFold outperforms existing programs for inter-β-strand residue contact prediction on TMB proteins, offering both higher average predictive accuracy as well as more consistent results. Moreover, the integration of these contact probabilities inside a stochastic contact map can be used to infer a more meaningful picture of the TMB folding landscape, which cannot be achieved with other methods. Partifold's predictions of B-values are competitive with recent methods specifically designed for this problem. Finally, we show that sampling TMBs from the Boltzmann ensemble matches the X-ray crystal structure better than single structure prediction methods. A webserver running partiFold is available at http://partiFold.csail.mit.edu/.

Proteins 2008; 71:1097–1112.

Key words: outer membrane proteins; transmembrane β-barrels; residue contact; structure modeling; structure prediction; Boltzmann partition function; stochastic contact map; ensembles; sampling.

INTRODUCTION

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 19 after the removal of homologous sequences. Some *in vitro* and *in vivo* mutation studies of omps[1,2] 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.

Since omps play a major role in cellular biology, computational efforts have in recent years focused on secondary structure prediction for these proteins, despite the paucity of data.[3–7] Almost all current methods use traditional machine learning approaches such as hidden Markov models (HMMs) and neural networks (NNs). However, these algorithms fail to incorporate the long-range interactions that are thought to be important in the folding of TMBs, as described in the 4-stage model.[8] Hence, new methods that evaluate the effect of residue interactions over long distances during the omp folding process are essential for studying this class of proteins. To address this issue, we recently introduced a novel model for TMBs[9] that uses a simplified representation of these structures and incorporates long-range pairwise residue contacts in multitape S-attribute grammars. This allowed us to compute the TMB supersecondary structure with the global minimum folding energy (m.f.e.). Our program, *transFold*, accurately predicts TMB

*Jérôme Waldispühl and Charles W. O'Donnell contributed equally to this work.
Grant sponsor: NSF; Grant number: DBI-0543506; Grant sponsor: NSP; Grant number: ITR(ASE+NIH)-(dms)-0428715.
†Correspondence to: Bonnie Berger, Department of Mathematics, MIT, Cambridge, MA.
E-mail: bab@mit.edu
Received 11 May 2007; Revised 26 July 2007; Accepted 3 August 2007
Published online 14 November 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21788

supersecondary structure and can be used to reproduce *in silico* the observations of *in vivo* mutation studies.[1,2]

One reason investigating the conformational landscape for a protein is important is because the native state may be quite different from the predicted minimum energy conformation; indeed, Zhang and Skolnick[10] 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. There has been substantial work on characterizing globular protein folding landscapes for lattice and nonlattice models—see Levitt and coworkers,[11–13] Mirny and Shakhnovich,[14] and Dill and coworkers.[15–19] Chiang et al.[20] have recently computed the partition function for simple two-helix bundles to study conformational changes in a 2D lattice model. Morozov has also investigated ways of computing a partition function to characterize DNA occupancy of nucleosomes and DNA-binding proteins.[21]

In this article, we characterize the folding landscape of TMBs by drawing parallels with RNA secondary structure prediction. The first RNA secondary structure prediction algorithm computing the minimum folding energy structure in the nearest neighbor energy model was introduced by Zuker and Stiegler.[22] Following this, McCaskill greatly advanced RNA structure prediction by developing algorithms capable of computing the Boltzmann partition function and the base pair probability of an RNA molecule.[23] Recently, a significant extension has been proposed by Ding and Lawrence, who described an algorithm for sampling RNA secondary structures according to their weight in the Boltzmann ensemble.[24] The transFold algorithm that we recently introduced to compute the minimum folding energy structure of TMBs is similar in spirit to the original Zuker/Stiegler algorithm, though it uses a more powerful grammatical model, a multitape S-attribute grammar. We show now that the powerful techniques established by McCaskill, Ding, and Lawrence can be adapted and applied to TMB structure prediction, thus establishing a bridge between RNA and TMB structure exploration. To do this, we take 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 Istrail who proved that the partition function of an Ising model can be computed in polynomial time given a 2D lattice.[25]

Here, we describe the first polynomial-time, recursive algorithm to compute the Boltzmann partition function of all β-barrel structures for a given outer membrane protein. From the partition function, 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, from the partition function we estimate statistical mechanical parameters such as ensemble free energy, average internal energy, heat capacity, etc. Rigorously defined stochastic contact maps, sampling, and thermodynamic parameters give us insight into the folding landscape of outer membrane proteins—an insight which cannot be gained by methods solely dedicated to the prediction of the native state conformation. This approach also provides a unified framework that allows us to tackle a wide variety of structural prediction problems which were previously addressed by independent algorithms. This unified approach achieves a clear gain in accuracy, precluding the problem of contradictory predictions encountered when interpreting the results of multiple, independent algorithms.

In addition, we extend the state-of-the-art BETAWRAP energy model[26,27] already used with success for TMBs,[9] by introducing the notion of interstrand residue stacking pairs (stacking of two pairs of adjacent residues). This results in a certain improvement in predictive results, allowing our algorithms to benefit from a significant statistical signal, to our knowledge never before used for protein structure prediction.

In our results, we illustrate these advances by demonstrating (i) how stochastic contact maps can be used to perform accurate, state-of-the-art β-strand contact prediction, (ii) how X-ray crystal per-residue B-values can be predicted with an accuracy rivaling that of leading specific B-value prediction algorithms, and (iii) 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.

MATERIALS AND METHODS

Two-tape representation of transmembrane β-barrels

We provide a simple and unambiguous representation of transmembrane protein structure by modeling them with multitape context-free grammars.[9,28] In the case of TMBs, this modeling explicitly separates each of the antiparallel β-strand pairs involved in the barrel. The complete structure can then be described as a sequence of individual antiparallel pairings, including the closing strand pair. While the algorithmic concepts and routines presented in this article can be equally described without multitape context-free grammars, this representation provides a more concise conceptual description that still lends itself toward an efficient computational solution.

Grammars provide a versatile framework that can be easily adapted to match the needs of experimentalists. Indeed, experimental observations of putative residue contacts, for instance, can be used to constrain the ensemble of folds to respect some specific structural

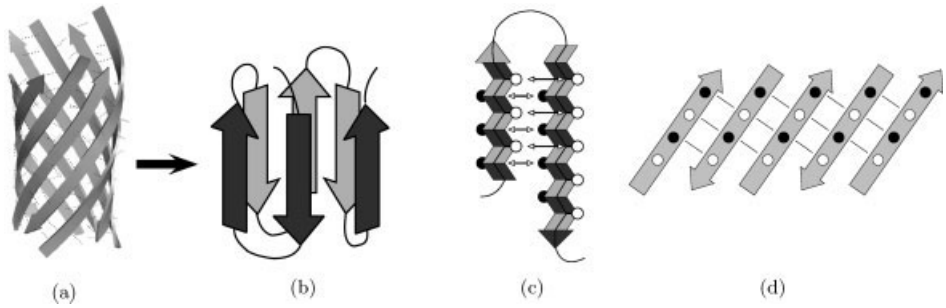


Figure 1

*Structure decomposition of transmembrane β-barrel. (**a**) The full structure of a transmembrane β-barrel, (**b**) overall shape of the channel, (**c**) antiparallel β-strands, and (**d**) inclination of TM β-strands across the membrane plane.*

features. Obviously, many others types of constraints can be designed, as was done by Waldispühl et al.[9] where some residues known to be present in extracellular loops were excluded from transmembrane strands.

To accurately represent TMBs using grammars (to agree with Schulz's summary[29]) we must describe three fundamental features of these structures: (i) the overall shape of the barrel (the number of TM β-strands and their relative arrangement), (ii) an exact description of the antiparallel 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. The modeling is based on an individual schematic representation of these features which will be merged hereafter. This decomposition of the structure into elementary units is illustrated in Figure 1.

A TMB must be decomposed into individual blocks of antiparallel β-strands, where each β-strand is involved in two distinct pairings—an exception being the "closing" strand pair involving the first and last β-strand. To handle this distinctly non-context-free feature, we employ a representation where the sequence is duplicated on a second tape, and pairings are made from one tape to the other. Figure 2 illustrates this representation, which is the foundation of the modeling introduced in Refs. 9 and 28 and motivates the designation of the ''2-tape representation.''

In this article, we introduce a new notation that allows us to generalize these models to compute critical features of the folding landscape. Each block can be represented as 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 strand on the first tape and i_2 and j_2 (s.t. $i_2 < j_2$) are those on the second tape.

We now consider an antiparallel pairing and the corresponding 4-tuple $\binom{i_1, j_1}{i_2, j_2}$. The left strand corresponds to the subsequence $[i_2 + 1, i_1]$, the right strand corresponds to $[j_2 + 1, j_1]$, and a loop to the subsequence $[i_1 + 1, j_2]$. Additionally, we assume that the rightmost amino acid at index i_1 of the left strand is paired with the leftmost residue at index $j_2 + 1$ of the right strand.

Although TM β-strands are not necessarily of the same length, the length of the residues in contact is

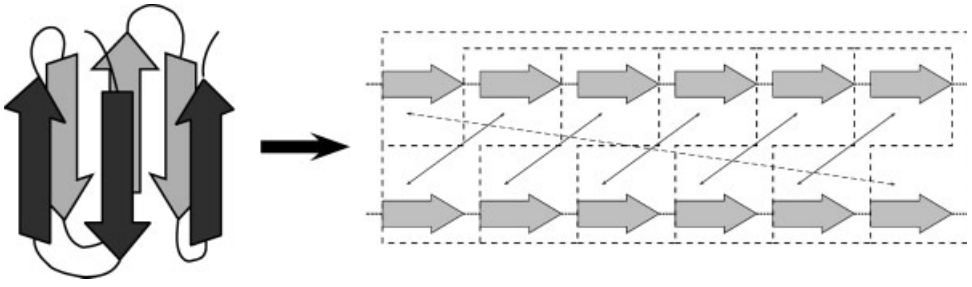


Figure 2

2-tape representation of a transmembrane β-barrel. The original input tape is duplicated and pairings are only allowed from one tape to the other. All pairings are antiparallel and indicated with arrows. The closing pair connects the first and last strands and is represented by the exterior block.

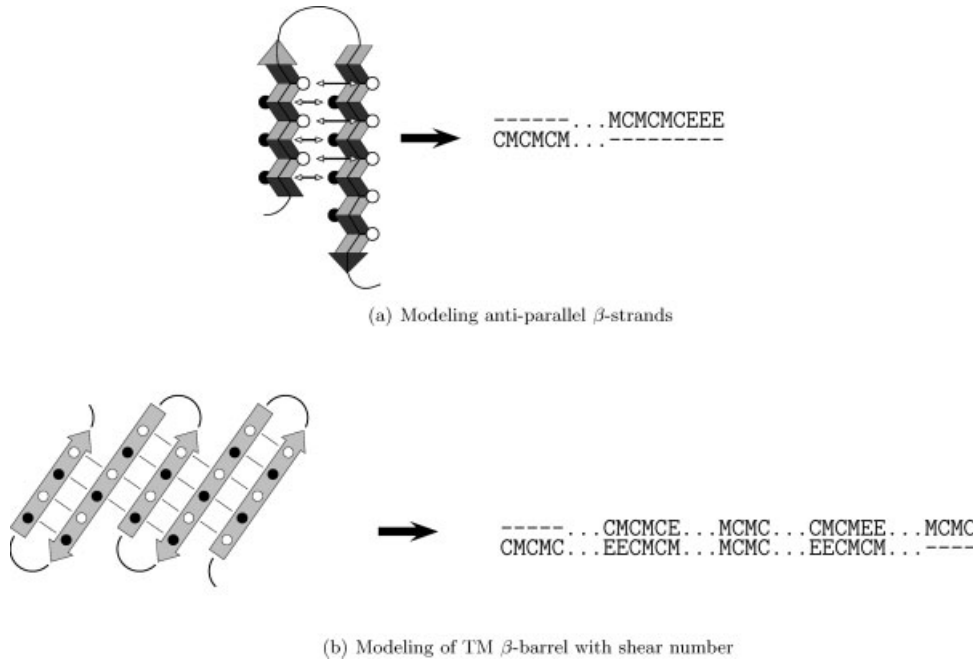


Figure 3

*(**a**) Representation of a TM β-strand pair with extension on left strand (i.e. extension). Residues annotated by M [resp. C] have side-chain facing the membrane [resp. channel], while those with E are unpaired β-strand residues which extend or reduce the strand. Dots "." represent the amino acids in the loop connecting the two strands, while dashes "-" are empty characters used to model the space available for the next pairing. (**b**) Representation of strand inclination using shear number. Reductions and extensions alternate around periplasmic loops (bottom) and extracellular loops in order to preserve the coherence of the orientation. The N-terminus of the protein sequence on the left diagram is at the right extremity.*

$L_c = \min(i_1 - i_2, j_1 - j_2)$ and the length of the strand extension is $L_e = |(i_1 - i_2) - (j_1 - j_2)|$. To avoid invalid configurations, only one strand from each pair can be extended. When an extension is done on the left strand, the right strand becomes shorter and the extension is then called a reduction; when an extension occurs on the right strand, the latter is longer and the operation is an extension.

The set $\mathcal{C}$ of residue–residue contacts involved in strand pairing can be defined as follows: $\mathcal{C} = \{(i_1 - k, j_2 + 1 + k) \mid 0 \leq k < L_c\}$. 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 $\mathcal{C}_0 = \left\{(i_1 - 2 \cdot k, j_2 + 1 + 2 \cdot k) \,\middle|\, 0 \leq k < \lfloor \frac{L_c}{2} \rfloor \right\}$ and $\mathcal{C}_1 = \left\{(i_1 + 1 - 2 \cdot k, j_2 + 2 \cdot k) \,\middle|\, 1 \leq k \leq \lfloor \frac{L_c}{2} \rfloor \right\}$ Assuming the location of the closest contact is known, we can also assign the nature of the milieu (i.e. membrane or channel).

For each block $\binom{i_1, j_1}{i_2, j_2}$ representing each distinct antiparallel pairing, we integrate these features by annotating each residue appropriately. β-strand residues with side-chains oriented toward the membrane are annotated with M, while those with side-chain oriented toward the channel are annotated with C. Unpaired β-strand residues are simply annotated E. An example of this modeling is given in Figure 3.

The inclination of strands through the membrane is modeled using a shear number. This feature is implemented with the help of strand extension. Indeed, strictly alternating reductions and extensions in consecutive strand pairs allows us to obtain the desired configuration. Without loss of generality, and in conjunction with experimental observations,[29] we assume that (i) the N-terminus is located on 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 extracellular side. Then, we restrict reductions to occur around periplasmic loops and extensions around extracellular loops. Figure 3 illustrates how to proceed.

It is worth note that, in principle, a similar 2-tape representation could be used to include other classes of β-barrel proteins domains as long as their structures followed similar topological rules. TMBs are well suited to the methodology given 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

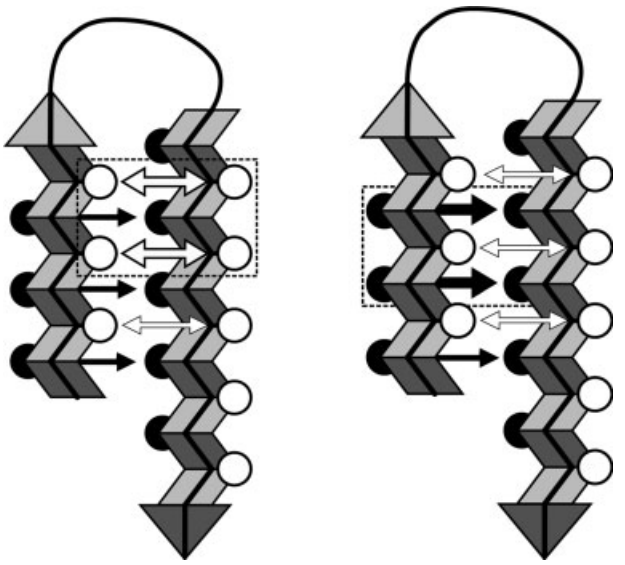


Figure 4
Stacking pairs of residues in antiparallel β-strands.

require more complicated rules (such as consecutive strands which are out of sequence order). These changes to the representation would affect the computational speed and tractability of our later techniques.

Energy model

Here we introduce a novel pseudo-energy model inspired by the classical RNA nearest neighbors model.[30] To describe this, we first introduce the notion of a stacking pair in a pair of β-strands. Intuitively, this consists of the stacking of two spatially adjacent pairs of H-bonding residues that have the same side-chain orientation. Figure 4 depicts such an arrangement. More specifically, consider an antiparallel β-strand pair and two residues, indexed i and j, such that i corresponds to an amino acid in the first strand and j to an amino acid in the second one. Then, assuming both pairs are H-bonded, the 2-tuple $((i,j), (i+2, j-2))$ is said to be a stacking pair of β-strand residues. The choice of the pair $(i+2, j-2)$ (as opposed to $(i+1, j-1)$) ensures that residues on the same face of the β-sheet are grouped since these are much closer in physical space and more likely to interact with one and another.

Using this idea, we say that the energy of a conformation is related to the sum of the knowledge-based statistical potentials associated with every unique stacking pair found within that structure. By this definition, stacking pairs improve upon the individual amino acid pairing potentials that have been used successfully in *transFold*[9] and BETAWRAP.[26,27]

Determining stacking potentials

To obtain statistical potentials for all possible amino acid stacking pairs, we compute the probability of observing adjacent pairs of stacked amino acids in solved β-sheet structures with characteristics closely matching those found in TMBs. Similar to prior methods,[9,26,27,31] we take the 50% nonredundant set of protein structures (PDB50) from the PDB[32] (taken April 4, 2007), and use STRIDE[33] to identify secondary structure features, solvent accessibility, and hydrogen bonds. Naturally, all solved structures of TMBs are removed as to not corrupt our testing (leaving 11,359 proteins in our set).

However, instead of simply counting the occurrences of β-sheet amino acid pairings in all known proteins, we restrict our search to better match the environment normally associated with TMB proteins. Namely, the barrel fold of TMBs are thought to consist of only antiparallel, amphipathic β-sheets, with a hydrophobic environment on the outer membrane side of the barrel and a hydrophilic environment commonly existing within. Therefore, we count the frequency of stacking pairs after extracting quadripeptide regions of antiparallel bonded β-strands that exhibit an amphipathic pattern. Alternating buried/exposed residues define amphipathicity, where a buried residue is required to have less than 4% the solvent accessible area as when that residue is in an extended G-X-G tripeptide,[34] and an exposed residue is required to have an area greater than 15%. For our data set, this results in approximately 143,000 unique experimentally observed stacking pairs that match a TMB environment. Using the stacking pairs frequency count, we estimate the conditional probability $P((X,Y)|(U,V))$ of observing the amino acid pair (X,Y) given an adjacent stacked pair (U,V). For consistency and to avoid parameter fitting, specific statistical bonuses have not been included in these potentials (e.g. special treatment of proline residues). Finer granularity information such as side-chain rotomers or atomic coordinates were also not included in this model, but may be integrated into a more sophisticated stacking potential model in the future.

Since a table of amino acid specific stacking pair potentials would require 20^4 entries, the only way to extract meaningful information from the PDB50 is to determine potentials based on a reduced residue alphabet. We investigated a number of reduced alphabet sets and decided upon the Wang and Wang 5-letter reduced alphabet[35] (cf. Section "Results and Discussion" for a description of the rationale behind this choice).

Energy calculation

Let i, j, $(i+2)$, and $(j-2)$ be the indices of two stacked amino acid pairs that are in contact, and let $x \in \{0,1\}$ be a variable which represents the type of environment in which such a contact occurs (which side of the amphipathic sheet). Specifically, $x = 0$ (resp. $x = 1$) when side-chain orientation is toward the channel interior (resp.

membrane). Let $E(i, j, x \mid i+2, j-2)$ denote the energy of the contact between residues ω_i and ω_j, with the environment x, given the adjacent stacked pair ω_{i+2} and ω_{j-2}.

Pairwise frequencies are transformed into an energy potential using the standard procedure (taking the negative logarithm—see Refs. 36 (pp 223–228) and 37 for details). Specifically, if $p_{i,j,x \mid i+2,j-2}$ is Boltzmann distributed, then $E(i,j,x \mid i+2, j-2) = -RT \log(p_{i,j,x \mid i+2,j-2}) - RT \log(Z_c)$. Here $\log(Z_c)$ is a statistical recentering constant that is chosen as a parameter. Further, although RT has no effect when computing the minimum folding energy structure,[9] this is not the case when computing the partition function for β-barrel structures. For this reason, our current software 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:

$$E = \sum_{(i,j)\in\mathcal{C}_0} E(i,j,0 \mid i+2, j-2) + \sum_{(i,j)\in\mathcal{C}_1} E(i,j,1 \mid i+2, j-2) \quad (1)$$

At the present time, our model does not contain any energy contribution for periplasmic or extracellular loops, although future work will indeed consider such loop energies. The method presented here to compute the β-barrel partition function can easily be adapted to handle loop energies in an extension of our pseudo-energy model, hence our approach has no intrinsic limitation.

Computing the partition function

Since a TMB structure can be represented as a sequence of antiparallel TM β-strand pairs, given any four indices i_1, i_2, j_1, j_2 and the environment x of the closing TM β-strand pair contact (i.e. "membrane" or "channel"), we can compute the energy $E(i_1, i_2, j_1, j_2, x)$ for the antiparallel β-strand pairing of $\omega_{i_1}, \ldots, \omega_{i_2}$ with $\omega_{j_1}, \ldots, \omega_{j_2}$. For all possible values of i_1, i_2, j_1, j_2, and x, we store the Boltzmann value $\exp(-E(i_1, i_2, j_1, j_2, x)/RT)$ 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 $\mathcal{O}(n^2)$,* where n represents sequence length.

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

Since the energy function is additive, we can decompose the energy of a TMB as the sum of the energy associated with each distinct antiparallel TM β-strand pair. Let n_s be the number of TM β-strands of the TMBs and let i_2^k [resp. $i_1^k - 1$] denote the index of the leftmost [resp. rightmost] residue of the k-th strand.† To simplify the algorithm description, in the following we will omit the parameter x used to indicate the environment of the first contact of an antiparallel TM β-strand pair. Therefore, the energy $E(s)$ of a given TMB structure s can be written as:

$$E(s) = E(i_1^{n_s}, i_2^{n_s}, i_1^1, i_2^1) + \sum_{k=1}^{n_s-1} E(i_1^k, i_2^k, i_1^{k+1}, i_2^{k+1}) \quad (3)$$

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 antiparallel TM β-strand pairs, i.e. a TMB without closure. This table can be dynamically filled using the following recursion:

$$Q_{sheet}\begin{pmatrix} i_1, j_1 \\ i_2, j_2 \end{pmatrix} = \sum_{(k_1,k_2)} Q_{sheet}(i_1, i_2, k_1, k_2) \cdot Q_{ap}(k_1, k_2, j_1, j_2) \quad (4)$$

Once filled, we use this array to compute the partition function Q_{tmb} over all TMBs. This operation consists of adding the contributions of the antiparallel β-strand pairs which close the extremities of the β-sheet. For this, we could use the values stored in Q_{ap}; however, in practice, we use a special array which is better suited to the special rules for this last β-strand pair.‡

$$Q_{tmb} = \sum_{(i_1,i_2)} \sum_{(j_1,j_2)} Q_{sheet}(i_1, i_2, j_1, j_2) \cdot Q_{ap}(j_1, j_2, i_1, i_2) \quad (5)$$

Note that in order to respect the pairwise orientation as well as strand inclination, the indices i_1, i_2 and j_1, j_2 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.

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 Ref. 38),

* Note that this bound can be decreased to $\mathcal{O}(n)$ if we bound the length of loops. However, since we use this table to compute the contribution of the closing strand pair, this feature is not considered.

† This notation is designed to respect the notation used for the strand pair block $\left(\begin{smallmatrix} i_1, j_1 \\ i_2, j_2 \end{smallmatrix}\right)$.

‡ The rules for the closing pair, explicitly described in Ref. 9, mainly consist of relaxing some constraints, and allowing extensions on both sides of the strand.

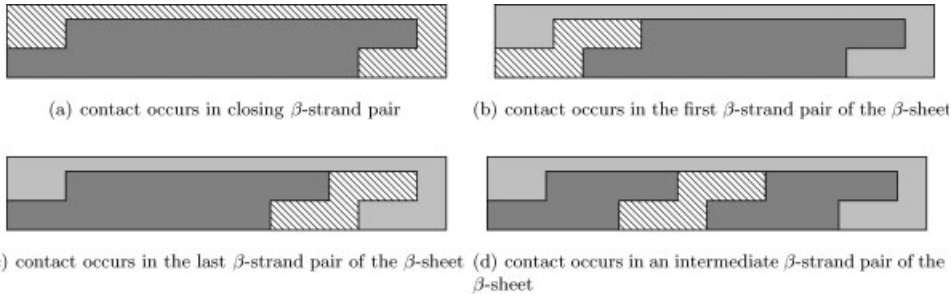


Figure 5

Decompositions of the transmembrane β-barrel, which allow us to isolate the antiparallel TM β-strand pair which contains the residue contact. The 2-tape block which corresponds to this strand pair is crosshatched. The blocks in dark and light gray respectively represent a TM β-sheet (i.e. a sequence of antiparallel TM β-strands) and the closing strand pair (i.e. the antiparallel β-strand pairs which close the sheet an form the barrel).

permit a better understanding of the folding landscape. For example, as shown in Ref. 39, 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.

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 of all, we need to characterize the antiparallel β-strand pairs which contain a given contact.

Property 1. *Let i and j ($i < j$) be two residues of two distinct consecutive antiparallel β-strands, and m and n (s.t. $i \leq m < n \leq j$) the two residues at the extremities of the connecting loop. Then, (i,j) are brought into contact if and only if $m + n = i + j$.*

It follows from this proposition that (i,j) is a valid contact if and only if the antiparallel β-strands $\binom{i_1,j_1}{i_2,j_2}$ verify $i + j = i_1 + j_2 + 1$ and $i_2 + 1 \leq i \leq i_1 < j_2 + 1 \leq j \leq j_1$.

To evaluate the residue pair probability $p(i,j)$, we must compute the partition function value over all TMB $Q(i,j)$ which contain this contact. Such TMB 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 5.

Let $\binom{i_1,j_1}{i_2,j_2}$ be an index of a block modeling an antiparallel TM β-strand pair. Then, we define $Q^{\text{close}}\binom{i_1,j_1}{i_2,j_2}$, $Q^{\text{first}}\binom{i_1,j_1}{i_2,j_2}$, $Q^{\text{last}}\binom{i_1,j_1}{i_2,j_2}$, and $Q^{\text{inter}}\binom{i_1,j_1}{i_2,j_2}$ to be the partition functions over all TMB structures which contain this antiparallel TM β-strand pair as, respectively, the pair closing the barrel (Fig. 5(a)), the first pair of the TM β-sheet (Fig. 5(b)), the last pair of the TM β-sheet (Fig. 5(c)), or any other intermediate pair (Fig. 5(d)). Formally:

$$Q^{\text{close}}\begin{pmatrix} i_1,j_1 \\ i_2,j_2 \end{pmatrix} = Q_{\text{sheet}}\begin{pmatrix} i_1,j_1 \\ i_2,j_2 \end{pmatrix} \cdot Q_{\text{ap}}\begin{pmatrix} j_1,i_1 \\ j_2,i_2 \end{pmatrix} \tag{7}$$

$$Q^{\text{first}}\begin{pmatrix} i_1,j_1 \\ i_2,j_2 \end{pmatrix} = \sum_{(y_1,y_2)} Q_{\text{ap}}\begin{pmatrix} i_1,j_1 \\ i_2,j_2 \end{pmatrix} \cdot Q_{\text{sheet}}\begin{pmatrix} j_1,y_1 \\ j_2,y_2 \end{pmatrix} \cdot Q_{\text{ap}}\begin{pmatrix} y_1,i_1 \\ y_2,i_2 \end{pmatrix} \tag{8}$$

$$Q^{\text{last}}\begin{pmatrix} i_1,j_1 \\ i_2,j_2 \end{pmatrix} = \sum_{(x_1,x_2)} Q_{\text{sheet}}\begin{pmatrix} x_1,i_1 \\ x_2,i_2 \end{pmatrix} \cdot Q_{\text{ap}}\begin{pmatrix} i_1,j_1 \\ i_2,j_2 \end{pmatrix} \cdot Q_{\text{ap}}\begin{pmatrix} j_1,x_1 \\ j_2,x_2 \end{pmatrix} \tag{9}$$

$$Q^{\text{inter}}\begin{pmatrix} i_1,j_1 \\ i_2,j_2 \end{pmatrix} = \sum_{\substack{(x_1,x_2) \\ (y_1,y_2)}} \left(Q_{\text{sheet}}\begin{pmatrix} x_1,i_1 \\ x_2,i_2 \end{pmatrix} \cdot Q_{\text{ap}}\begin{pmatrix} i_1,j_1 \\ i_2,j_2 \end{pmatrix} \cdot Q_{\text{sheet}}\begin{pmatrix} j_1,y_1 \\ j_2,y_2 \end{pmatrix} \cdot Q_{\text{ap}}\begin{pmatrix} y_1,x_1 \\ y_2,x_2 \end{pmatrix} \right) \tag{10}$$

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

$$Q(i,j) = \sum_{\substack{(i_1,i_2) \\ (j_1,j_2)}}^{i+j=i_1+j_2+1} \left(Q^{\text{close}}\begin{pmatrix} j_1,i_1 \\ j_2,i_2 \end{pmatrix} + Q^{\text{first}}\begin{pmatrix} i_1,j_1 \\ i_2,j_2 \end{pmatrix} + Q^{\text{last}}\begin{pmatrix} i_1,j_1 \\ i_2,j_2 \end{pmatrix} + Q^{\text{inter}}\begin{pmatrix} i_1,j_1 \\ i_2,j_2 \end{pmatrix} \right) \tag{11}$$

Finally, the Boltzmann probability $p(i,j)$ of a contact between the residues at indices i and j can be obtained by computing the value $p(i,j) = \frac{Q(i,j)}{Q_{\text{tmb}}}$. The contact map of a TMB can be immediately derived from this

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

Improved computing of the stochastic contact map

The formidable time requirement for a brute force algorithm to compute Eq. (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 present a simple strategy using additional dynamic tables, which allows us 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 which contains the residue contact is not involved, the product of the partition function of two substructures is realized over all possible configurations (i.e. $Q_x\left(\begin{smallmatrix} i_1, k_1 \\ i_2, k_2 \end{smallmatrix}\right) \cdot Q_y\left(\begin{smallmatrix} k_1, j_1 \\ k_2, j_2 \end{smallmatrix}\right)$ is computed over all possible pairs of indices (k_1,k_2)). In Eq. (10), the pairs of indices (x_1, x_2) and (j_1, j_2) are used for different residue contacts since the pair (i_1, i_2) varies. Thus, we can precompute the values of $Q_{\text{sheet}}\left(\begin{smallmatrix} y_1, j_1 \\ y_2, j_2 \end{smallmatrix}\right) \cdot Q_{\text{ap}}\left(\begin{smallmatrix} j_1, i_1 \\ j_2, i_2 \end{smallmatrix}\right)$ over all possible (y_1, y_2) and store them in a dynamic table for later retrieval. Given (i_1,i_2) and (j_1,j_2), let Q_{tail} be the array storing the values $\sum_{(k_1,k_2)} Q_x\left(\begin{smallmatrix} i_1, k_1 \\ i_2, k_2 \end{smallmatrix}\right) \cdot Q_y\left(\begin{smallmatrix} k_1, j_1 \\ k_2, j_2 \end{smallmatrix}\right)$. This table can be filled in time $\mathcal{O}(n^3)$. Then, in place of Eq. (10), we now have Eq. (12).

$$Q^{\text{inter}}\begin{pmatrix} i_1, j_1 \\ i_2, j_2 \end{pmatrix} = \sum_{(i_1,i_2)} Q_{\text{sheet}}\begin{pmatrix} x_1, i_1 \\ x_2, i_2 \end{pmatrix} \cdot Q_{\text{ap}}\begin{pmatrix} i_1, j_1 \\ i_2, j_2 \end{pmatrix} \cdot Q_{\text{tail}}\begin{pmatrix} j_1, x_1 \\ j_2, x_2 \end{pmatrix}. \tag{12}$$

Equations (8) and (9) need not be improved, since there is no redundancy in those cases. The time complexity for computing the entire contact map $p(i,j)$ is now $\mathcal{O}(n^4)$. However, an additional observation allows us save an additional factor n in the time complexity: when a TMB structure is considered in one of the Eqs. 7–10, the TM β-strand pair which contains the contact (i,j) also involves many other contacts. Hence, instead of using these equations to compute the values $Q(i,j)$ (and $p(i,j)$) 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 the contact map of a TMB, which runs in time $\mathcal{O}(n^3)$.

Although not explicitly mentioned thus far, we should emphasize that we can also compute the contact probability $p_x(i,j)$ for a specific environment x—that is membrane or channel (cf. Section "Energy Model" for 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.

Rigorous sampling of transmembrane β-barrels

Ding and Lawrence[24] introduced a powerful technique to sample RNA secondary structures according to their weight in the Boltzmann ensemble. This method has been successfully applied to uncover critical features of the distribution of structures over the RNA folding landscape, as well as in biologically important applications such as gene knock-down experiments.[§]

Inspired by this work, we design a rigorous sampling algorithm for TMBs. Given an amino acid sequence s, we are able to randomly generate, according to the distribution of structures in the Boltzmann ensemble, low energy TMB structures for s. By sampling, we expect to be able to efficiently estimate nontrivial features concerning the ensemble of potential TMB folds, with the long-term goal of 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 6. First, The "closing" antiparallel strand pair is sampled according to the weight of all TMBs that contain it over all possible TMB. Then, we sample each antiparallel 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 correctness of the algorithm is ensured by construction of the dynamic table in Eqs. (4) and (5).

RESULTS AND DISCUSSION

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 demonstrate the flexibility of the approach and evaluate the reliability of these predictions by comparing individual contact predictions and B-value predictions against known X-ray crystal structures. We further perform whole structure sampling to show the benefits of ensemble modeling over single structure prediction, and the possibilities for structural exploration provided by these techniques. A fully-functional web server is available at

[§] By analyzing Boltzmann samples of messenger RNA (mRNA), likely single-stranded regions of mRNA can be determined. Such single-stranded regions are good targets for hybridization by small interfering RNAs (siRNA).

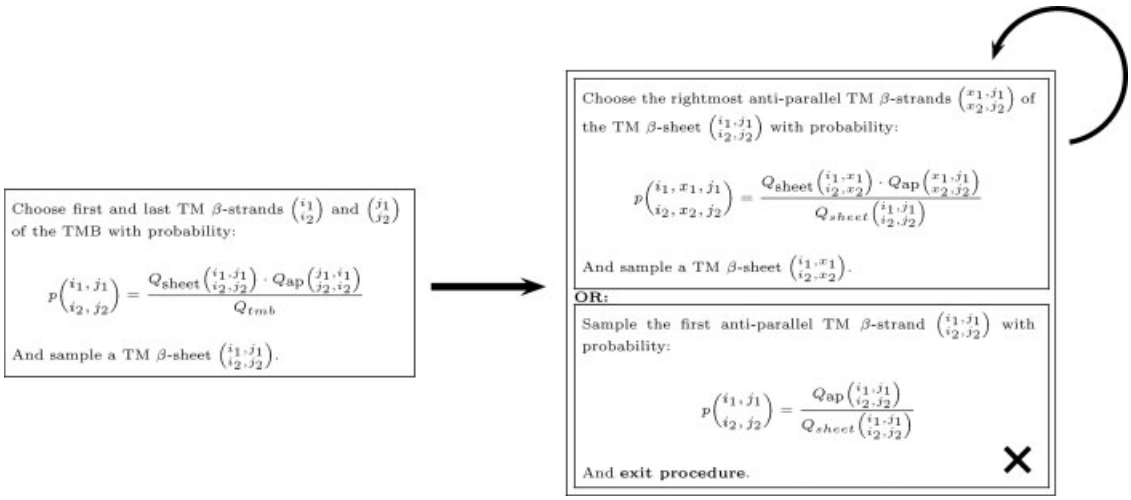


Figure 6

Sampling procedure: First, the first and last TM β-strands of the barrel are sampled (left box). Then, we sample the remaining TM β-sheet by iteratively sampling the rightmost antiparallel β strand of the remaining sequence, until we finally sample the first β-strand pair of the sheet.

http://partiFold.csail.mit.edu which displays the comprehensive set of our input data and results, as well as generates the same predictions and visuals for arbitrary TMBs.

Dataset

Very few TMBs have experimentally-derived structures deposited in the PDB. After removing homologous sequences, and focusing on monomeric, amphipathic TMBs without any plugging domains, we find in our test set 8 proteins with known X-ray crystal structures (PDB codes: 1QJ8, 1P4T, 1QJP, 1THQ, 1K24, 1QD6, 1TLY, and 1I78—Figure 7). Larger omps such as porins have been excluded since they typically exist in trimer, and can contain short α-helical loops which are critical for stabilization. Similarly, a number of TMBs are found to have large plug domains within the barrel itself, likely stabilizing the structure in an irregular, possibly dynamic

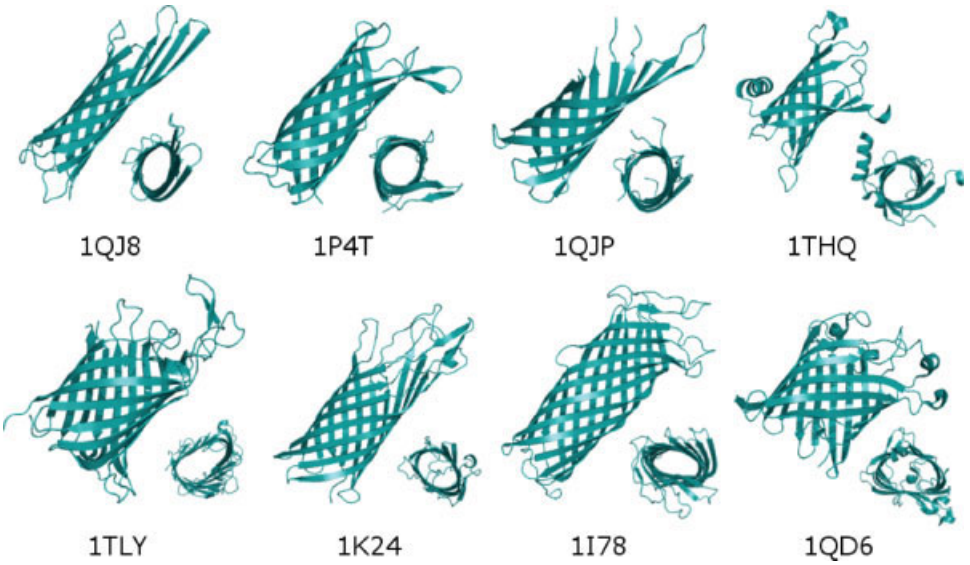


Figure 7

3D structures of the eight known monomeric, amphipathic, plug-domain-free TMBs as shown from the side and top. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Figure 8

Illustrative representations of stochastic contact predictions. Left: Stochastic contact map for 1P4T. 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) 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 are shown in red. Right: 2D representation (unrolled β-barrel) of 1QJ8 X-ray crystal structure 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 red, low probability in cyan). The leftmost β-strand is repeated on the right to allow the barrel to close, labeled dup.

fashion. Given *a priori* knowledge of such configurations, it may be possible to adjust our model to provide accurate predictions, however, the current energy function has not been formulated with this goal.

This paucity of experimental structural information is in fact recurrent in TMB structure prediction research. For instance, only eight structures were used to train and evaluate PROFtmb, a state-of-the-art genomic-level TMB existence predictor.[5]

We further divide our test set of eight to distinguish short (<200aa: 1QJ8, 1P4T, 1QJP, 1THQ) and long (>200aa: 1TLY, 1K24, 1I78, 1QD6) proteins, and apply two different choices of grammar constraints in a similar manner as was done by Waldispühl et al.[9] (A full listing is available on the webserver.) This matches an observed link between the length of the peptide sequence and the length, number, and sheer of the strands that make up the barrel.

How to evaluate predictions of ensembles

The class of predictions enabled by *partiFold* embody whole-ensemble properties of a protein. The stochastic contact map generated (cf. Section "Computing the Residue Contact Probability") reflects the likelihood of two β-strand amino acids pairing in the (estimated) Boltzmann distribution of conformations, and not one single minimum structure. Figure 8 depicts two ways to view information from a stochastic contact map. On the left, the full contact map of 1P4T is shown, identifying the probability of contact for all possible pairs of residues across all conformations in the Boltzmann distribution. On the right, a single structure is chosen (in this case the X-ray structure of 1QJ8), 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 (red) or a low (cyan) probability in the Boltzmann distributed ensemble. From this, substructures may be identified from their relative likelihood of pairing.

Unfortunately, validation of our results is limited by the availability of a *single* solved X-ray crystal structure for each test protein. Therefore, we focus validation on the task of single contact prediction of X-ray crystal structures even though much more information can be obtained from our results about the nature of the folding landscape, suggesting future experimental directions. Nonetheless, single contact prediction remains an important concern when reconstructing 3D models,[40–42] and *partiFold* still performs well in this capacity.

We define a set of single contact predictions 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

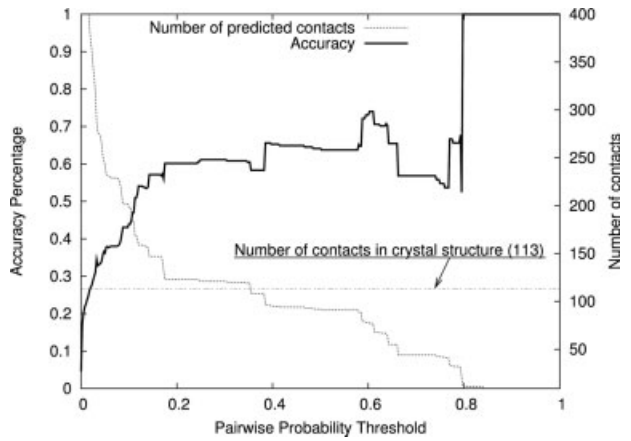


Figure 9

Predicting residue contact probabilities in 1QJ8. Plot of prediction accuracy as a function of size of predicted contact set.

STRIDE.[33] To evaluate our contact predictions we rely on three standard measures: the coverage (i.e., sensitivity), where coverage $= \frac{\text{number of correctly predicted contacts}}{\text{number of observed contacts}}$, the accuracy (i.e., positive predictive value), where accuracy $= \frac{\text{number of correctly predicted contacts}}{\text{number of predicted contacts}}$, and the F-measure, where F-measure $= \frac{2 \cdot \text{Coverage} \cdot \text{Accuracy}}{\text{Coverage} + \text{Accuracy}}$.

To demonstrate how these metrics would apply to this type of contact prediction, we refer to Figure 9 depicting the accuracy of contact prediction for 1QJ8 as a function of the size of the predicted set (e.g. p_t). Here one finds a high predictive accuracy ($\approx$ 60–70%) when the number of contact predictions made is roughly the number of contacts in the X-ray crystal structure ($\approx$ 100–120 pairs). 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.

Stacking pairs outperform single pair potentials

Here we justify our choice for using the Wang and Wang 5-letter reduced alphabet[35] for the amino acid stacking pair potentials described in Section "Determining Stacking Potentials". Following preliminary study, five alphabets were

Reduced alphabet	Group 1	Group 2	Group 3	Group 4	Group 5
Wang & Wang 5-letter (WW5)	CMFI LVWY	ATH	GP	DE	SNQ RK
Wang & Wang Variant 5-letter (WWV5)	CMFI	LVWY	ATGS	NQDE	HPRK
Chemical differentiation 4-letter (Chem4)	IVL	FYWH	KRDE	GACS	TMQNP
Li 4-letter (Li4)	CFYW	MLIV	GPA TS	NHQE DRK	-
C Murphy 4-letter (Mur4)	LVI MC	AGS TP	FYW	EDNQ KRH	-

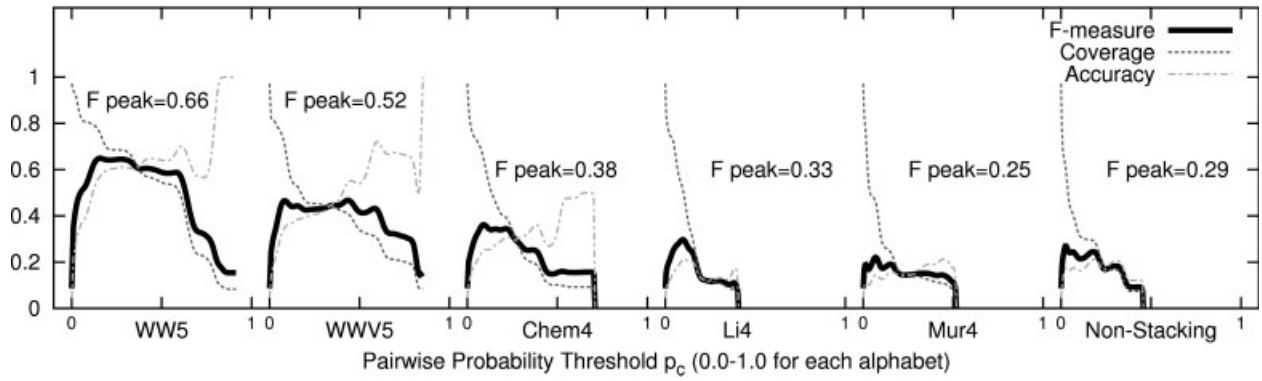
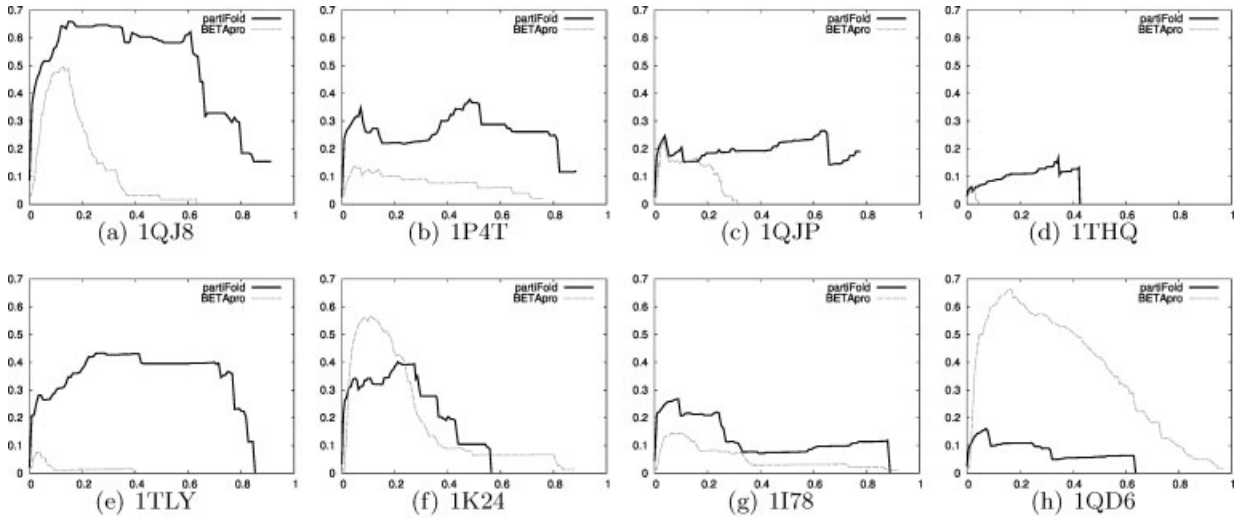


Figure 10

Above: Amino acid groupings for reduced alphabets selected and tested. Below: Smoothed F-measure/coverage/accuracy plots for the 1QJ8 protein across five reduced alphabets, plus the non-reduced, non-stacking energy potential previously used in Ref. 9 for comparison.



F-measure peak	1QJ8	1P4T	1QJP	1THQ	1TLY	1K24	1I78	1QD6
partiFold	**0.66**	**0.38**	**0.27**	**0.18**	**0.43**	0.40	**0.27**	0.16
BETApro	0.49	0.14	0.22	0.05	0.08	**0.56**	0.15	**0.66**

Figure 11

F-measure scores (y-axis) comparing partiFold (black) and BETApro (gray) as a function of number of contacts predicted (i.e. all contacts with contact probability greater than any threshold p_t along x-axis). Bold entries in table show higher performance.

selected to represent a broad range of residue classifications, and their predictive abilities were fully tested on our available protein structures. We present results for the protein 1QJ8 in Figure 10. The original energy parameters from Waldispühl et al.[9] are also included for comparison.

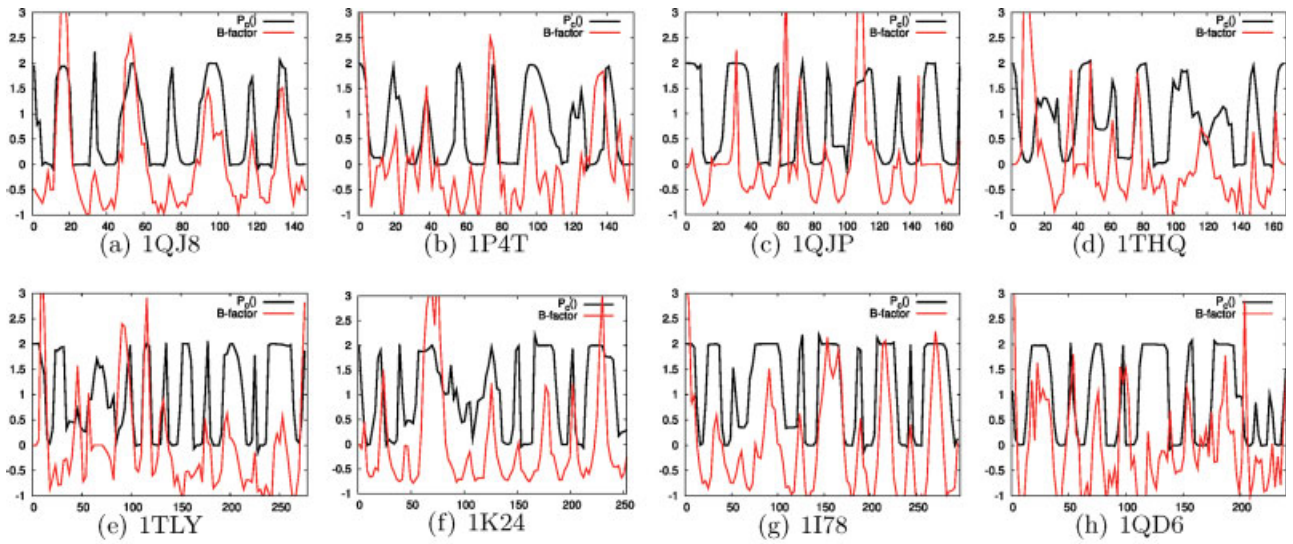
These plots show that the Wang and Wang alphabet offers the highest combination of coverage and accuracy for contact prediction, though a few other alphabets offer decent accuracy for a smaller coverage. The exact statistical potentials derived using the Wang and Wang alphabet can be found on the web server previously mentioned. The (i, j, k, l) tuples had a mean value of $\text{Avg}[\ln(p(i, j, k, l))] = -1.15$ for membrane facing residues and $\text{Avg}[\ln(p(i, j)] = -0.78$ for channel facing residues, while the standard deviations were 2.75 and 1.77, respectively.

Interestingly, the majority of stacking pair potential alphabets outperformed the nonstacking pair potentials used by Waldispühl et al.,[9] supporting the hypothesis that stacking pairs better describe the energy potential of TMBs. Experimentation on other proteins revealed varied results, though the Wang and Wang alphabet tends to remain the best candidate. One reason for this may be the biophysically important segregation of aspartic and glutamic acids into their own residue classes, reducing stacking charge clashes.

Ensemble approach improves individual contact prediction

To test the accuracy of our algorithm, in Figure 11 we compare its residue contact prediction abilities (cf. Section "Computing the Residue Contact Probability") in matching the contacts of solved X-ray crystal structures with the abilities of BETApro,[43] a recent stochastic secondary and super-secondary structure predictor made specifically for β-sheets (although not precisely for TMBs since the graph-based approach used does not support the barrel closure created by the pairing of the extremal β-strands of the β-sheet). This algorithm was arguably the top performer of the CASP7 inter-residue contact predictions competition,[41] and most closely resembles *partiFold*'s ability to stochastically predict β-strand contacts. Other general contact predictors, such as SVMcon[44] showed reduced average performance when compared with BETApro on our particular test set, and offered many more high probability contacts than can exist in real structures.

It should be noted that while BETApro does provide a stochastic contact map of β-strand interactions, its interaction probabilities are not related to a Boltzmann distribution of conformations, but rather based on a sophisticated neural network and graph algorithm that aims to predict a single structure. Its energy model also appears to not be common across all proteins, and, unlike partiFold, incorporates secondary structure and solvent accessibility profiles of the target amino acid sequence. Finally, BETApro was designed for, and trained on, globular proteins, and it does not support important aspects of β-barrel architectures such as circular β-sheets. Thus, during comparison, one must keep in mind that BETApro was not designed specifically for TMBs.



Correlation coefficient	1QJ8	1P4T	1QJP	1THQ	1TLY	1K24	1I78	1QD6
partiFold	**0.76**	0.26	0.34	0.05	**0.25**	**0.49**	**0.64**	0.27
PROFBval	0.62	**0.43**	**0.54**	**0.18**	**0.25**	0.40	0.56	**0.57**

Figure 12

Contact probability profile (black, y-axis) and normalized B-value curve (red, y-axis) for partiFold as a function of residue index from left to right (x-axis). Because of the simple shape of most TMBs, experimental B-values tend to oscillate from high to low. Regions of B-value curves which are flat at 0 represent residues missing from the X-ray crystal structure (e.g. 1QJP residues 146–159, 1THQ residues 38–47, etc.). Bold entries in table show higher performance.

A comparison of F-measure scores (cf. Section "How to Evaluate Predictions of Ensembles") is plotted in Figure 11. The range of peak scores shown varies from 0.16–0.66, which indicates good coverage and accuracy when considered against F-measure scores reported for CASP7 inter-residue contact predictions of 0.02–0.09.[41,44] For all but two proteins tested, our predictor strictly improves upon the results of BETApro, with a median peak score of 0.33 versus 0.19. More importantly, *partiFold* provides more consistent results across all proteins, and maintains flattened curves, indicative of good separation between high probability contacts and noise.

The performance of 1K24 and 1QD6 can be directly attributed to their inclusion of extracellular structural components outside of the barrel (see Figure 7). Since our current model focuses only on the barrel fold of a TMB, extra β-sheets and α-helices can be missed, as in 1K24 and 1THQ, degrading performance (the latter much more strikingly due to its already short sequence). In 1QD6, a large number of 3_{10} and α-helical structures cap the β-barrel and partially interact with the β-sheet walls, creating an small environment inconsistent with our energy model and interfering with *partiFold* predictions. Alteration of constraints as in Waldispühl et al.[9] results in an improved peak score (0.30–0.40), but would require *a priori* evidence of such a configuration. BETApro uses a more complicated and less transparent model that uses secondary structure annotation to identify such regions. Future versions of *partiFold* may include this as an option.

Consistent with prior predictors (including BETApro), our algorithm does not yet model bulges in β-sheets, and suffers slightly in performance where bulges exist. However, of the proteins tested, only 8 of 76 β-strand pairs contained bulges (type C or W[45]). Across β-sheets in general, only 14% of paired strands have bulges, and of those, 90% have only a single bulge.[43] Therefore, the impact of bulges on Figure 11 should be minimal. However, the possibility that our approach can aide in bulge discovery is a subject of ongoing research.

Residue flexibility can be predicted from contact probability profile

We show in Figure 12 how the stochastic contact map (cf. Section "Computing the Residue Contact Probability") generated by *partiFold* can be used to predict per-residue flexibility and entropy. To a first approximation, this flexibility can correlate with the Debye-Waller factor (a.k.a. the B-value) found in X-ray crystal structures.[46] This demonstrates an important purpose of computing the Boltzmann partition function: to provide a biologi-

cally-relevant grounding for the prediction of experimentally testable macroscopic and microscopic properties.

Predicting residue B-values 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.[47] In our context, flexible regions are strong candidates for loop regions connecting antiparallel TM β-strands that extend either into the extracellular or intracellular milieu.

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 in Figure 12 compare this against the normalized B-value, $B_{\text{norm}} = \frac{B-\langle B \rangle}{\sigma}$, a ratio commonly used for such a comparison.[47] 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 a greater chance for flexibility. Similarly, residues with a positive B-value are considered flexible or disordered while others are considered rigid.

Computing the cross-correlation coefficient between the P_c and B-value of our test proteins, we find that partiFold compares well against PROFBval,[47] a leading edge algorithm tuned specifically for B-value prediction. In fact, the more generally applicable *partiFold* method improves upon or matches 4 of the 8 TMBs. We have computed the per-residue contact entropy (defined as $S_i = \sum_{j=1}^{n} - p_{i,j} \log(p_{i,j})$) for the same test proteins and found similar results.

Protein	largest cluster: centroid cov.	largest cluster: centroid acc.	largest cluster: top sample cov.	largest cluster: top sample acc.	largest cluster: size	m.f.e. cov.	m.f.e. acc.
1QJ8	**0.65**	**0.67**	0.78	0.82	375	0.65	0.58
1QJP	**0.38**	**0.34**	0.33	0.33	422	0.19	0.21
1P4T	**0.20**	**0.18**	0.41	0.39	309	0.13	0.14
1THQ	**0.11**	0.09	0.11	0.09	358	0.08	0.11
1TLY	0.32	0.33	0.32	0.34	303	0.36	0.40
1K24	**0.15**	**0.17**	0.53	0.58	428	0.09	0.08
1I78	**0.17**	**0.24**	0.23	0.32	373	0.17	0.13
1QD6	**0.14**	**0.14**	0.22	0.22	568	0.05	0.06

Protein	"best" cluster: centroid cov.	"best" cluster: centroid acc.	"best" cluster: top sample cov.	"best" cluster: top sample acc.	"best" cluster: size	m.f.e. cov.	m.f.e. acc.
1QJ8	**0.65**	**0.67**	0.78	0.82	375	0.65	0.58
1QJP	**0.38**	**0.34**	0.33	0.33	422	0.19	0.21
1P4T	**0.43**	**0.38**	0.42	0.39	109	0.13	0.14
1THQ	**0.20**	**0.16**	0.20	0.16	40	0.08	0.11
1TLY	**0.37**	0.38	0.37	0.39	15	0.36	0.40
1K24	**0.31**	**0.34**	0.31	0.35	24	0.09	0.08
1I78	**0.22**	**0.31**	0.27	0.36	53	0.17	0.13
1QD6	**0.14**	**0.14**	0.22	0.22	568	0.05	0.06

Figure 13

Coverage and accuracy of contacts when compared against X-ray crystal structure. Centroid representative structure scores are given as well as the top performing sample in that given cluster. Bold numbers show the trend of improvement in the centroid structure's coverage and accuracy over that of the m.f.e. structure. Above: Largest cluster produced when sampling 1000 TMB structures. Below: "Best" cluster produced, as defined by the cluster containing the centroid conformation with the minimal $d_c()$, but no fewer than 15 samples.

Boltzmann sampling improves whole structure prediction

To demonstrate 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 "Rigorous Sampling of Transmembrane β-barrels") to map the landscape defined by the Boltzmann partition function. This also illustrates how *partiFold* can be used to 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. In this set of clusters, we show in Figure 13 that some individual clusters tend to match the X-ray crystal structure better than the single minimum folding energy (m.f.e.) structures.

In this examination we sample 1000 TMB structures and group them into 10 clusters according to hierarchical clustering. Similar to prior methods,[24] for each cluster we designate a centroid representative conformation that is chosen as the structure with the minimal total distance to all other structures in the set. To facilitate this clustering, we introduce a metric of contact distance: $d_c(S_1, S_2) = |\mathcal{C}_1| + |\mathcal{C}_2| - 2 \cdot |\{\mathcal{C}_1 \cap \mathcal{C}_2\}|$, where $\mathcal{C}_1$ and $\mathcal{C}_2$ are the sets of contact in S_1 and S_2 (which represents the minimal number of contacts to be removed and added to pass from S_1 to S_2 or vice versa).

Figure 13 reports the coverage and accuracy of contact predictions for the largest cluster produced and for the cluster who's centroid structure best matches the X-ray crystal structure contacts (minimizing $d_c()$, labeled "best"), ignoring clusters smaller than 15. Both centroid scores and scores for the highest coverage and accuracy sample ("top sample") within that cluster are listed. Comparing coverage and accuracy scores, surprisingly the centroid structures of *both* the largest and "best" cluster usually outperform the scores obtained by the minimum folding energy structure. This is despite the fact that in five of the cases the "best" cluster is not the largest cluster produced (e.g. 1THQ and 1I78). From this we see that the minimum folding energy structure does not always best describe the structure found by X-ray crystallography. This might even suggest that alternate conformations might be found in the Boltzmann distribution with high probability, although a more sophisticated energy model, including, for instance, an explicit term

for the entire connecting loops, would be required to understand this result. In future work, we intend to improve upon these simple clustering techniques and further explore these implications on the folding landscape.

CONCLUSIONS

In this article, we present the first set of algorithms for computing the Boltzmann partition function and stochastic contact maps of TMBs. From these calculations we establish techniques to perform individual contact prediction, B-value prediction, and whole structure sampling from the Boltzmann low energy ensemble. Unlike other approaches that aggregate numerous, complex techniques for a single predictive goal, the algorithms presented here use a simple, biophysically meaningful model that is capable of generating predictions for a broad range of molecular properties. This gives our method the important benefit of transparency when interpreting results. Accompanying these new algorithms, we introduce an energy model for TMBs that incorporates the information present in intrastrand stacked pairs of amino acids, resulting in higher predictive accuracy than prior nonstacked pair models.

The reliability and accuracy of our method is verified to be good by comparing its individual contact and B-value predictions against two of the forefront algorithms that are exclusively designed to handle these tasks. Specifically focusing on the study of omps, *partiFold* is able to offer significant improvement in accuracy over BETApro in our tests. The only cases where *partiFold* does not outperform BETApro can be attributed to extracellular non-β-barrel regions that BETApro distinguishes using additional information. This same method is also able to perform quite comparably to PROFBval for B-value prediction; a fact that speaks to the generality of the Boltzmann ensemble approach. Indeed, grounding these predictions on a unified model provides a framework for dependable results, whereas independent algorithms might present unresolvable contradictions.

Notably, the results presented also show that sampling from the Boltzmann distribution of conformations can lead to a much better characterization of the X-ray crystal structure than by computing the minimum folding energy structure alone. Sampling can also suggest alternate structures that might exist *in vivo*. This highlights our intention that *partiFold* also serve as a useful exploratory tool. To this end, the software has been designed to allow experimental observations to constrain the ensemble of folds that can be adopted, providing a natural way to combine experimental techniques with theoretical predictions when investigating omp structures or substructures.

The *partiFold* suite of tools is still in development and freely available online at the URL http://partiFold.csail.mit.edu. In future work, we expect to provide efficient methods for applying these algorithms to larger multimeric outer-membrane proteins, and to expand the range of environmental parameters that can be manipulated within the algorithm.

ACKNOWLEDGMENT

We graciously thank Nathan Palmer for insightful discussions and initial testing of the energy model.

REFERENCES

1. Wimley WC, White SH. Reversible unfolding of β-sheets in membranes: A calorimetric study. J Mol Biol 2004;342:703–711.
2. Koebnik R. Membrane assembly of the *Escherichia Coli* outer membrane protein ompa: exploring sequence constraints on transmembrane β-strands. J Mol Biol 1999;285:1801–1810.
3. Martelli PL, Fariselli P, Krogh A, Casadio R. A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. Bioinformatics 2002;S46–S53. Proceedings of ISMB'02.
4. Liu Q, Zhu Y-S, Wang B-H, Li Y-X. A HMM-based method to predict the transmembrane regions of β-barrel membrane proteins. Comput Biol Chem 2003;27:69–76.
5. Bigelow H, Petrey D, Liu J, Przybylski D, Rost B. Predicting transmembrane beta-barrels for entire proteomes. Nucleic Acids Res 2004;32:2566–2577.
6. Navjyot KN, Harpreet K, Raghava GPS. Prediction of transmembrane regions of β-barrel proteins using ANN- and SVM-based methods. Proteins: Struct Funct Bioinform 2004;56:11–18.
7. Gromiha M, Suwa M. A simple statistical method for discriminating outer membrane proteins with better accuracy. Bioinformatics 2005;21:961–968.
8. Tamm LK, Hong H, Liang B. Folding and assembly of β-barrel membrane proteins. Biochimica et Biophysica Acta–Biomemb 2004;1666:250–263.
9. Waldispühl J, Berger B, Clote P, Steyaert J-M. Predicting transmembrane β-barrels and inter-strand residue interactions from sequence. Proteins: Struct Func Bioinform 2006;65:61–74.
10. Zhang Y, Skolnick J. SPICKER: A clustering approach to identify near-native protein folds. J Comput Chem 2004;25:865–871.
11. Park B, Levitt M. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. J Mol Biol 1996;258:367–392.
12. Fain B, Levitt M. A novel method for sampling alpha-helical protein backbones. J Mol Biol 2001;305:191–201.
13. Fain B, Levitt M. Funnel sculpting for in silico assembly of secondary structure elements of proteins. Proc Natl Acad Sci USA 2003;100:10700–10705.
14. Mirny L, Shakhnovich E. Protein folding theory: from lattice to all-atom models. Annu Rev Biophys Biomol Struct 2001;30:361–396.
15. Dill KA, Phillips AT, Rosen JB. Protein structure and energy landscape dependence on sequence using a continuous energy function. J Comput Biol 1997;4:227–239.
16. Miller DW, Dill KA. Ligand binding to proteins: the binding landscape model. Protein Sci 1997;6:2166–2179.
17. Amato NM, Dill KA, Song G. Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures. J Comput Biol 2003;10:239–255.
18. Hockenmaier J, Joshi AK, Dill KA. Routes are trees: The parsing perspective on protein folding. Proteins: Struct Funct Bioinform 2007;66:1–15.

19. Voelz VA, Dill KA. Exploring zipping and assembly as a protein folding principle. Proteins: Struct Funct Bioinform 2007;66:877–888.
20. Chiang D, Joshi AK, Dill KA. A grammatical theory for the conformational changes of simple helix bundles. J Comput Biol 2006; 13:27–42.
21. Alexandre V. Morozov, The Rockefeller University, personal communication, April 25, 2007.
22. Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 1981;9:133–148.
23. McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 1990;29:1105–1119.
24. Ding Y, Lawrence CE. A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res 2003;31:7280–7301.
25. 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: ACM Press, editor. Proceedings of the 32nd ACM symposium on the theory of computing (STOC00), 2000. pp. 87–96.
26. Bradley P, Cowen L, Menke M, King J, Berger B. Betawrap: successful prediction of parallel β-helices from primary sequence reveals an association with many microbial pathogens. Proc Nat Acad Sci, USA 2001;98:14819–14824.
27. Cowen L, Bradley P, Menke M, King J, Berger B. Predicting the beta-helix fold from protein sequence data. J Comput Biol 2002; 9:261–276.
28. Waldispühl J, Steyaert J-M. Modeling and predicting all-α transmembranaire proteins including helix–helix pairing. Theor Comput Sci (special issue on Pattern Discovery in the Post Genome) 2005; 67–92.
29. Schulz G. β-barrel membrane proteins. Curr Opin Struct Biol 2000; 10:443–447.
30. Xia T, SantaLucia J Jr, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry 1999;37:14719–14735.
31. Kernytsky A, Rost B. Static benchmarking of membrane helix prediction. Nucleic Acids Res 2003;31:3642–3644.
32. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I, Bourne P. The protein data bank. Nucleic Acids Res 2000;28:235–242.
33. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins 1995;23:566–579.
34. Chotia C. The Nature of the Accessible and Buried Surfaces in Proteins. J Mol Biol 1975;105:1–14.
35. Jun Wang, Wei Wang. A computational approach to simplifying the protein folding alphabet. Nat Struct Biol 1999;6:1033–1038.
36. Clote P, Backofen R. Computational molecular biology: an introduction. Wiley; 2000. 279p.
37. Manfred J. Sippl. Calculation of conformational ensembles from potentials of mean force. J Mol Biol 1990;213:859–883.
38. Dill KA, Bromberg S. Molecular driving forces. New York: Garland Science, Taylor & Francis; 2003.
39. Clote P, Waldispühl J, Behzadi B, Steyaert J-M. Energy landscape of k-point mutants of an RNA molecule. Bioinformatics 2005;21: 4140–4147.
40. Grana O, Baker D, MacCallum RM, Meiler J, Punta M, Rost ML, Tress B, Valencia A. Casp6 assessment of contact prediction. Proteins 2005;61:214–224.
41. Critical assessment of techniques for protein structure prediction. http://predictioncenter.org/casp7/.
42. Punta B, Rost B. Profcon: novel prediction of long-range contacts. Bioinformatics 2005;21:2960–2968.
43. Cheng J, Baldi P. Three-stage prediction of protein β-sheets by neural networks, alignments and graph algorithms. Proc ISMB 2005, Bioinformatics 2005;21(suppl 1):i75–i84.
44. Cheng J, Baldi P. Improved residue contact prediction using support vector machines and a large feature set. BMC Bioinform 2007; 8:113–121.
45. Chan AW, Gail Hutchinson E, Harris D, Thornton JM. Identification, classifiction, and analysis of β-bulges in proteins. Protein Sci 1993;2:1574–1590.
46. Rhodes G. Crystallography made crystal clear, 2nd ed. San Diego: Academic Press; 2000.
47. Schlessinger A, Rost B. Protein flexibility and rigidity predicted from sequence. Proteins: Struct Funct Bioinform 2005;61: 115–126.