

# A generative, probabilistic model of local protein structure

Wouter Boomsma\*, Kanti V. Mardia<sup>†</sup>, Charles C. Taylor<sup>†</sup>, Jesper Ferkinghoff-Borg<sup>‡</sup>, Anders Krogh\*, and Thomas Hamelryck\*<sup>§</sup>

\*Bioinformatics Centre, Department of Biology, University of Copenhagen, Ole Maaloes Vej 5, 2200 Copenhagen N, Denmark; <sup>†</sup>Department of Statistics, University of Leeds, Leeds, West Yorkshire LS2 9JT, United Kingdom; and <sup>‡</sup>DTU Elektro, Technical University of Denmark, 2800 Lyngby, Denmark

Edited by David Baker, University of Washington, Seattle, WA, and approved March 14, 2008 (received for review February 27, 2008)

Despite significant progress in recent years, protein structure prediction maintains its status as one of the prime unsolved problems in computational biology. One of the key remaining challenges is an efficient probabilistic exploration of the structural space that correctly reflects the relative conformational stabilities. Here, we present a fully probabilistic, continuous model of local protein structure in atomic detail. The generative model makes efficient conformational sampling possible and provides a framework for the rigorous analysis of local sequence–structure correlations in the native state. Our method represents a significant theoretical and practical improvement over the widely used fragment assembly technique by avoiding the drawbacks associated with a discrete and nonprobabilistic approach.

conformational sampling | directional statistics | probabilistic model | TorusDBN | Bayesian network

Protein structure prediction remains one of the greatest challenges in computational biology. The problem itself is easily posed: predict the three-dimensional structure of a protein given its amino acid sequence. Significant progress has been made in the last decade, and, especially, knowledge-based methods are becoming increasingly accurate in predicting structures of small globular proteins (1). In such methods, an explicit treatment of local structure has proven to be an important ingredient. The search through conformational space can be greatly simplified through the restriction of the angular degrees of freedom in the protein backbone by allowing only angles that are known to appear in the native structures of real proteins. In practice, the angular preferences are typically enforced by using a technique called fragment assembly. The idea is to select a set of small structural fragments with strong sequence–structure relationships from the database of solved structures and subsequently assemble these building blocks to form complete structures. Although the idea was originally conceived in crystallography (2), it had a great impact on the protein structure-prediction field when it was first introduced a decade ago (3). Today, fragment assembly stands as one of the most important single steps forward in tertiary structure prediction, contributing significantly to the progress we have seen in this field in recent years (4, 5).

Despite their success, fragment-assembly approaches generally lack a proper statistical foundation, or equivalently, a consistent way to evaluate their contributions to the global free energy. When a fragment-assembly method is used, structure prediction normally proceeds by a Markov Chain Monte Carlo (MCMC) algorithm, where candidate structures are proposed by the fragment assembler and then accepted or rejected based on an energy function. The theoretical basis of MCMC is the existence of a stationary probability distribution dictating the transition probabilities of the Markov chain. In the context of statistical physics, this stationary distribution is given by the conformational free energy through the Boltzmann distribution. The problem with fragment-assembly methods is that it is not possible to evaluate the proposal probability of a given structure, which makes it difficult to ensure an unbiased sampling (which requires the property of detailed balance). Local

free energies could, in principle, be assigned to individual fragments, but there is no systematic way to combine them into a local free energy for an assembly of fragments. In fact, because of edge effects, the assembly process often introduces spurious local structural motifs that are not themselves present in the fragment library (3).

Significant progress has been made in the probabilistic modeling of local protein structure. With HMMSTR, Bystroff and coworkers (6) introduced a method to turn a fragment library into a probabilistic model but used a discretization of angular space, thereby sacrificing geometric detail. Other studies focused on strictly geometric models (7, 8). For these methods, the prime obstacle is their inability to condition the sampling on a given amino acid sequence. In general, it seems that none of these models has been sufficiently detailed or accurate to constitute a competitive alternative to fragment assembly. This is reflected in the latest CASP (critical assessment of techniques for protein structure prediction) exercise, where the majority of best performing *de novo* methods continue to rely on fragment assembly for local structure modeling (5).

Recently, we showed that a first-order Markov model forms an efficient probabilistic, generative model of the  $C\alpha$  geometry of proteins in continuous space (9). Although this model allows sampling of  $C\alpha$  traces, it is of limited use in high-resolution *de novo* structure prediction, because this requires the representation of the full atomic detail of a protein's backbone, and the mapping from  $C\alpha$  to backbone geometry is one-to-many. Consequently, this model also cannot be considered a direct alternative to the fragment-assembly technique.

In the present study, we propose a continuous probabilistic model of the local sequence–structure preferences of proteins in atomic detail. The backbone of a protein can be represented by a sequence of dihedral angle pairs,  $\phi$  and  $\psi$  (Fig. 1) that are well known from the Ramachandran plot (10). Two angles, both with values ranging from  $-180^\circ$  to  $180^\circ$ , define a point on the torus. Hence, the backbone structure of a protein can be fully parameterized as a sequence of such points. We use this insight to model the angular preferences in their natural space using a probability distribution on the torus and thereby avoid the traditional discretization of angles that characterizes many other models. The sequential dependencies along the chain are captured by using a dynamic Bayesian network (a generalization of a hidden Markov model), which emits angle pairs, amino acid labels, and secondary structure labels. This allows us to directly sample structures compatible with a given sequence and resample parts of a structure while maintaining consistency

Author contributions: W.B. and T.H. designed research; W.B. performed research; K.V.M. and C.C.T. contributed new reagents/analytic tools; and W.B., J.F.-B., A.K., and T.H. wrote the paper.

The authors declare no conflict of interest.

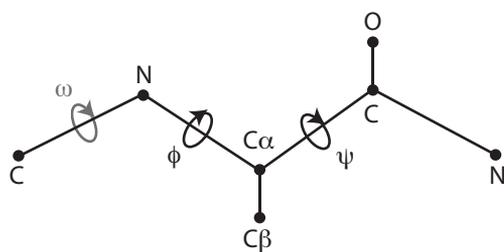
This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

<sup>§</sup>To whom correspondence should be addressed. E-mail: thamelry@binf.ku.dk.

This article contains supporting information online at [www.pnas.org/cgi/content/full/0801715105/DCSupplemental](http://www.pnas.org/cgi/content/full/0801715105/DCSupplemental).

© 2008 by The National Academy of Sciences of the USA



**Fig. 1.** The  $\phi$ ,  $\psi$  angular degrees of freedom in one residue of the protein backbone. The  $\omega$  dihedral angle can be assumed to be fixed at  $180^\circ$  (*trans*) or  $0^\circ$  (*cis*).

along the entire chain. In addition, the model makes it possible to evaluate the likelihood of any given structure. Generally, the sampled structures will not be globular, but will have realistic local structure, and the model can thus be used as a proposal distribution in structure-prediction simulations. The probabilistic and generative nature of the model also makes it directly applicable in the framework of statistical mechanics. In particular, because the probability before and after any resampling can be evaluated, unbiased sampling can be ensured.

We show that the proposed model accurately captures the angular preferences of protein backbones and successfully reproduces previously identified structural motifs. Finally, through a comparison with one of the leading fragment-assembly methods, we demonstrate that our model is highly accurate and efficient, and we conclude that our approach represents an attractive alternative to the use of fragment libraries in *de novo* protein-structure prediction.

## Results and Discussion

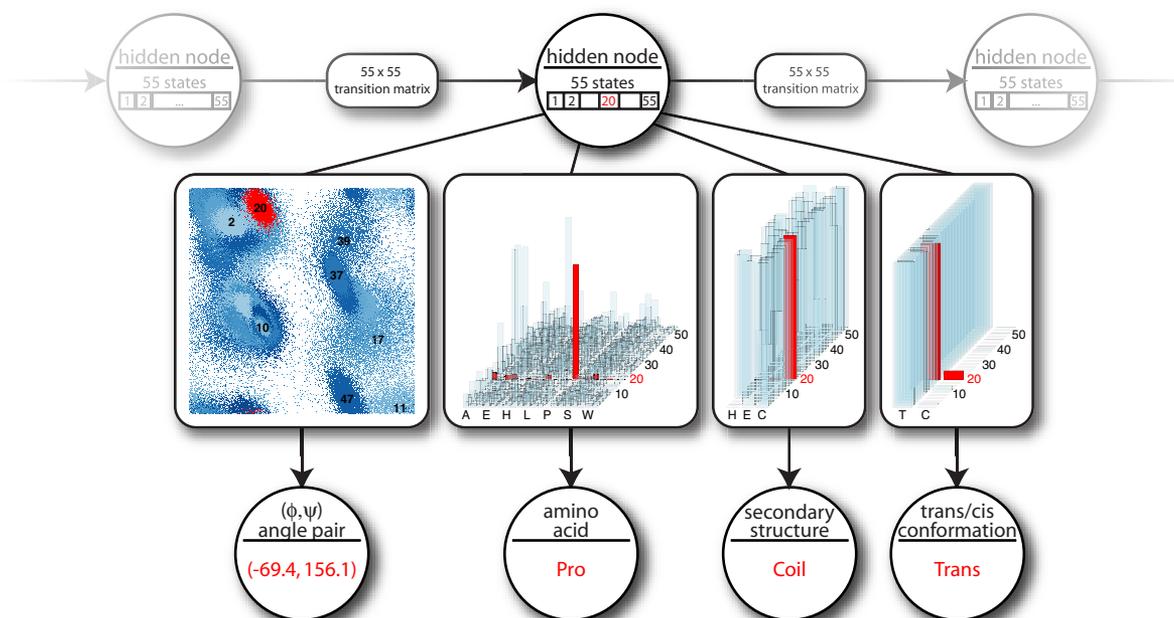
**TorusDBN—A Model of Protein Local Structure.** Considering only the backbone, each residue in a protein chain can be represented by using two angular degrees of freedom, the  $\phi$  and  $\psi$  dihedral bond

angles (Fig. 1). The bond lengths and all remaining angles can be assumed to have fixed values (11). Even with this simple representation, the conformational search space is extremely large. However, as Ramachandran and coworkers (10) noted in 1963, not all values of  $\phi$  and  $\psi$  are equally frequent, and many combinations are never observed because of steric constraints. In addition, strong sequential dependencies exist between the angle pairs along the chain. We define it as our goal to model precisely these local preferences.

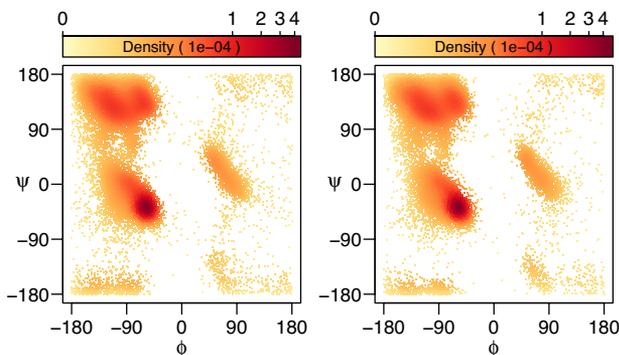
We begin by stating a few necessary conditions for the model. First, we require that, given an amino acid sequence, our model should produce protein backbone chains with plausible local structure. In particular, the parameterization used in our model should be sufficiently accurate to allow direct sampling and the construction of complete protein backbones. Note that we do not expect sampled structures to be correctly folded globular proteins—we only require them to have realistic local structure. Secondly, it should be possible to seamlessly replace any stretch of a protein backbone with an alternative segment, thus making a small step in conformational space. Finally, we require that it is possible to compare the probability of a newly sampled candidate segment with the probability of the original segment, which is needed to enforce the property of detailed balance in MCMC simulations.

The resulting model is presented in Fig. 2. Formulated as a dynamic Bayesian network (DBN), it is a probabilistic model that ensures sequential dependencies through a sequence of hidden nodes. A hidden node represents a residue at a specific position in a protein chain. It is a discrete node that can adopt 55 states (see *Methods*). Each of these states, or  $h$  values, corresponds to a distinct emission distribution over dihedral angles [ $d = (\phi, \psi)$ ], amino acids ( $a$ ), secondary structure ( $s$ ), and the *cis* or *trans* conformation of the peptide bond ( $c$ ). The angular emissions are modeled by bivariate von Mises distributions, whereas the  $\omega$  dihedral angle (Fig. 1) is fixed at either  $180^\circ$  or  $0^\circ$ , depending on the *trans/cis* flag. Note that this model can also be regarded as a hidden Markov model with multiple outputs.

The joint probability of the model is a sum over each possible



**Fig. 2.** The TorusDBN model. The circular nodes represent stochastic variables, whereas the rectangular boxes along the arrows illustrate the nature of the conditional probability distribution between them. The lack of an arrow between two nodes denotes that they are conditionally independent. A hidden node emits angle pairs, amino acid information, secondary structure labels (H, helix; E, strand; C, coil) and *cis/trans* information. One arbitrary hidden node value is highlighted in red and demonstrates how the hidden node value controls which mixture component is chosen.



**Fig. 3.** Ramachandran plots displaying the distribution of the 42,654 angle pairs in the test set (*Left*), and an equal number of angle pairs from sampled proteins from the model (*Right*).

hidden node sequence  $\mathbf{h} = \{h_1, \dots, h_N\}$ , where  $N$  denotes the length of the protein:

$$P(d, a, s, c) = \sum_{\mathbf{h}} P(d|\mathbf{h})P(a|\mathbf{h})P(s|\mathbf{h})P(c|\mathbf{h})P(\mathbf{h}) = \sum_{\mathbf{h}} \prod_i P(d_i|h_i)P(a_i|h_i)P(s_i|h_i)P(c_i|h_i)P(h_i|h_{i-1}).$$

The four types of emission nodes ( $d$ ,  $a$ ,  $s$ , and  $c$ ) can each be used either as input or output. In most cases, some input information is available (e.g., the amino acid sequence), and the corresponding emission nodes are subsequently fixed to specific values. These nodes are referred to as observed nodes. Sampling from the model then involves two steps: (i) sampling a hidden node sequence conditioned on the set of observed nodes and (ii) sampling emission values for the unobserved nodes conditioned on the hidden-node sequence. The first step is most efficiently solved by using the forward-backtrack algorithm (12, 9) [see [supporting information \(SI\) Text](#)], which allows for the resampling of any segment of a chain. This resembles fragment insertion in fragment assembly-based methods, but the forward-backtrack approach has the advantage that it ensures a seamless resampling that correctly handles the transitions at the ends of the segment. Once a particular sequence of hidden node values has been obtained, emission values for the unobserved nodes are drawn from the corresponding

conditional probability distributions (step ii). This is illustrated in Fig. 2, where the emission probability distributions for a particular  $h$  value are highlighted.

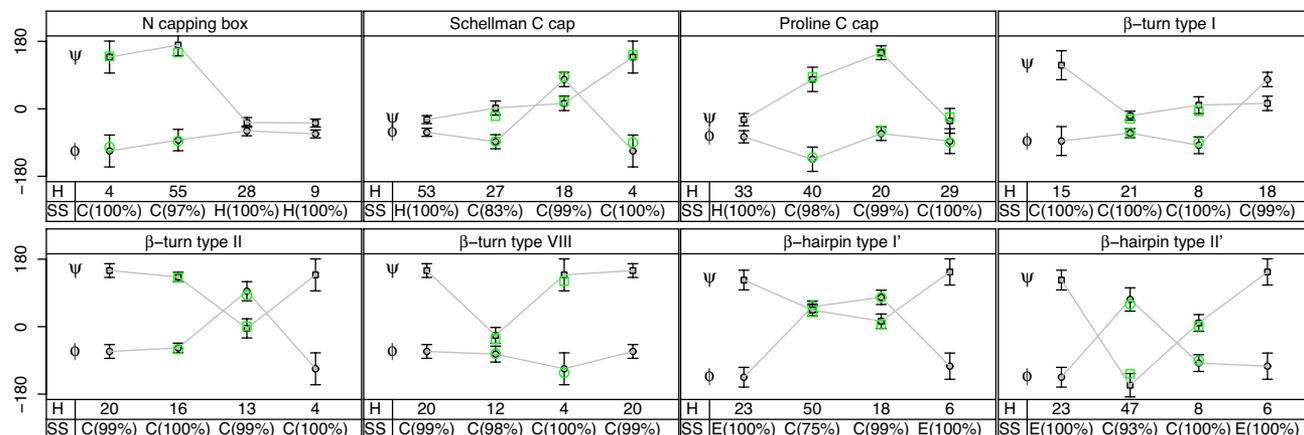
The parameters of the model were estimated from the SABmark 1.65 (13) dataset (see *Methods*). From the 1,723 proteins, 276 were excluded during training and used for testing purposes (test set).

We conducted a series of experiments to evaluate the model's performance. Throughout this article, we will be comparing the results obtained with our model (TorusDBN) to the results achieved with one of the most successful fragment assembly-based methods currently available, the Rosetta fragment assembler (3). Because our interest in this study is limited to modeling local protein structure, we exclusively enabled Rosetta's initial fragment-assembly phase, disabling any energy evaluations apart from clash detection. In all cases, as input to Rosetta, we used the amino acid sequence of the query structure, multiple sequence information from PSI-BLAST (14), and a predicted secondary structure sequence using PSIPRED (15).

**Angular Preferences.** As a standard quality check of protein structure, a Ramachandran plot is often used by crystallographers to detect possible angular outliers. We investigated how closely the Ramachandran plot of samples from our model matched the Ramachandran plot for the corresponding native structures.

For each protein in the test set, we extracted the amino acid sequence, and calculated a predicted secondary structure labeling using PSIPRED. We then sampled a single structure using the sequence and secondary structure labels as input and summarized the sampled angle pairs in a 2D histogram. Fig. 3 shows the histograms for the test set and the samples, respectively. The results are strikingly similar. Although the experiment reveals little about the detailed sequence-structure signal in our model, it provides a first indication that a mixture of bivariate von Mises distributions is an appropriate choice to model the angular preferences of the Ramachandran plot.

We proceeded with a comparison to Rosetta. For each protein in the test set, we created a single structure using Rosetta's fragment assembler and compared the resulting histogram to that of the test set. Also in this case, the produced plot is visually indistinguishable from the native one (plot not shown). However, by using the Kullback-Leibler (KL) divergence, a standard measure of distance between probability distributions, it becomes clear that the Ramachandran plot produced by the TorusDBN is closer to native than the plot produced by Rosetta (see *SI Text* and [Table S1](#)).



**Fig. 4.** Hidden node paths corresponding to well known structural motifs. The angular preferences for the hidden node paths are illustrated by using the mean  $\phi$  ( $\psi$ ) value as a circle (square), with the error bars denoting 1.96 standard deviations of the corresponding bivariate von Mises distribution. Because the angular distributions are approximately Gaussian at high concentrations, this corresponds to  $\approx 95\%$  of the angular samples from that distribution. In cases where ideal angular preferences for these motifs are known from the literature, they are specified in green. H, hidden node sequence; SS, secondary structure labeling (H, helix; E, strand; C, coil), with corresponding emission probabilities in parentheses.

**Table 1. Amino acid propensities for turn motifs calculated by using TorusDBN**

Name	Position	Input		Output		
		$(\phi, \psi)$	SS	AA		
$\beta$ -Turn type I	1	(-60, -30)	C	P (3.2130)	S (1.5816)	E (1.3680)
	2	(-90, 0)	C	D (2.4864)	N (2.1854)	S (1.5417)
$\beta$ -Turn type II	1	(-60, 120)	C	P (3.9598)	K (1.4291)	E (1.4234)
	2	(80, 0)	C	G (10.6031)	N (1.0152)	
$\beta$ -Turn type VIII	1	(-60, -30)	C	P (3.4599)	S (1.3431)	D (1.3290)
	2	(-120, 120)	C	V (1.9028)	I (1.8459)	F (1.3373)
$\beta$ -Hairpin type I'	1	(60, 30)	C	N (5.9596)	D (2.3904)	H (1.6610)
	2	(90, 0)	C	G (12.4208)		
$\beta$ -Hairpin type II'	1	(60, -120)	C	G (11.2226)		
	2	(-80, 0)	C	N (2.9914)	D (2.8430)	H (1.5844)

The propensity of a particular amino acid (columns 5–7) at a certain position (column 2) in a motif (column 1) is calculated as the posterior probability  $P(a|d, s)$  divided by the probability of that amino acid according to the stationary distribution  $P(a)$  of the model. Angular and secondary structure input are listed in columns 3 and 4. The three most preferred amino acids (with propensities  $>1$ ) are reported.

**Structural Motifs.** The TorusDBN models the sequential dependencies along the protein backbone through a first-order Markov chain of hidden states. In such a model, we expect longer range dependencies to be modeled as designated high-probability paths through the model.

By manually inspecting the paths of length 4 with highest probability according to the model (based on their transition probabilities), we indeed recovered several well known structural motifs. Fig. 4 demonstrates how eight well known structural motifs appear as such paths in the model. Both the emitted angle pairs (Fig. 4) and the amino acid preferences (Fig. S1) have good correspondence with the literature (16, 17) (see *SI Text*). All reported paths are among the 0.25% most probable 4-state paths in the model (out of the  $55^4$  possible paths).

Often, structural motifs will arise from combinations of several hidden node paths. By summing over the contributions of all possible paths [posterior decoding (18)], it is possible to extract this information from the model. To illustrate, we reversed the analysis of the structural motifs, by giving the ideal angles and secondary structure labeling of a motif as input to the model, and calculating the posterior distribution over amino acids at each position. Table 1 lists the top three preferred amino acids for each position in the different  $\beta$ -turn motifs. All of these amino acids have previously been reported to have high propensities at their specific positions (17).

**Sampling Structures.** We conclude with a demonstration of the model's performance beyond the scope of well defined structural

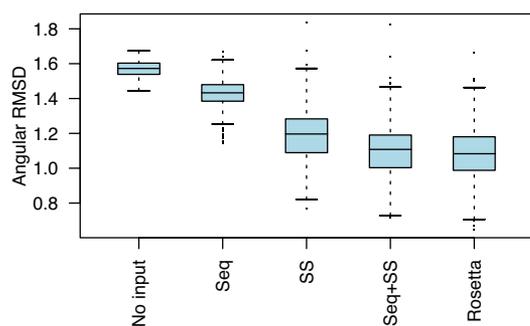
motifs. In the context of *de novo* structure prediction, the role of the model is that of a proposal distribution, where repeated resampling of angles should lead to an efficient exploration of conformational space. In this final experiment, we therefore sampled dihedral angles for the proteins in our test set and investigated how closely the sampled angles match those of the native state.

For each protein in the test set, 100 structures were sampled, and the average angular deviation was recorded (see *SI Text*). This was done for an increasing amount of input to the model. Initially, samples were generated without using input information, resulting in unrestricted samples from the model. We then included the amino acid sequence of the protein, a predicted secondary structure labeling (using PSIPRED), and, finally, a combination of both. We ran the same test with Rosetta's fragment assembler for comparison.

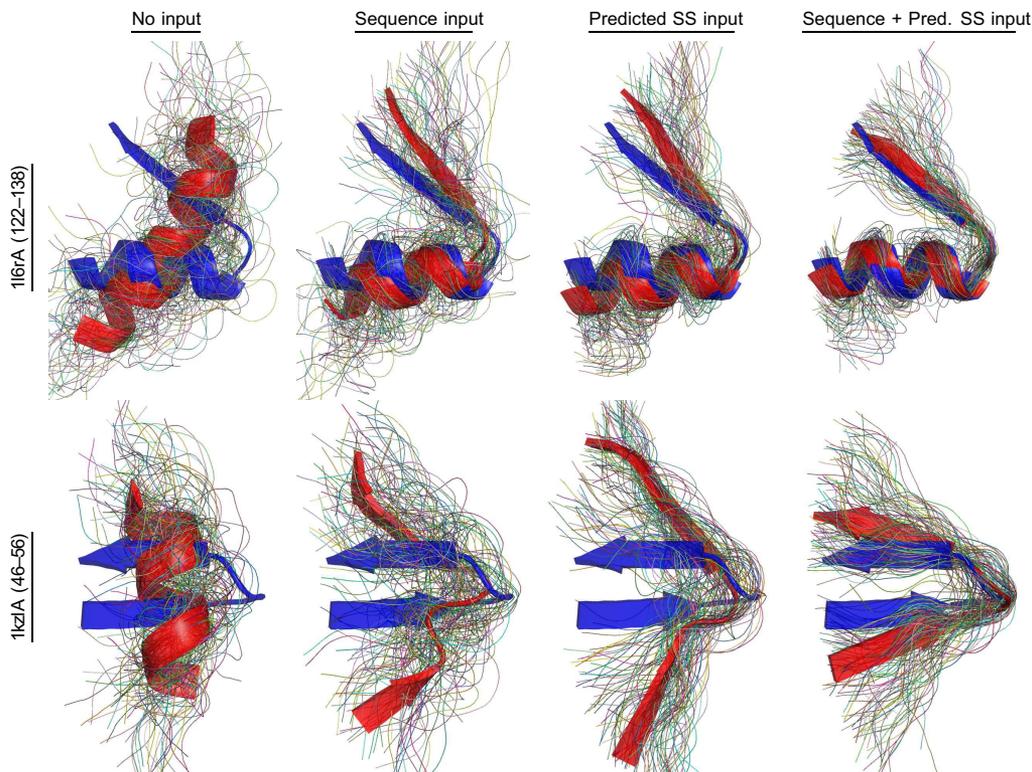
Fig. 5 shows the distribution of the average angular distance over all proteins in the test set. Clearly, as more information becomes available, the samples lie more closely around the native state. When both amino acid and secondary structure information is used, the performance of the TorusDBN approaches that of the fragment assembler in Rosetta. Recall that Rosetta also uses both amino acid and secondary structure information in its predictions but, in addition, incorporates multiple sequence information directly, which TorusDBN does not. In this light, our model performs remarkably well in this comparison. The time necessary to generate a single sample, averaged over all of the proteins in the test set, was 0.08 s for our model and 1.30 s for Rosetta's fragment assembler. All experiments were run on a 2,800 MHz AMD Opteron processor.

To illustrate the effect of the different degrees of input, we include a graphical view of two representative fragments extracted from the samples on the test set (Fig. 6). Note how the sequence and secondary structure input provide distinct signals to the model. In the hairpin motif, the sequence-only signal creates structures with an excess of coil states around the hairpin, whereas the inclusion of only secondary structure input gets the secondary structure elements right but fails to make the turn correctly. Finally, with both inputs, the secondary structure boundaries of the motif are correct, and the quality of the turn is enhanced through the knowledge that the sequence motif Asp-Gly is found at the two coil positions, which is common for a type I' hairpin (17).

**Additional Evaluations.** We conducted several additional experiments to evaluate other aspects of the model. First, we performed a detailed evaluation of TorusDBN's performance on local structure motifs using the I-sites library (19) (*SI Text* and Figs. S2–S4).



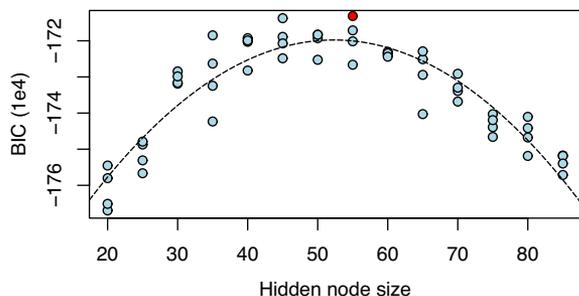
**Fig. 5.** Box-plots of the average angular deviation in radians (see *SI Text*) between native structures from the test set, and 100 sampled structures. From left to right, an increasing amount of information was given to the model: No input data, amino acid input data (Seq), predicted secondary structure input data (SS), and a combination of both (Seq+SS). The rightmost box corresponds to candidate structures generated by the fragment assembler in Rosetta.



**Fig. 6.** Two representative examples of samples generated by TorusDBN on the proteins in the test set (116rA, position 122–138 and 1kzIA, position 46–56). Each image contains the native structure in blue and a cloud of 100 sampled structures. The sampled structure with minimum average distance to all other samples is chosen as representative and highlighted in red. From left to right, an increasing amount of input is given to the model. Note, that the leftmost structures are sampled without any input information and are therefore not specific to these proteins. They are included here merely as a null model. Figures were created by using Pymol (29).

Second, we compared TorusDBN directly to HMMSTR in the recognition of decoy structures from native (*SI Text* and [Tables S2 and S3](#)), and finally, the length distributions of secondary structure elements in samples were analyzed (*SI Text* and [Fig. S5](#)). All these studies lend further support to the quality of the model.

**Potential Applications.** In closing, we list a few potential applications for the described model. First and foremost, it is in the context of *de novo* predictions that we expect the greatest benefits from our model. Seamless resampling and probability evaluations of proposed structures should provide a better sampling of conformational space, allowing calculations of thermodynamical averages in MCMC simulations (20). There are, however, several other potential areas of application. (i) Homology modeling, where the model is potentially useful as a proposal distribution for loop closure tasks; (ii) quality verifi-



**Fig. 7.** BIC values for models with varying hidden node size. For each size, four independent models were trained. The model used for our analyses is highlighted in red.

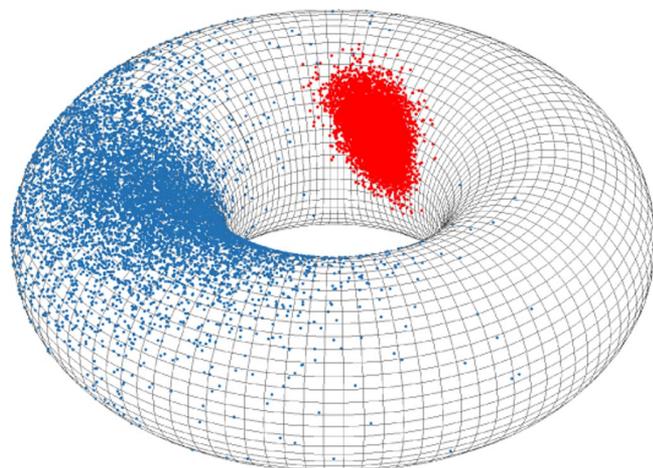
cation of experimentally determined protein structures, where it is likely that the sequential signal in our model constitutes an advantage over the current widespread use of Ramachandran plots to detect outliers; and (iii) protein design, where the model might be used to predict or sample amino acid sequences that are locally compatible with a given structure (as was demonstrated for short motifs in [Table 1](#)).

## Methods

**Parameter Estimation.** The model was trained by using the Mocapy DBN toolkit (21). As training data, we used the SABmark 1.65 twilight protein dataset, which for each different SCOP-fold provides a set of structures with low sequence similarity (13). Training was done on structures from 180 randomly selected folds (1,447 proteins, 226,338 observations), whereas the remaining 29 folds (276 proteins, 42,654 observations) were used as a test set. Amino acid, *trans/cis* peptide bond, and angle pair information was extracted directly from the training data, whereas secondary structure was computed by using DSSP (22).

Because the hidden node values are inherently unobserved, an algorithm capable of dealing with missing data is required. Here, we used a stochastic version of the well known expectation-maximization (EM) algorithm (23, 24). The idea behind stochastic EM (25, 26) is to first fill in plausible values for all unobserved nodes (E-step), and then update the parameters as if the model was fully observed (M-step). Just as with classic EM, these two steps are repeated until the algorithm converges. In our case, for each observation in the training set, we sampled a corresponding  $h$  value, using a single sweep of Gibbs sampling: in random order, all  $h$  values were resampled based on their current left and right neighboring  $h$  values and the observed emission values at that residue. Computationally, stochastic EM is more efficient than classic EM. Furthermore, on large datasets, stochastic EM is known to avoid convergence to local maxima (26).

The optimal size of the hidden node (i.e., the number of states that it can adopt) is a hyperparameter that is not automatically estimated by the EM procedure. We optimized this parameter by training models for a range of sizes, evaluating the likelihood for each model using the forward algorithm (18). Because the training procedure is stochastic in nature, we repeated this proce-



**Fig. 8.** Samples from two bivariate von Mises distributions, corresponding to two hidden nodes states. The red samples ( $h$ -value 20) represent a highly concentrated distribution ( $\kappa_1 = 65.4$ ,  $\kappa_2 = 45.7$ ,  $\kappa_3 = 17.3$ ,  $\mu = -66.2$ ,  $\nu = 149.6$ ), whereas the blue samples ( $h$ -value 39) are drawn from a less concentrated distribution ( $\kappa_1 = 3.6$ ,  $\kappa_2 = 1.9$ ,  $\kappa_3 = -0.8$ ,  $\mu = 67.4$ ,  $\nu = 96.2$ ).

ture several times. The best model was selected by using the Bayesian Information Criterion (BIC) (27), a score based on likelihood, which penalizes an excess of parameters and thereby avoids overfitting (see *SI Text*). As displayed in Fig. 7, the BIC reaches a maximum at a hidden node size of  $\approx 55$ . The model, however, appears to be quite stable with regard to the choice of this parameter. Several of the experiments in our study were repeated with different  $h$  size models (size 40–80) without substantially affecting the results.

- Dill KA, Ozkan SB, Weikel TR, Chodera JD, Voelz VA (2007) The protein folding problem: When will it be solved? *Curr Opin Struct Biol* 17:342–346.
- Jones TA, Thirup S (1986) Using known substructures in protein model building and crystallography. *EMBO J* 5:819–822.
- Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol* 268:209–225.
- Chikenji G, Fujitsuka Y, Takada S (2006) Shaping up the protein folding funnel by local interaction: Lesson from a structure prediction study. *Proc Natl Acad Sci USA* 103:3141–3146.
- Jauch R, Yeo H, Kolarik P, Clarke N (2007) Assessment of CASP7 structure predictions for template free targets. *Proteins* 69 Suppl 8:57–67.
- Bystrhoff C, Thorsson V, Baker D (2000) HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J Mol Biol* 301:173–190.
- Edgoose T, Allison L, Dowe DL (1998) An MML classification of protein structure that knows about angles and sequence. *Pac Symp Biocomput* 3:585–96.
- Camproux AC, Tuffery P, Chevrolat JP, Boisvieux JF, Hazout S (1999) Hidden Markov model approach for identifying the modular framework of the protein backbone. *Protein Eng Des Sel* 12:1063–1073.
- Hamelryck T, Kent JT, Krogh A (2006) Sampling realistic protein conformations using local structural bias. *PLoS Comput Biol* 2:e131.
- Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. *J Mol Biol* 7:95–99.
- Engh RA, Huber R (1991) Accurate bond and angle parameters for x-ray protein structure refinement. *Acta Crystallogr A* 47:392–400.
- Cawley SL, Pachter L (2003) HMM sampling and applications to gene finding and alternative splicing. *Bioinformatics* 19 Suppl 2:ii36–ii41.
- Van Walle I, Lasters I, Wyns L (2005) SABmark—a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics* 21:1267–1268.

**Angular Probability Distribution.** The Ramachandran plot is well known in crystallography and biochemistry. The plot is usually drawn as a projection onto the plane, but because of the periodicity of the angular degrees of freedom, the natural space for these angle pairs is on the torus. To capture the angular preferences of protein backbones, a mixture of Gaussian-like distributions on this surface is therefore an appropriate choice. We turned to the field of directional statistics for a bivariate angular distribution with Gaussian-like properties that allows for efficient sampling and parameter estimation. From the family of bivariate von Mises distributions, we chose the cosine variant, which was especially developed for this purpose by Mardia *et al.* (28). The density function is given by

$$f(\phi, \psi) = c(\kappa_1, \kappa_2, \kappa_3) \exp(\kappa_1 \cos(\phi - \mu) + \kappa_2 \cos(\psi - \nu) - \kappa_3 \cos(\phi - \mu - \psi + \nu)). \quad [1]$$

The distribution has five parameters:  $\mu$  and  $\nu$  are the respective means for  $\phi$  and  $\psi$ ,  $\kappa_1$  and  $\kappa_2$  their concentration, and  $\kappa_3$  is related to their correlation (Fig. 8). The parameters can be efficiently estimated by using a moment-estimation technique. Efficient sampling from the distribution is achieved by rejection sampling, using a mixture of two von Mises distributions as a proposal distribution (see *SI Text*).

**Availability.** The TorusDBN model is implemented as part of the backboneDBN package, which is freely available at <http://sourceforge.net/projects/phaistos/>.

**ACKNOWLEDGMENTS.** We thank Mikael Borg, Jes Frellsen, Tim Harder, Kasper Stovgaard, and Lucia Ferrari for valuable suggestions to the paper; John Kent for discussions on the angular distributions; Christopher Bystrhoff for help with HMMSTR and the newest version of I-sites; and the Bioinformatics Centre and the Zoological Museum, University of Copenhagen, for use of their cluster computer. W.B. was supported by the Lundbeck Foundation, and T.H. was funded by Forskningsrådet for Teknologi og Produktion (“Data Driven Protein Structure Prediction”).

- Altschul SF, *et al.* (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
- Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 292:195–202.
- Aurora R, Rose GD (1998) Helix capping. *Protein Sci* 7:21–38.
- Hutchinson EG, Thornton JM (1994) A revised set of potentials for  $\beta$ -turn formation in proteins. *Protein Sci* 3:2207–2216.
- Durbin R, Eddy SR, Krogh A, Mitchison G (1998) *Biological Sequence Analysis* (Cambridge Univ Press, Cambridge, UK).
- Bystrhoff C, Baker D (1998) Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol* 281:565–577.
- Winther O, Krogh A (2004) Teaching computers to fold proteins. *Phys Rev E* 70:30903.
- Hamelryck T (2007) *Mocapy: A Parallelized Toolkit for Learning and Inference in Dynamic Bayesian Networks. Manual* (Univ of Copenhagen, Copenhagen).
- Kabsch W, Sander C (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637.
- Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B* 39:1–38.
- Ghahramani Z (1998) Learning dynamic Bayesian networks. *Lect Notes Comput Sci* 1387:168–197.
- Diebolt J, Ip EHS (1996) *Markov Chain Monte Carlo in Practice*, eds Gilks WR, Richardson S, Spiegelhalter DJ (Chapman & Hall/CRC), pp 259–273.
- Nielsen SF (2000) The stochastic EM algorithm: Estimation and asymptotic results. *Bernoulli* 6:457–489.
- Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464.
- Mardia KV, Taylor CC, Subramaniam GK (2007) Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. *Biometrics* 63:505–512.
- DeLano WL (2002) *The PyMOL User's Manual* (DeLano Scientific, San Carlos, CA).