Bayesian Segmental Models with Multiple Sequence Alignment Profiles for Protein Secondary Structure and Contact Map Prediction

Wei Chu, Zoubin Ghahramani, Alexei Podtelezhnikov, David L. Wild

The authors would like to thank the Institute of Applied Mathematics (IPAM) at UCLA, where part of this work was carried out. This work was supported by the National Institutes of Health Grant Number 1 P01 GM63208. Wei Chu and Zoubin Ghahramani are with Gatsby Computational Neuroscience Unit, University College London, London WC1N 3AR, UK. Alexei Podtelezhnikov and David L. Wild are with the Keck Graduate Institute of Applied Life Sciences, Claremont, CA 91711, USA. All the correspondence should be addressed to David L. Wild, whose email address is *david_wild@kgi.edu*.

Abstract

In this paper, we develop a segmental semi-Markov model (SSMM) for protein secondary structure prediction which incorporates multiple sequence alignment profiles with the purpose of improving the predictive performance. The segmental model is a generalization of the hidden Markov model where a hidden state generates segments of various length and secondary structure type. A novel parameterized model is proposed for the likelihood function that explicitly represents multiple sequence alignment profiles to capture the segmental conformation. Numerical results on benchmark data sets show that incorporating the profiles results in substantial improvements and the generalization performance is promising. By incorporating the information from long range interactions in β-sheets, this model is also capable of carrying out inference on contact maps. This is an important advantage of probabilistic generative models over the traditional discriminative approach to protein secondary structure prediction. The web server of our algorithm and supplementary materials are available at http://public.kgi.edu/$\sim$wild/bsm.html.

Index Terms

Bayesian segmental semi-Markov models, generative models, protein secondary structure, contact maps, multiple sequence alignment profiles, parametric models.

I. Introduction

Protein secondary structure prediction remains an important step on the way to full tertiary structure prediction in both fold recognition (threading) and ab-initio methods, as well as providing useful information for the design of site directed mutagenesis experiments to elucidate protein function. A variety of approaches have been proposed to derive the secondary structure of a protein from its amino acid sequence as a classification problem. Beginning with the seminal work of Qian and Sejnowski [1], many of these methods have utilized neural networks. A major improvement in the prediction accuracy of these methods was made by Rost and Sander [2], who proposed a prediction scheme using multi-layered neural networks, known as PHD. The key novel aspect of this work was the use of evolutionary information in the form of profiles derived from multiple sequence alignments instead of training the networks on single sequences. Another kind of alignment profile, position-specific scoring matrices (PSSM) derived by the iterative search procedure PSI-BLAST [3], has been used in neural network prediction methods to achieve further improvements in accuracy [4] [5].

All the above approaches treat the secondary structure prediction problem as a supervised discriminative classification problem. An alternative approach is to treat the problem from the perspective of generative models. One of the first applications of hidden Markov models (HMMs) to the secondary structure prediction problem was described by Delcher et al. [6]. Generalized HMMs with explicit state duration, also known as segmental semi-Markov models, have been widely applied in the field of gene identification [7] [8] [9] [10]. Recently, Schmidler [11] presented an interesting statistical generative model for protein structure prediction, based on a segmental semi-Markov model (SSMM) [12] for sequence-structure relationships. The SSMM is a generalization of hidden Markov models that allows each hidden state to generate a variable length sequence of observations. One advantage of such a probabilistic framework is that it is possible to incorporate varied sources of sequence information using a joint sequence-structure probability distribution based on structural segments. Secondary structure prediction can then be formulated as a general Bayesian inference problem. However, the secondary structure prediction accuracy of the SSMM as described by Schmidler [11] falls short of the best contemporary discriminative methods. As suggested by Schmidler et al. [11], incorporation of multiple alignment profiles into the model might be a plausible way to improve the performance.

Whilst we draw heavily on the work of Schmidler et al. [11] [13] [14], our paper makes several original contributions to the SSMM model, which can be summarized as follows:

1) We incorporate evolutionary information in the form of multiple sequence alignment profiles into the SSMM model. We observed that this extension results in substantial improvements in performance ($\approx 7\%$) over Schmidler's original model [11] and makes the SSMM approach competitive with other contemporary methods for secondary structure prediction. We believe that this is a significant result given the interest in improving structure prediction techniques in the field.
2) We propose a novel parameterized model as the likelihood function for the SSMM model in order to exploit the information provided by the multiple sequence alignment profiles. Schmidler et al. [13] used lookup tables for conditional discrete random variables in the

SSMM model. This limits the dependency modelling since the number of parameters of the lookup table grows exponentially with the length of the dependency window. The number of parameters in the model we propose (see equation 8) grows linearly with the length of the dependency window. We also apply the concept of a "product of experts" [15] to combine contributions from the segmental dependency and helical capping signals in the likelihood evaluation (see equation 4).

3) We propose a novel definition of beta sheet contact space that allows for efficient sampling of long-range contacts under the realistic constraint that one amino acid can have at most two hydrogen bonds and unpaired beta strands are not possible. This ability to infer contact maps is one of the principal advantages of a probabilistic modelling approach over the traditional discriminative approach to protein secondary structure prediction.

The paper is organized as follows. We describe the Bayesian framework of the SSMM in section II. In section III we extend the model to incorporate long range interactions, and point out the capability to infer contact maps. In section IV we discuss the issue of parameter estimation in detail. In section V we describe a general sampling scheme for prediction. In section VI we present the results of numerical experiments, and conclude in section VII.

II. Bayesian Modelling Framework

The key concept underlying our modelling approach is the notion of proteins as collections of local structural fragments, or segments, which may be shared by unrelated proteins - an approach which has gained increasing currency in the protein modelling community in recent years [16], [17]. The modelling framework we adopt is that of the segmental semi-Markov model (SSMM) [11], [12] - a generalization of hidden Markov models that allows each hidden state to generate a variable length sequence of the observations. For uniformity, we follow the definitions first established by Schmidler et al. [13] in the context of protein secondary structure.

The observation sequence includes both a residue sequence and a multiple alignment profile for each protein chain, and is denoted as $O = [O_1, O_2, \ldots, O_i, \ldots, O_n]$. The associated secondary structure can be fully specified in terms of segment locations and segment types. The segment locations can be identified by the positions of the last residue of these segments, denoted as

$e = [e_1, e_2, \ldots, e_m]$, where m is the number of segments. We use three secondary structure types. The set of secondary structure types is denoted as $\mathcal{T} = \{H, E, C\}$ where H is used for α-helix, E for β-strand and C for Coil. The sequence of segment types can be denoted as $T = [T_1, T_2, \ldots, T_i, \ldots, T_m]$ with $T_i \in \mathcal{T}\ \forall i$. In Figure 1, we present an illustration for the specification of the secondary structure of an observed sequence. Based on a set of protein chains with known secondary structure, we learn an explicit probabilistic model for sequence-structure relationships in the form of a segmental semi-Markov model.

In this approach, the segment types are regarded as a set of discrete variables, known as states. Each of the segment types possesses an underlying generator, which generates a variable-length sequence of observations, i.e. a segment. A schematic depiction of the SSMM is presented in Figure 2 from the perspective of generative models. The variables (m, e, T) describe the secondary structure segmentation of the sequence. In a Bayesian framework the secondary structure prediction problem then consists of computing the posterior probability, $\mathcal{P}(m, e, T|O)$ for an observed sequence O. For this purpose we need to define the prior probability $\mathcal{P}(m, e, T)$ and the likelihood $\mathcal{P}(O|m, e, T)$. This Bayesian framework is described in more detail in the following sections.

A. Multiple alignment profiles

In our model, each of the primary sequences of amino acid residues we are given, denoted as $R = [R_1, R_2, \ldots, R_i \ldots, R_n]$ with $R_i \in \mathcal{A}$ where $1 \leq i \leq n$ and $\mathcal{A}$ is the set of 20 amino acids, is associated with a profile derived by multiple sequence alignment [18] or PSI-BLAST [3].

- ***Multiple Sequence Alignment Profiles***: for a sequence of amino acid residues, we employ the techniques of pairwise sequence comparison to search a non-redundant protein sequence database for several other sequences which are similar enough at the sequence level to be evolutionarily related. These homologs are then aligned using standard multiple sequence alignment techniques [18]. Ideally, a row of aligned residues occupy similar structural positions and all diverge from a common ancestral residue. By counting the number of occurrences of each amino acid at each location, we obtain an alignment profile. Formally, the alignment profile $M = [M_1, M_2, \ldots, M_i, \ldots, M_n]$ is a sequence of 20×1 vectors, where

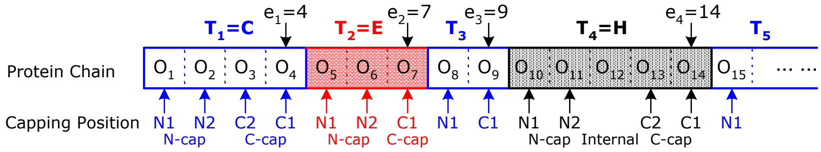


Fig. 1. Presentation of the secondary structure of a protein chain in terms of segments (adapted from Schmidler et al. [13]). The square blocks denote our observations of these amino acid residues. The rectangular blocks with solid borders denote the segments. The model represents the segment type $T = [C, E, C, H, \ldots]$ and the segmental endpoints $e = [4, 7, 9, 14, \ldots]$. Capping positions specify the N- and C-terminal positions within a segment. Here, both the N-capping and C-capping length are fixed at 2, and then $\{N1, N2, \text{Internal}, C2, C1\}$ are used to indicate the capping positions within a segment.

M_i contains the occurrence counts for the 20 amino acids at location i.

- ***Profiles from PSI-BLAST***: PSI-BLAST [3] is a gapped-version of BLAST that uses an effective scheme for weighting the contribution of different numbers of specific residues at each position in the sequence via a intermediate sequence profile, known as position-specific score matrix (PSSM). Jones [4] explored the idea of using this PSSM as a direct input to a secondary structure prediction method rather than extracting the homologous sequences and then producing a multiple sequence alignment. The PSSM from PSI-BLAST is a matrix with $20 \times n$ elements, where n is the length of the sequence, and each element represents the log-likelihood of the particular amino acid substitution at that position. The profile matrix elements can be mapped to relative occurrence counting by using the standard logistic function: $\frac{1}{1+\exp(-x)}$.

B. Prior Distribution

The prior distribution for the variables describing secondary structure $\mathcal{P}(m, e, T)$ is factored as

$$\mathcal{P}(m, e, T) = \mathcal{P}(m)\mathcal{P}(e, T|m) = \mathcal{P}(m) \prod_{i=1}^{m} \mathcal{P}(e_i|e_{i-1}, T_i)\mathcal{P}(T_i|T_{i-1}). \tag{1}$$

The segment type depends on the nearest previous neighbour in the sequence through the state transition probabilities $\mathcal{P}(T_i|T_{i-1})$, which are specified by a 3×3 transition matrix. $\mathcal{P}(e_i|e_{i-1}, T_i)$, more exactly $\mathcal{P}(l_i|T_i)$ where $l_i = e_i - e_{i-1}$, is the segmental length distribution of the type T_i.[1]

[1] $e_0 = 0$ is introduced as an auxiliary variable.

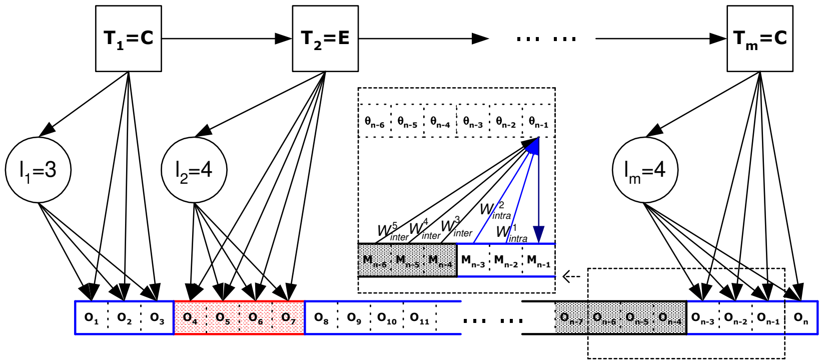


Fig. 2. The segmental semi-Markov model illustrated as generative processes. A variable-length segment of observations associated with random length l_i is generated by the state T_i. The observations within a segment need not be fully correlated, while there might be dependencies between the residues in adjacent segments. The dashed rectangle denotes the dependency window with length 5 for the observation O_{n-1}. In the enlarged dependency window, θ_{n-1} is a vector of latent variables that defines the multinomial distribution in which we observe M_{n-1}, while θ_{n-1} is assumed to be dependent on $M_{n-6}, \ldots, M_{n-2}$.

Note that the prior on length implicitly defines a prior on the number of segments m for a sequence of a given length. A uniform prior can be assigned for m, i.e. $\mathcal{P}(m) \propto 1$, as this does not have much effect on inference.

C. Likelihood Function

The likelihood is the probability of observing the sequence of alignment profiles given the set of random variables $\{m, e, T\}$. Generally, the probability of the observations can be evaluated as a product of the segments specified by $\{m, e, T\}$:

$$\mathcal{P}(O|m, e, T) = \prod_{i=1}^{m} \mathcal{P}(S_i|S_{-i}, T_i) \tag{2}$$

where $S_i = O_{[e_{i-1}+1:e_i]} = [O_{e_{i-1}+1}, O_{e_{i-1}+2}, \ldots, O_{e_i}]$ is the i-th segment, and $S_{-i} = [S_1, S_2, \ldots, S_{i-1}]$. The likelihood function $\mathcal{P}(S_i|S_{-i}, T_i)$ for each segment can be further factorized as a product of the conditional probabilities of individual observations,

$$\mathcal{P}(S_i|S_{-i}, T_i) = \prod_{k=e_{i-1}+1}^{e_i} \mathcal{P}(O_k|O_{[1:k-1]}, T_i) \tag{3}$$

where O_k is the pair of $\{R_k, M_k\}$. R_k is a column vector with 20 elements in which only one element is 1, indicating the amino acid type of the k-th residue, while others are 0, and M_k is the count vector obtained from the alignment profile. The likelihood function for each residue should be capable of capturing the core features of the segmental composition in the protein structure.

Schmidler [13] proposed a helical segment model with lookup tables to capture helical capping signals [19] and hydrophobicity dependency [20] in segmental residues. Aydin et al. [21] introduce an improved dependency model by considering the statistically significant correlation patterns at segment borders. The number of free parameters in these approaches is exponential with the length of dependency window. To overcome this drawback, Chu et al. [22] proposed an extended sigmoid belief network with parameterization for likelihood evaluation, in which the parameter number grows linearly with the length of the dependency window. However, these methods were designed to use the primary sequence only, and their secondary structure prediction accuracy still falls short of the best contemporary methods. Incorporation of multiple alignment profiles into the model could be a plausible way to improve the performance.

The plausibility of a protein structure should be evaluated from various perspectives, such as segmental dependency [20], helical capping signals [19] and steric restrictions [23] etc. It is hard to incorporate all the relevant perspectives by using a single model as the likelihood function. Therefore we adopt the concept of a "product of experts" [15] for likelihood evaluation. In the present work, we introduce two experts for the segmental dependency and the helical capping signals respectively. One is a novel parameterized model for segmental dependency that explicitly represents the multiple sequence alignment profile; another is a set of discrete distributions that captures helical capping signals. The conditional probabilities of individual observations can be evaluated in the form of a product of experts, i.e.

$$\mathcal{P}(O_k|O_{[1:k-1]},T_i) = \mathcal{P}(M_k|M_{[1:k-1]},T_i)\mathcal{P}(R_k|R_{[1:k-1]},T_i) \tag{4}$$

More details are given in the following.

1) An Expert for Segmental Dependency: The existence of correlated side chain mutations in α-helices has been well studied [20] [24]. These correlations in non-adjacent sequence positions are induced by their spatial proximity in the folded protein molecule and provide an important source of information about the underlying structure. We propose a novel parametric model to capture the segmental dependency by exploiting the information in the multiple sequence alignment profile.

a) Multinomial Distribution: We assume that M_k comes from a multinomial distribution with 20 possible outcomes and outcome probabilities θ_k, a 20×1 vector. The outcomes refer to the types of amino acids occurring at the current residue position, while M_k is a 20×1 vector counting the occurrence of these outcomes. Thus, the probability of getting M_k can be evaluated by

$$\mathcal{P}(M_k|\theta_k, T_i) = \frac{\Gamma(\sum_a M_k^a + 1)}{\prod_a \Gamma(M_k^a + 1)} \prod_{a \in \mathcal{A}} (\theta_k^a)^{M_k^a} \tag{5}$$

where $\mathcal{A}$ is the set of 20 amino acids, M_k^a is the element in M_k for the amino acid a, and θ_k^a denotes the probability of the outcome a with the constraint $\sum_a \theta_k^a = 1$. $\Gamma(\cdot)$ is the Gamma function defined as $\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t)\, dt$.[2]

b) Dirichlet Prior: As shown in the dependency window of Figure 2, the multinomial distribution at the k-th residue is dependent upon preceding observations within the dependency window, the segment type, and the current capping position within the segment. The underlying causal impact on the current multinomial distribution, where we observed M_k, can be captured by a prior distribution over the latent variables θ_k. A natural choice for the prior distribution over θ_k is a Dirichlet, which has also been used to define priors for protein family HMMs [25]. In our case, this can be explicitly parameterized by weight matrices with *positive elements* as follows:

$$\mathcal{P}(\theta_k|M_{[1:k-1]}, T_i) = \frac{\Gamma(\sum_a \gamma_k^a)}{\prod_a \Gamma(\gamma_k^a)} \prod_{a \in \mathcal{A}} (\theta_k^a)^{\gamma_k^a - 1} \tag{6}$$

where $\boldsymbol{\gamma}_k$ is a 20×1 vector defined as

$$\boldsymbol{\gamma}_k = W^0 + \sum_{j=1}^{\ell_k} W_{intra}^j M_{k-j} + \sum_{j=\ell_k+1}^{\ell} W_{inter}^j M_{k-j} \tag{7}$$

with ℓ is the length of dependency window,[3] $\ell_k = \min(k - e_{i-1} - 1, \ell)$,[4] and a 20×1 weight vector W^0 is used for local contributions. Weight matrices W_{intra} and W_{inter} of size

[2] Note that $\Gamma(x+1) = x!$ for positive integers x.

[3] The window length may be specified individually for segment types.

[4] $\min(a, b)$ means a if $a \leq b$, otherwise b.

20×20 are used to capture both intra-segmental and inter-segmental dependency respectively, where the superscript denotes the residue interval. The constraint $\gamma_k^a > 0\ \forall a$ is guaranteed by constraining the weight variables to have positive values. In total we have three sets of weights for $\tau \in \mathcal{T}$ individually. For a segment type τ, we get the set of weight parameters, $\boldsymbol{W}_\tau = \{W^0, W^1_{intra}, \ldots, W^\ell_{intra}, W^1_{inter}, \ldots, W^\ell_{inter}\}$.

c) Dirichlet-Multinomial Distribution: The quantity of interest, $\mathcal{P}(M_k|M_{[1:k-1]}, T_i)$ in (3), can be finally obtained as an integral over the space of the latent variables θ_k, which is given by

$$\begin{aligned} \mathcal{P}(M_k|M_{[1:k-1]}, T_i) &= \int_{\theta_k} \mathcal{P}(M_k|\theta_k, T_i)\mathcal{P}(\theta_k|M_{[1:k-1]}, T_i)\, d\theta_k \\ &= \frac{\Gamma(\sum_a \gamma_k^a) \cdot \prod_a \Gamma(\gamma_k^a + M_k^a)}{\Gamma\left(\sum_a(\gamma_k^a + M_k^a)\right) \cdot \prod_a \Gamma(\gamma_k^a)} \cdot \frac{\Gamma(\sum_a M_k^a + 1)}{\prod_a \Gamma(M_k^a + 1)} \end{aligned} \tag{8}$$

where $\Gamma(\cdot)$ denotes the Gamma function, and $\boldsymbol{\gamma}_k$ is defined as in (7).

2) An Expert for Helical Capping Signals: Helical capping signals [19] refer to the preference for particular amino acids at the N- and C-terminal ends which terminate helices through side chain-backbone hydrogen bonds or hydrophobic interactions. Thus amino acid distributions around the segment ends differ significantly from those of the internal positions, which provide important information for identifying α-helix segments in protein sequences.

The component $\mathcal{P}(R_k|R_{[1:k-1]}, T_i)$ in (3) can be simply modelled as $\mathcal{P}(R_k|T_i)$, which represents the probability of observing R_k at the particular capping position in segments with type T_i. The capping position of each residue within a segment can be determined uniquely (see Figure 1 for an illustration).[5] The probability distribution of amino acids on a specific capping position c in segments with type τ, denoted as $\mathcal{P}_c(R|\tau)$, can be directly estimated from the training data set, where $c \in \{N1, N2, \ldots, \text{Internal}, \ldots, C2, C1\}$, $R \in \mathcal{A}$ and $\tau \in \mathcal{T}$.

In summary, the segmental likelihood function we proposed can be explicitly written as

$$\mathcal{P}(O|m, e, T) = \prod_{i=1}^{m} \mathcal{P}(S_i|T_i, S_{-i}) = \prod_{i=1}^{m} \prod_{k=e_{i-1}+1}^{e_i} \mathcal{P}(M_k|M_{[1:k-1]}, T_i)\mathcal{P}_c(R_k|T_i) \tag{9}$$

where $\mathcal{P}(M_k|M_{[1:k-1]}, T_i)$ is defined as in (8), and $\mathcal{P}_c(R_k|T_i)$ is the position-specific distribution of capping signals.

[5] Note that we have used two sets of positioning indices for each residue: a sequential number k where $1 \le k \le n$, and a capping position *cap* where *cap* $\in \{\mathrm{N1}, \mathrm{N2}, \ldots, \mathrm{Internal}, \ldots, \mathrm{C2}, \mathrm{C1}\}$.

Winther and Krogh [26] have demonstrated that optimized potential functions learned from training data can provide very strong restrictions on the spatial arrangement of protein folding. As a very promising direction for future work, the introduction of an additional "steric expert" into our likelihood function could provide global restrictions on secondary structure and fulfill the potential of the Bayesian segmental model for tertiary structure prediction.

D. Posterior Distribution

All inferences about the segmental variables (m, e, T) defining secondary structure are derived from the posterior probability $\mathcal{P}(m, e, T|O)$. Using Bayes' theorem,

$$\mathcal{P}(m, e, T|O) = \frac{\mathcal{P}(O|m, e, T)\mathcal{P}(m, e, T)}{\mathcal{P}(O)} \tag{10}$$

where $\mathcal{P}(O) = \sum_{\{m,e,T\}} \mathcal{P}(O|m, e, T)\mathcal{P}(m, e, T)$ as the normalizing factor. From the posterior distribution over segmental variables $\mathcal{P}(m, e, T|O)$, we can obtain two different ways of estimating the secondary structure of a given sequence:

- The most probable segmental variables in the posterior distribution: $\arg\max\limits_{m,e,T} \mathcal{P}(m, e, T|O)$, known as the MAP estimate;
- The posterior distribution of the segment type at each residue: $\mathcal{P}(T_{O_i}|O)$ where we denote T_{O_i} as the segment type at the i-th observation. The marginal posterior mode estimate is defined as $\arg\max\limits_{T} \mathcal{P}(T_{O_i}|O)$.

The Viterbi and forward-backward algorithms for SSMM [27] [13] can be employed for the MAP and marginal posterior mode estimate respectively.

III. Incorporating Long Range Interactions in β-Sheets

We have set up a Bayesian framework to predict secondary structure. However, secondary structure might be affected not only by local sequence information, but also by long range interactions between distal regions of the amino acid sequence. This is particularly important in the case of β sheets, which are built up from several interacting regions of β-strands. The strands align so that the NH groups on one strand can form hydrogen bonds with the CO groups on the distal strand and vice versa. The alignment can happen in two ways: either the

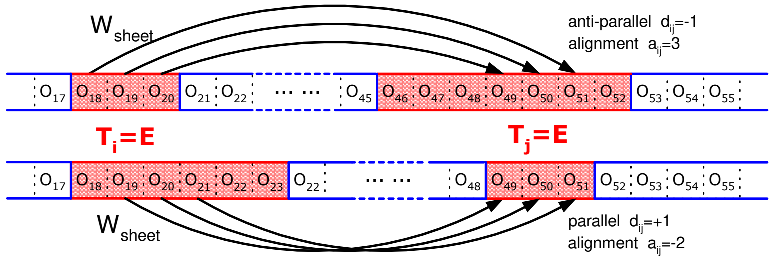


Fig. 3. Anti-parallel (top), and parallel (bottom), pairs of interacting segments, S_i and S_j. d_{ij} is the binary variable for alignment direction, and a_{ij} is the integer variable for alignment position. A weight matrix W_{sheet} is introduced to capture the distal residue interactions.

direction of the polypeptide chain of β-strands is identical, a *parallel* β-sheet, or the strand alignment is in the alternative direction, an *anti-parallel* β-sheet. In Figure 3, we present the two cases for a pair of interacting segments, S_i and S_j with $i < j$. A binary variable is used to indicate alignment direction; $d_{ij} = +1$ for parallel and $d_{ij} = -1$ for anti-parallel. A integer variable a_{ij} is used to indicate the alignment position. The endpoint of S_i, known as e_i, is used as the origin, and then a_{ij} is defined as the shift between e_i and e_j for parallel cases, while for anti-parallel cases it is the shift between e_i and the beginning point of S_j, i.e. $e_{j-1} + 1$.[6] The challenge for a predictive approach is how to introduce these long range interactions into the model. Incorporating these long-range dependencies into the SSMM model was pioneered by Schmidler et al. [11] [14]. In this section, we extend our parametric model to incorporate information on long range interactions in β-sheets.

A. *Prior Specification for Distal Interactions*

A set of random variables is introduced to describe the long range interactions, collected as $\mathcal{I} = \{\{S_j \leftrightarrow S_{j'}, d_{jj'}, a_{jj'}\}_{j=1}^{r}\}$, where r is the number of interacting pairs and $\{S_j \leftrightarrow S_{j'}, d_{jj'}, a_{jj'}\}$ is a pair of interacting segments together with their alignment information. We can expand the prior probability as $\mathcal{P}(m, e, T, \mathcal{I}) = \mathcal{P}(\mathcal{I}|m, e, T)\mathcal{P}(m, e, T)$, where $\mathcal{P}(m, e, T)$ is defined as in

[6] We assume interaction parts to be contiguous, e.g. excluding the case of β-bulges.

(1) and the conditional probability $\mathcal{P}(\mathcal{I}|m,e,T)$ can be further factored as

$$\mathcal{P}(\mathcal{I}|m,e,T) = \mathcal{P}(r|k)\mathcal{P}(\{S_j \leftrightarrow S_{j'}\}_{j=1}^{r}) \prod_{j=1}^{r} \mathcal{P}(d_{jj'}|S_j \leftrightarrow S_{j'})\mathcal{P}(a_{jj'}|S_j \leftrightarrow S_{j'}, d_{jj'}) \qquad (11)$$

where r is the number of interacting pairs, k is the number of β-strands, and $\{S_j \leftrightarrow S_{j'}\}_{j=1}^{r}$ denotes a combination for β-strands to form r interacting pairs. Various specifications for these distributions in (11) are applicable provided that they satisfy $\sum_{\mathcal{I}} \mathcal{P}(\mathcal{I}|m,e,T) = 1$. In the present work, we assumed a uniform distribution, $\mathcal{P}(\{S_j \leftrightarrow S_{j'}\}_{j=1}^{r}) = \frac{1}{c(r,k)}$ if the combination is valid, where $c(r,k)$ is the total number of *valid combinations*, otherwise $\mathcal{P}(\{S_j \leftrightarrow S_{j'}\}_{j=1}^{r}) = 0$. A *valid combination* requires that each β-strand interacts with at least one and at most two other strands. This constraint comes from the chemical structure of amino acids, i.e. the CO and NH groups. $\mathcal{P}(r|k)$, $\mathcal{P}(d_{jj'}|S_j \leftrightarrow S_{j'})$ and $\mathcal{P}(a_{jj'}|S_j \leftrightarrow S_{j'}, d_{jj'})$ are discrete distributions depending on the distance between the two β-strands and their lengths, which were learned from training data by counting the relative occurrence frequencies.

B. Joint Segmental Likelihood

It is straightforward to extend the parametric model (8) to include long range interactions in β-sheets, which can be regarded as an extension of the dependency window to include the distal pairing partners. We introduce another 20×20 weight matrix W_{sheet} to capture the correlation between distal interacting pairs. The segmental likelihood function (3) for the β-strands can be enhanced as

$$\mathcal{P}(S_i|S_{-i}, T_i = E, \mathcal{I}) = \prod_{k=e_{i-1}+1}^{e_i} \frac{\Gamma(\sum_a \tilde{\gamma}_k^a) \prod_a \Gamma(\tilde{\gamma}_k^a + M_k^a)}{\Gamma\left(\sum_a (\tilde{\gamma}_k^a + M_k^a)\right) \prod_a \Gamma(\tilde{\gamma}_k^a)} \frac{\Gamma(\sum_a M_k^a + 1)}{\prod_a \Gamma(M_k^a + 1)} \mathcal{P}_c(R_k|T_i = E) \qquad (12)$$

with $\tilde{\gamma}_k = \gamma_k + \sum_{\{k^*\}} W_{sheet} M_{k^*}$ where γ_k is defined as in (7) and $\{k^*\}$ denotes the set of interacting residues of O_k that can be determined by $\mathcal{I}$.

C. β-Sheet Contact Maps

Contact maps represent the pairwise, inter-residue contacts, as a symmetrical, square, boolean matrix. Pollastri and Baldi [28] have previously applied ensembles of bidirectional recurrent neural network architectures to the prediction of such contact maps. In this section, we describe the capability of this parametric SSMM model to carry out inference on contact maps, which has also been explored by Schmidler et al. [14] [11]. This capability is one of the advantages of the probabilistic modelling approach over the traditional discriminative approach (e.g. neural networks) to protein secondary structure prediction. β-sheets are built up from pairs of β-strands with hydrogen bonds, which are prominent features in contact maps. The set of β-sheet interactions is associated with a *β-sheet contact map* defined by a $n \times n$ matrix $\mathcal{C}$ whose ij-th entry $\mathcal{C}^{ij}$ defined as

$$\mathcal{C}^{ij}(\mathcal{I}) = \begin{cases} 1 & \text{if } O_i \text{ and } O_j \text{ are paired in } \mathcal{I}; \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

We may estimate the marginal predicted $\mathcal{C}$ from the posterior distribution of $\mathcal{P}(m, e, T, \mathcal{I}|O)$, given by

$$\mathcal{P}(\mathcal{C}^{ij} = 1|O) = \sum_{m,e,T,\mathcal{I}} \mathcal{C}^{ij}(\mathcal{I})\, \mathcal{P}(m, e, T, \mathcal{I}|O) \tag{14}$$

where the indicator function $\mathcal{C}^{ij}(\mathcal{I})$ is defined as in (13). Using the samples we have collected from the distributions $\mathcal{P}(m, e, T|O)$ and $\mathcal{P}(\mathcal{I}|m, e, T)$ (see Section V and Appendix A for details), (14) can be estimated by

$$\mathcal{P}(\mathcal{C}^{ij} = 1|O) = \sum_{m,e,T} \sum_{\mathcal{I}} \mathcal{C}^{ij}(\mathcal{I})\, \mathcal{P}(m, e, T, \mathcal{I}|O) \approx \frac{1}{\mathcal{N}} \sum_{\{m,e,T\}} \sum_{\{\mathcal{I}\}} \mathcal{C}^{ij}(\mathcal{I})\, \frac{\mathcal{P}(O|m, e, T, \mathcal{I})}{\sum_{\{\mathcal{I}\}} \mathcal{P}(O|m, e, T, \mathcal{I})} \tag{15}$$

where the samples $\{\mathcal{I}\}$ are collected from $\mathcal{P}(\mathcal{I}|m, e, T)$, and $\mathcal{N}$ samples of $\{m, e, T\}$ are from $\mathcal{P}(m, e, T|O)$.

IV. PARAMETER ESTIMATION

The probabilistic model we describe above has five classes of latent variables, parameters and hyperparameters, which are inferred or specified in different ways:

- Latent variables related to the location and length of secondary structure elements $\{m, e, T\}$
  - number of segments: m
  - the end points of each segment that specify the segment lengths: e
  - secondary structure classes of each segment: T

We infer these latent variables by sampling from the posterior distribution $\mathcal{P}(m, e, T|O)$ (see Section V for details).

- Latent variables related to distal interactions in β-sheets $\{\mathcal{I}\}$
  - number of interacting pairs: r
  - the interacting pairs of β-strands: $\{S_j \leftrightarrow S_{j'}\}$
  - orientation indicators: $\{d_{jj'}\}$
  - the indicators of alignment positions: $\{a_{jj'}\}$

  These interacting variables can be sampled in the conditional distribution $\mathcal{P}(\mathcal{I}|m, e, T)$ (see Section V and Appendix A for details).
- Parameters that specify discrete distributions
  - State transition probabilities for $\mathcal{P}(T_i|T_{i-1})$ as defined in (1)[7]
  - Segmental length distributions $\mathcal{P}(e_i|e_{i-1}, T_i)$ as defined in (1)
  - Position-specific distributions of amino acids $\mathcal{P}_c(R|T_i)$ as defined in (9) for capping signals
  - The conditional distribution of the number of interacting pairs $\mathcal{P}(r|k)$ as defined in (11)
  - The conditional distribution of the orientation indicators $\mathcal{P}(d_{jj'}|S_j, S_{j'})$ as defined in (11)[8]
  - The conditional distribution of the alignment position $\mathcal{P}(a_{jj'}|S_j, S_{j'}, d_{jj'})$as defined in (11)

  These parameters specifying discrete distributions can be directly estimated by their relative frequency of occurrence in the training data set.[9] We present the results of state transition probabilities and segmental length distributions, estimated from our training data, in Figure 4 and Figure 5 respectively as an illustration. $\mathcal{P}(\{S_j, S_{j'}\}_{j=1}^{r})$ defined as in (11) is uniformly distributed.

[7]The initial state probabilities $\mathcal{P}(T_0)$ can simply be set to be equal.

[8]The distribution is actually only conditional on the distance between the β-strand pair and the segment lengths of the two β-strands.

[9]An appropriate prior might be used for smoothing.

- Weight parameters in the likelihood function for segmental dependency (8) and (12) were estimated by penalized maximum likelihood, where the regularization factor can be estimated from the data (see Section IV-A below for details).
- Model parameters
  - N-capping and C-capping length for capping signals.

    The capping components result in amino acid distributions at the end-segment positions which differ significantly from the overall distribution. In Table I, we present the Kullback-Leibler divergence between the overall amino acid distribution and the distribution at capping positions. We observed that N-capping and C-capping with length 4 can include most of the significant divergence, and so we fix the N-capping and C-capping length both to be 4.
  - The length of dependency window in (7).

    Crooks and Brenner [29] have examined the entropy densities of protein primary and secondary structure sequences, and the local inter-sequence mutual information density. They found that the inter-sequence interactions important to secondary structure prediction are short-ranged, usually within 5 neighbouring residues and so we decide to fix the window length at 5 in the present work.

A. *Estimates on Weight Parameters*

The weight parameters consist of three sets for different segmental types, i.e. $\{\boldsymbol{W}_\tau\}$ for $\tau \in \mathcal{T}$. For each segment type τ, there are $|\mathcal{A}|^2\ell$ parameters of W_{intra}'s, $|\mathcal{A}|^2\ell$ parameters of W_{inter}'s and $|\mathcal{A}|$ parameters in the vector W^0, where the types of amino acid residues $|\mathcal{A}| = 20$ and the length of dependency window $\ell = 5$ in the present work. Thus the total number of weight parameters is 4020. β-strands have an additional $|\mathcal{A}|^2$ parameters in W_{sheet} if the long-range interactions are incorporated.

The maximum a posteriori (MAP) estimate of its associated weights $\boldsymbol{W}_\tau$ can be obtained as

$$\arg\max_{\boldsymbol{W}_\tau} \mathcal{P}(\{O, m, e, T\}|\boldsymbol{W}_\tau)\mathcal{P}(\boldsymbol{W}_\tau) \tag{16}$$

under the condition of positive elements, where $\mathcal{P}(\boldsymbol{W}_\tau)$ is the prior probability usually specified by $\mathcal{P}(\boldsymbol{W}_\tau) \propto \exp(-\frac{C}{2}\|\boldsymbol{W}_\tau\|_2^2)$ with $C \geq 0$, and $\mathcal{P}(\{O, m, e, T\}|\boldsymbol{W}_\tau)$ is the product of the

TABLE I

KULLBACK-LEIBLER DIVERGENCE BETWEEN THE OVERALL AMINO ACID DISTRIBUTION AND THE DISTRIBUTION AT CAPPING POSITIONS. THE DIVERGENCE BETWEEN TWO DISTRIBUTIONS $\mathcal{Q}$ AND $\mathcal{P}$ IS EVALUATED BY $\sum_R \mathcal{Q}(R) \log\left(\frac{\mathcal{Q}(R)}{\mathcal{P}(R)}\right)$. HERE, $R \in \mathcal{A}$, $\mathcal{P}(R)$ IS THE AMINO ACID DISTRIBUTION AT CAPPING POSITIONS, AND $\mathcal{Q}(R)$ IS THE OVERALL SEGMENTAL DISTRIBUTION. BOLD FACE WAS USED TO INDICATE DIFFERENCES ABOVE THE CUTOFF 0.01.

Capping Position	α-helix	β-strand	Coil
N1	**0.0290**	0.0059	0.0064
N2	**0.0535**	0.0093	0.0043
N3	**0.0136**	0.0069	0.0008
N4	**0.0478**	0.0038	0.0018
N5	0.0030	0.0076	0.0032
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C5	0.0091	0.0050	0.0037
C4	**0.0181**	0.0018	0.0025
C3	0.0086	0.0089	0.0019
C2	0.0044	0.0028	0.0015
C1	0.0066	**0.0155**	0.0046

joint probabilities over all protein chains in training data set. The optimal $\boldsymbol{W}_\tau$ is therefore the minimizer of the negative logarithm of (16), which can be obtained by

$$\min_{\boldsymbol{W}_\tau} \mathcal{L}(\boldsymbol{W}_\tau) = -\sum_{\{O\}}\sum_{\{\tau\}} \ln \mathcal{P}(S_i|S_{-i},\tau) + \frac{C}{2}\|\boldsymbol{W}_\tau\|_2^2 \tag{17}$$

subject to $w > 0, \forall w \in \boldsymbol{W}_\tau$, where $\sum_{\{O\}}$ means the sum over all the protein chains, $\sum_{\{\tau\}}$ denotes the sum over all the segments of type τ, and $\mathcal{P}(S_i|S_{-i},\tau)$ is defined as in (3). A set of auxiliary variables $\mu = \ln w$ can be introduced to convert the constrained optimization problem into an unconstrained problem. The derivatives of $\mathcal{L}(\boldsymbol{W}_\tau)$ with respect to μ are given as follows:

$$\frac{\partial \mathcal{L}(\boldsymbol{W}_\tau)}{\partial \mu} = w\Big(\sum_{\{O\}}\sum_{\{\tau\}}\sum_{k=e_{i-1}+1}^{e_i} \psi_k^T \frac{\partial \boldsymbol{\gamma}_k}{\partial w} + Cw\Big) \tag{18}$$

where $\boldsymbol{\gamma}_k$ is defined as in (7), and $\psi_k = \frac{\partial - \ln \mathcal{P}(M_k|M_{[1:k-1]},\tau)}{\partial \boldsymbol{\gamma}_k}$ is a 20×1 vector whose a-th element is

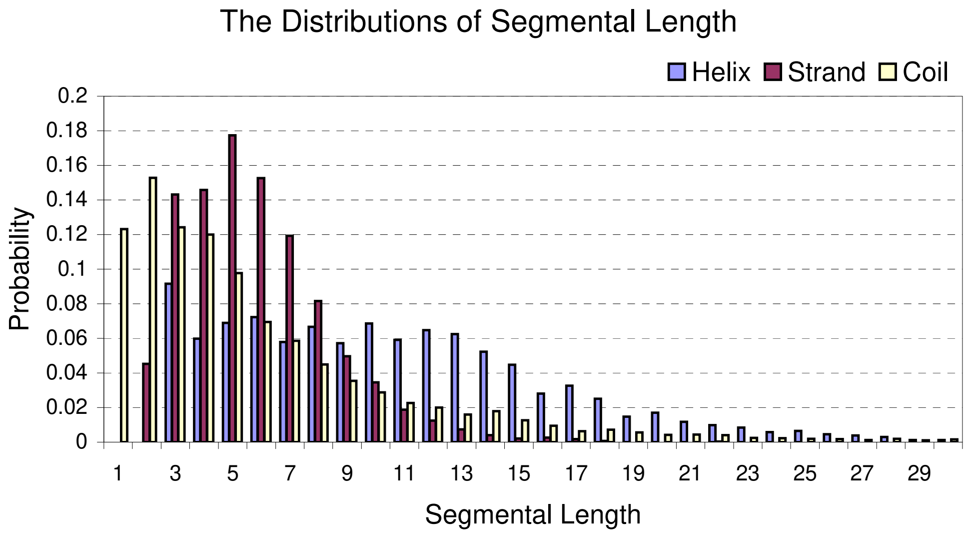


Fig. 4. The distributions of segmental length for the three segment types, $\mathcal{P}(e_i|e_{i-1}, T_i)$ defined as in (1). Note that the three distributions are quite different as pointed by Schmidler et al. [13].

$$\psi_k^a = \Psi(\gamma_k^a) - \Psi(\gamma_k^a + M_k^a) + \Psi(\textstyle\sum_a(\gamma_k^a + M_k^a)) - \Psi(\textstyle\sum_a(\gamma_k^a))$$

where $\Psi(x) = \frac{d}{dx}\ln(\Gamma(x))$ is known as the digamma function. Then standard gradient-based optimization methods are employed to minimize (17).

The optimal value of the regularization factor C in the regularized functional $\mathcal{L}(\boldsymbol{W}_\tau)$ was determined by standard k-fold cross validation [30] [31]. We carried out 7-fold cross validation as follows. The original training data were randomly partitioned into 7 almost equal folds with each fold having an almost equal percentage of different segments and amino acid residues. Given a particular value of C, one fold was left out as a validation set in turn, the weight parameters were estimated by minimizing $\mathcal{L}(\boldsymbol{W}_\tau)$ $\forall\tau$ over the protein chains in the other 6 folds, and the resulting model was tested on the left-out fold to obtain the validation error. The average of the validation errors on the 7 left-out folds indicates the predictive performance of the regularization factor C. We tried the set of C values: $C = \{10^{-3}, 10^{-2}, ..., 10^{+2}\}$, and found the best validation performance was achieved when $C = 0.01$. The optimal weight parameters in the model were finally obtained by optimizing on the whole training data set with the best C value.

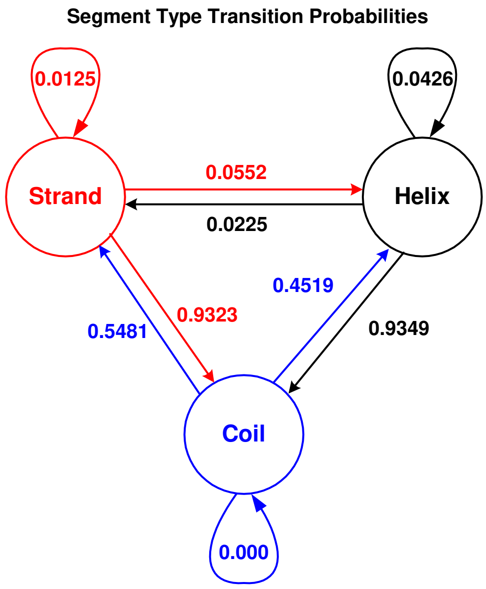


Fig. 5. The segment type transition probabilities, $\mathcal{P}(T_i|T_{i-1})$ defined as in (1). The self-transitions are obtained from the annotations in the training database.

It is possible to specify different values of C for the segment types, but it increases the computational cost of cross validation massively. Approximate Bayesian techniques could also be used to further specify different C values on weight matrices individually, while the computational difficulty lies in evaluating the integral over the high-dimensional weight space. This is an interesting and worthwhile issue for further investigation.

V. Sampling Scheme for Prediction

Without the incorporation of long range interactions, the quantities of the segmentation variables can be inferred exactly by the Viterbi and forward-backward algorithms in the segmental semi-Markov framework (see [27] or [13] for details). Generally, the introduction of long range interactions into the segmental model makes exact calculation of posterior probabilities intractable. Markov Chain Monte Carlo (MCMC) algorithms can be applied here to obtain approximate inference. Schmidler [11] attempted to collect samples from the joint posterior distribution $\mathcal{P}(m, e, T, \mathcal{I}|O)$, although the dependency between (m, e, T) and $\mathcal{I}$ makes it complicated to maintain detailed balance for the Metropolis proposals. However, the main

Metropolis-Hastings scheme is still applicable here. The main advantage of our approach over Schmidler's method is that we collect samples of long-range contacts under the realistic constraint that one amino acid can have at most two hydrogen bonds and unpaired beta strands are not possible.

In our approach, the latent variables of segmentation $\{m, e, T\}$ are sampled from the posterior distribution $\mathcal{P}(m, e, T|O)$ with MCMC, keeping the weight parameters and the model parameters fixed. The posterior distribution $\mathcal{P}(m, e, T|O)$ is proportional to the joint distribution $\mathcal{P}(m, e, T, O)$. The joint distribution can be evaluated as

$$\mathcal{P}(m, e, T, O) = \mathcal{P}(m, e, T) \prod_{\{S_i : T_i \neq E\}} \mathcal{P}(S_i|S_{-i}, T_i) \sum_{\mathcal{I}} \mathcal{P}(\mathcal{I}|m, e, T) \prod_{\{S_i : T_i = E\}} \mathcal{P}(S_i|S_{-i}, T_i, \mathcal{I}) \tag{19}$$

where the individual terms (from left to right on the righthand side) are defined as in (1), (9), (11) and (12) respectively. Note that only the β-strands are in the interaction set $\mathcal{I}$. Following Schmidler et al. [11] [14], we define a set of Metropolis proposals for the construction of a Markov chain on the space of segmentations $\mathcal{V} = (m, e, T)$:

- *Segment split*: propose $\mathcal{V}^* = (m^*, e^*, T^*)$ with $m^* = m + 1$ by splitting segment S_k into two new segments (S_{k^*}, S_{k^*+1}) with $k \sim \text{Uniform}[1 : m]$, $e_{k^*} \sim \text{Uniform}[e_{k-1} + 1 : e_k - 1]$, $e_{k^*+1} = e_k$, $T_{k^*} \sim \text{Uniform}[H, E, L]$, and $T_{k^*+1} \sim \text{Uniform}[H, E, L]$.[10]
- *Segment merge*: propose $\mathcal{V}^* = (m^*, e^*, T^*)$ with $m^* = m - 1$ by merging the two segments S_k and S_{k+1} into one new segment S_{k^*} with $k \sim \text{Uniform}[1 : m - 1]$, $e_{k^*} = e_{k+1}$, and $T_{k^*} \sim \text{Uniform}[H, E, L]$.
- *Type change*: propose $\mathcal{V}^* = (m, e, T^*)$ with $T^* = [T_1, \ldots, T_{k-1}, T_k^*, T_{k+1}, \ldots, T_m]$ where $T_k^* \sim \text{Uniform}[H, E, L]$.
- *Endpoint change*: propose $\mathcal{V}^* = (m, e^*, T)$ with $e^* = [e_1, \ldots, e_{k-1}, e_k^*, e_{k+1}, \ldots, e_m]$ where $e_k^* \sim \text{Uniform}[e_{k-1} + 1 : e_{k+1} - 1]$.

The acceptance probability for *Type change* and *Endpoint change* depends on the ratio of likelihood $\frac{\mathcal{P}(\mathcal{V}^*, O)}{\mathcal{P}(\mathcal{V}, O)}$, where the likelihood is defined as in (19). *Segment split* and *Segment merge* jumps between segmentations of different dimension are accepted or rejected according to

[10]Here $\sim \text{Uniform}[H, E, L]$ denotes uniformly sampling in the set $\{H, E, L\}$.

a reversible-jump Metropolis criteria. According to the requirement of detailed balance, the acceptance probability for a new proposal $\mathcal{V}^*$ should be $\rho(\mathcal{V},\mathcal{V}^*) = \frac{\mathcal{P}(\mathcal{V}^*,O)}{\mathcal{P}(\mathcal{V},O)} \times \frac{\mathcal{P}(\mathcal{V}\leftarrow\mathcal{V}^*)}{\mathcal{P}(\mathcal{V}^*\leftarrow\mathcal{V})}$. Therefore, the acceptance probability for *Segment split* and *Segment merge* should respectively be

$$\begin{aligned}
\rho_{split(k)}(\mathcal{V},\mathcal{V}^*) &= \frac{\mathcal{P}(\mathcal{V}^*,O)}{\mathcal{P}(\mathcal{V},O)} \times |\mathcal{T}| \cdot (e_k - e_{k-1} - 1) \\
\rho_{merge(k)}(\mathcal{V},\mathcal{V}^*) &= \frac{\mathcal{P}(\mathcal{V}^*,O)}{\mathcal{P}(\mathcal{V},O)} \times \frac{1}{|\mathcal{T}| \cdot (e_{k+1} - e_{k-1} - 1)}
\end{aligned} \tag{20}$$

where $\mathcal{P}(\mathcal{V},O)$ is defined as in (19) and $|\mathcal{T}| = 3$ denotes the number of segment types.

Due to the factorizations in (19), only the changed segments require evaluation in computing the acceptance probability for the new proposal $\mathcal{V}^*$. Once the β-strands are changed in the new proposal, the interacting set $\mathcal{I}$ is changed too. The joint segmental likelihood of the β-strands has to be calculated again, which is a sum $\sum_{\mathcal{I}} \mathcal{P}(\mathcal{I}|m,e,T) \prod_{\{S_i:T_i=E\}} \mathcal{P}(S_i|S_{-i},T_i,\mathcal{I})$. Although the set $\mathcal{I}$ is composed of finite elements, it might be too expensive to enumerate all of them for the marginalization. We again apply sampling methods here to approximate the sum by randomly walking in the distribution $\mathcal{P}(\mathcal{I}|m,e,T)$ that is defined as in (11). A sampling scheme is described in Appendix A for this purpose. These samples can be reused to estimate the β-sheet contact map as in (15).

In the model training discussed in Section IV-A, we need to solve three optimization problems to estimate the weight parameters by gradient-descent methods. This is required tens of times in cross validation.[11] Once the optimal regularization parameter is found, we solve the minimization problems once more to get the final weight parameters. The cost on counting the occurrence frequencies of these discrete distributions is relatively negligible. With the incorporation of long-range interactions, we employed the sampling scheme described above to collect 10000 samples from the posterior distribution.[12] To approximate the marginalization over the interacting set, we randomly collected 40 samples in the β-sheet space $\mathcal{P}(\mathcal{I}|m,e,T)$.

Ideally, the inference problem should be formulated as a Bayesian hierarchical model and all quantities, which includes the latent variables, parameters and model parameters, could be sampled from the joint posterior distribution by MCMC methods [32]. However, the

[11] In the present work, it is required 6×7 times for 7-fold cross validation on 6 different C values.

[12] The first 1000 samples in the Markov chain were discarded for "burn in".

computational cost could be prohibitively expensive, since it would involve sampling in several high-dimensional spaces jointly. For example, the three spaces of the weight parameters contain over ten thousand variables. Laplace approximation could also be applied to carry out this integration, despite the demanding requirement of inversion of large matrices. Nevertheless, we believe that approximate Bayesian inference could achieve a genuine improvement, which is well worth further investigation.

VI. Results and Discussion

We implemented the proposed algorithm in ANSI C. In this implementation, the length of dependency window was fixed at 5, and the length of N- and C-capping was fixed at 4, and the regularization factor C was fixed at 0.01. We normalized the M_i vectors so that $\sum_a M_i^a = 1$ for both the multiple sequence alignment profile and the PSSM based profile. We used the following quantities as performance measures:

- Overall 3-state Accuracy Q_3,
- Sensitivity $Q^{obs} = \frac{TruePositive}{TruePositive+FalseNegative}$,
- Positive Predictive Value $Q^{pred} = \frac{TruePositive}{TruePositive+FalsePositive}$,
- Matthew's correlation $\mathcal{C} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FP)}}$ defined by Matthews [33],
- Segment Overlap Measure (SOV) as defined by Zemla et al. [34].

A. Validation on CB513

The data set we used is CB513, a non-redundant set of 513 non-homologous protein chains with structures determined to a resolution of ≤ 2.5Å generated by Cuff and Barton [35].[13] This data set has been used as a common benchmark for a number of different secondary structure prediction algorithms. We used 3-state DSSP definitions of secondary structure [36], calculated from the PDB files.[14] We removed the proteins that are shorter than 30 residues, or longer than

[13]The data set and the multiple sequence alignments profiles generated by Cuff and Barton [5], can be accessed at http://www.compbio.dundee.ac.uk/$\sim$www-jpred/data/.

[14]In DSSP definitions, H and G were assigned as α-helix segments, E and B were assigned as β-strands, and the others were assigned as coil. The segments with only one residue were also labelled as coil.

550 residues, following [5], to leave 480 proteins for 7-fold cross validation. Seven folds were created randomly, and validation outputs were carried out on the left-out fold while the weight parameters were optimized on the other 6 folds with a fixed regularization factor in turn. The optimal value of C was chosen as 0.01 after carrying out a 7-fold cross validation experiment on the dataset, as described in Section IV. In the following, we report the validation results using $C = 0.01$ in the 7-fold cross validation experiment. We used two kinds of alignment profiles: the multiple sequence alignment profiles (MSAP) used by Cuff and Barton [5], and position-specific score matrices (PSSM) as in [4]. For comparison purposes, we also implemented the algorithm proposed by Schmidler et al. [13], which uses the single sequence information only. The validation results are recorded in Table II. We also cite the results reported by Cuff and Barton [5] in Table III for reference. The results obtained from our model show a substantial improvement over those of Schmidler et al. [13] on all evaluation criteria. Compared with the performance of the neural network methods with various alignment profiles as shown in Table III, the prediction accuracy of our model is also competitive.[15] Due to small sample errors and the variation due to changes in secondary structure assignment by different methods, reported accuracies separated by less than about two percentage points are unlikely to be statistically significant [37] [38] and our results are comparable to many other prediction methods which have been tested on this benchmark data. Crooks and Brenner [29] point out that this is probably due to the fact that most contemporary methods for secondary structure prediction all utilize local sequence correlations, which contain only about one quarter of the total information necessary to determine secondary structure.

We did observe that the marginal posterior mode is more accurate than the MAP estimate, which shows that averaging over all the possible segmentations helps. According to the class definitions of the Structural Classification of Proteins database (SCOP) [39], we divided the 480 chains of CB513 into four groups: α, β, α/β and $\alpha + \beta$. The validation results of marginal posterior mode estimate on these groups are recorded separately in Table IV. We note that the

[15] It is also possible to further improve performance by constructing smoothers over current predictive outputs as Cuff and Barton [5] did in their Jury networks.

performance on α/β and α proteins is relatively better than that on $\alpha + \beta$ and β.

Note that validation results of CB513 data set, reported in Table II, III and IV, are different from predictive (blind) test results and could be slightly over-optimistic estimates of the generalization performance. We have included validation results so as to make a fair comparison with the results of Cuff and Barton [5] and Schmidler et al. [13], who also used this type of 7-fold cross validation procedure (see Table III in [5]). Cuff and Barton's results [5] are based on the choice of many model parameters (e.g. window size, number of hidden neurons, a "conservation number", etc.), although the procedure to determine these parameters is unclear from their paper. In order to have a truly unbiased estimate we have also carried out extensive predictive blind tests on the CASP data sets in the next subsection.

B. Blind Test on CASP Targets

The meetings of Critical Assessment of Techniques for Protein Structure Prediction (CASP) facilitate large-scale experiments to assess protein structure prediction methods. To perform a blind test experiment, we extracted protein chains from the latest three meetings from the public web page of the Protein Structure Prediction Center (accessible at http://predictioncenter.llnl.gov/). We used the model parameters specified in Section IV and fixed the regularization factor C to the optimal value 0.01 obtained from the cross validation experiments. Then the weight parameters of our model were optimized using all the 480 chains from CB513 and their PSSM profiles. We also prepared a larger training dataset using the CulledPDB list with the percentage identity cutoff 25%, the resolution cutoff 1.8 angstroms, and the R-factor cutoff 0.25.[16] There are 2147 chains in this expanded list. We used the same model settings, and optimized the weights parameters on the subset of 1814 chains.[17] The predictive results of the marginal posterior mode estimate of our two models on these CASP target proteins are reported in Table V, indexed by meeting, along with the marginal posterior mode estimate of Schmidler et al.'s algorithm [13]. The detailed predictive results of CASP5 can be found on our supplementary

[16] The protein list is accessible at http://dunbrack.fccc.edu/Guoli/pisces_download.php.

[17] The reduction was caused by removing the protein chains that are shorter than 30 residues, or longer than 550 residues, following [5].

TABLE II

VALIDATION RESULTS FOR SECONDARY STRUCTURE PREDICTION ON 480 PROTEIN SEQUENCES FROM CB513. "SEQUENCE ONLY" DENOTES THE ALGORITHM OF SCHMIDLER [13]; MSAP DENOTES OUR APPROACH USING MULTIPLE SEQUENCE ALIGNMENT PROFILES; PSSM DENOTES OUR APPROACH USING POSITION SPECIFIC SCORE MATRICES. Q_3 DENOTES THE OVERALL ACCURACY. $Q^{obs} = \frac{TruePositive}{TruePositive+FalseNegative}$ AND $Q^{pred} = \frac{TruePositive}{TruePositive+FalsePositive}$. $\mathcal{C}$ DENOTES MATTHEWS' CORRELATION COEFFICIENT DEFINED BY MATTHEWS [33]. SOV DENOTES THE SEGMENT OVERLAP MEASURE [34]. THE SUBSCRIPTS DENOTE THE SECONDARY STRUCTURE TYPE. MAP DENOTES THE MOST PROBABLE POSTERIOR ESTIMATE, WHILE MARG DENOTES MARGINAL POSTERIOR MODE ESTIMATE.

	Sequence Only		with MSAP		with PSSM	
	MAP	MARG	MAP	MARG	MAP	MARG
Q_3	59.23%	65.08%	68.11%	71.31%	62.54%	72.23%
Q_H^{obs}	66.34%	66.73%	78.17%	78.71%	66.74%	71.56%
Q_E^{obs}	20.74%	46.32%	41.40%	57.61%	26.18%	54.93%
Q_C^{obs}	72.80%	73.19%	73.28%	72.11%	77.45%	81.54%
Q_H^{pred}	61.87%	68.64%	69.91%	73.51%	66.82%	79.90%
Q_E^{pred}	56.45%	58.88%	70.15%	67.67%	72.16%	70.96%
Q_C^{pred}	57.77%	64.72%	66.06%	70.94%	58.54%	67.91%
$\mathcal{C}_H$	0.3709	0.4621	0.5457	0.5927	0.4350	0.6085
$\mathcal{C}_E$	0.2194	0.3809	0.4306	0.5120	0.3359	0.5203
$\mathcal{C}_C$	0.2821	0.3945	0.4253	0.4820	0.3338	0.5064
SOV_H	48.10%	61.18%	61.56%	67.83%	50.03%	68.58%
SOV_E	58.55%	64.41%	68.78%	74.26%	57.39%	70.71%
SOV_C	33.13%	60.90%	53.08%	69.42%	39.50%	68.92%
SOV	37.50%	60.97%	58.42%	63.79%	48.54%	68.31%

webpage http://public.kgi.edu/∼wild/bsm.html. We also cite the average performance of the participants from the CASP5 website for comparative purposes.

The results of these blind tests indicate that our algorithm based on generative modelling gives comparable results to other contemporary methods.[18] The performance of Q_3 and SOV on the target proteins of CASP5 are shown in Figure 6. We found that the model trained on the larger

[18] The predictive results produced by other contemporary methods, indexed by CASP meeting, are available at http://predictioncenter.llnl.gov/

TABLE III

THE RESULTS OF 7-FOLD CROSS VALIDATION ON 480 PROTEINS OF CB513 REPORTED BY [5], ALONG WITH OUR RESULTS. Q_3 DENOTES THE OVERALL ACCURACY.

METHOD DESCRIPTION	Q_3
NETWORKS USING FREQUENCY PROFILE FROM CLUSTALW	71.6%
NETWORKS USING BLOSUM62 PROFILE FROM CLUSTALW	70.8%
NETWORKS USING PSIBLAST ALIGNMENT PROFILES	72.1%
ARITHMETIC SUM BASED ON THE ABOVE THREE NETWORKS	73.4%
NETWORKS USING PSIBLAST PSSM	75.2%
OUR ALGORITHM WITH MSAP OF [5]	71.3%
OUR ALGORITHM WITH PSIBLAST PSSM	72.2%

dataset can achieve better generalization performance, especially on SOV.

C. Prediction of Contact Maps

In the inference with long range interactions, we approximated the marginalization over the β-sheet space by randomly collecting 40 samples in $\mathcal{P}(\mathcal{I}|m, e, T)$ as described in Appendix A. We present the trace plot of ten test proteins in the MCMC sampling to show the convergence of the Markov chains, and compare the results to those without long range interactions in Figure 7. We found that the Markov chains converge well after 6000 samples in all cases.

We prepared a dataset with long range interaction information specified by the Protein Data Bank (PDB) files. The dataset, a subset of CB513, is composed of 198 protein chains along with β-sheet definitions.[19] This reduction in size was caused by the incompleteness in the long range interaction information in many of the original PDB files. In MCMC sampling we collected 9000 samples. 30-fold cross validation was carried out on this subset. Surprisingly, we have not yet observed significant improvement on secondary structure prediction accuracy in the sampling results over exact inference without long range interactions. A potential reason might be the small size of the training data set used in this set of experiments, which we will investigate

[19]The list of these proteins can be found at http://public.kgi.edu/∼wild/bsm.html.

TABLE IV

VALIDATION RESULTS OF MARGINAL POSTERIOR MODE ESTIMATE FOR SECONDARY STRUCTURE PREDICTION ON 480 PROTEIN SEQUENCES FROM CB513, CATEGORIZED BY STRUCTURAL CLASSES OF PROTEINS (SCOP). MSAP DENOTES OUR APPROACH USING MULTIPLE SEQUENCE ALIGNMENT PROFILES; PSSM DENOTES OUR APPROACH USING POSITION SPECIFIC SCORE MATRICES. Q_3 DENOTES THE OVERALL ACCURACY. $Q^{obs} = \frac{TruePositive}{TruePositive+FalseNegative}$ AND $Q^{pred} = \frac{TruePositive}{TruePositive+FalsePositive}$. $\mathcal{C}$ DENOTES MATTHEWS' CORRELATION COEFFICIENT DEFINED BY MATTHEWS [33]. SOV DENOTES THE SEGMENT OVERLAP MEASURE [34]. THE SUBSCRIPTS DENOTE THE SECONDARY STRUCTURES.

	α proteins (65 chains)		β proteins (92 chains)		α/β proteins (129 chains)		$\alpha+\beta$ proteins (87 chains)	
	MSAP	PSSM	MSAP	PSSM	MSAP	PSSM	MSAP	PSSM
Q_3	73.97%	74.78%	68.35%	69.95%	72.97%	74.47%	69.81%	70.72%
Q_H^{obs}	79.53%	76.02%	64.45%	55.82%	81.02%	73.36%	76.91%	69.72%
Q_E^{obs}	47.45%	45.60%	57.06%	52.71%	61.90%	63.44%	55.92%	52.39%
Q_C^{obs}	67.50%	76.48%	78.00%	86.82%	69.48%	79.94%	72.66%	82.44%
Q_H^{pred}	85.36%	89.13%	46.01%	56.58%	75.72%	82.14%	71.19%	77.49%
Q_E^{pred}	25.00%	25.45%	77.43%	81.60%	65.67%	69.97%	68.85%	74.84%
Q_C^{pred}	67.35%	66.00%	71.22%	68.17%	72.82%	70.00%	69.17%	65.63%
$\mathcal{C}_H$	0.5328	0.5694	0.4424	0.4792	0.5945	0.6131	0.5770	0.5910
$\mathcal{C}_E$	0.2917	0.2894	0.4842	0.4965	0.5517	0.5894	0.4880	0.5100
$\mathcal{C}_C$	0.4956	0.5339	0.4416	0.4654	0.4985	0.5340	0.4652	0.4858
SOV_H	67.04%	71.23%	65.44%	65.74%	71.77%	72.94%	66.84%	69.02%
SOV_E	74.36%	75.91%	60.25%	52.11%	82.95%	77.97%	77.04%	75.54%
SOV_C	75.98%	76.85%	66.54%	65.57%	73.32%	76.03%	65.17%	62.88%
SOC	59.98%	67.62%	66.82%	69.03%	64.96%	70.71%	62.55%	67.69%

further by training the model on a larger data set. However, this observation is consistent with the findings of Cline et al. [40] and Crooks et al. [41], who examined the mutual information content of interacting amino acid residues distantly separated by sequence but proximate in three-dimensional structure, and concluded that, for the purposes of tertiary structure prediction, these interactions were essentially uninformative. The analysis of Cline et al. [40] and Crooks et al. [41] also suggests that a modification to our method, which captures distal interactions between *secondary structure elements* rather than amino acid residues, should provide a distinct improvement.

TABLE V

PREDICTIVE RESULTS OF MARGINAL POSTERIOR MODE ESTIMATE OF OUR ALGORITHM USING PSSM ON THE PROTEIN DATA OF CASP. CASP 3 HAS 36 CHAINS, CASP 4 HAS 40 CHAINS, AND CASP 5 HAS 56 CHAINS. "SEQUENCE ONLY" DENOTES THE ALGORITHM OF SCHMIDLER [13]; "CB513 WITH PSSM" DENOTES OUR MODEL TRAINED ON THE 480 CHAINS FROM CB513 WITH PSSM PROFILES; "CULLEDPDB WITH PSSM" DENOTES OUR MODEL TRAINED ON THE 1814 CHAINS OF CULLEDPDB DATA. Q_3 DENOTES THE OVERALL ACCURACY. $Q^{obs} = \frac{TruePositive}{TruePositive+FalseNegative}$ AND $Q^{pred} = \frac{TruePositive}{TruePositive+FalsePositive}$. $\mathcal{C}$ DENOTES MATTHEWS' CORRELATION COEFFICIENT DEFINED BY MATTHEWS [33]. SOV DENOTES THE SEGMENT OVERLAP MEASURE [34].

	SEQUENCE ONLY			CB513 WITH PSSM			CULLEDPDB WITH PSSM		
	CASP3	CASP4	CASP5	CASP3	CASP4	CASP5	CASP3	CASP4	CASP5
Q_3	63.92%	66.29%	67.13%	72.03%	74.87%	74.49%	72.28%	74.68%	74.86%
Q_H^{obs}	61.55%	67.64%	67.47%	78.61%	85.17%	85.85%	71.13%	77.42%	77.53%
Q_E^{obs}	42.45%	44.92%	48.15%	58.30%	58.74%	60.74%	56.19%	58.42%	60.46%
Q_C^{obs}	78.57%	77.13%	77.46%	75.63%	73.69%	72.40%	82.78%	81.13%	80.54%
Q_H^{pred}	64.27%	72.49%	71.25%	71.17%	77.88%	73.73%	78.30%	82.82%	79.74%
Q_E^{pred}	69.81%	65.06%	67.96%	80.12%	77.07%	79.41%	78.67%	76.54%	78.18%
Q_C^{pred}	62.03%	61.79%	63.97%	69.29%	70.58%	73.06%	66.73%	67.35%	69.92%
$\mathcal{C}_H$	0.4200	0.4721	0.4904	0.5961	0.6573	0.6442	0.6111	0.6509	0.6451
$\mathcal{C}_E$	0.4072	0.4157	0.4478	0.5766	0.5804	0.6022	0.5589	0.5784	0.5966
$\mathcal{C}_C$	0.3881	0.4344	0.4378	0.4956	0.5362	0.5330	0.5050	0.5445	0.5472
SOV_H	62.58%	59.63%	65.60%	70.31%	69.24%	73.35%	70.52%	70.49%	74.31%
SOV_E	56.08%	60.39%	65.61%	67.55%	73.55%	77.23%	66.81%	70.69%	75.61%
SOV_C	61.55%	60.92%	60.87%	66.13%	67.73%	71.51%	71.90%	74.52%	73.73%
SOC	65.26%	65.39%	66.55%	64.91%	64.85%	68.18%	74.52%	68.95%	73.08%

However, it is interesting that we can infer β-sheet contacts based on the predicted secondary structure. We present predicted contact maps in Figure 8 as an example, where the the grey scale indicates the probability $\mathcal{P}(\mathcal{C}^{ij} = 1|O)$. It can be seen that, in the case of 1PGA (Protein G), which contains 2 parallel and 2 anti-parallel β-strands, and 1DTX (α-dendrotoxin, see Figure 9a), which contains 2 anti-parallel β-strands, the position and direction of the β-strands are predicted correctly, but have a shorter range than in the true contact maps. The false positive predictions in the case of 1DTX (α-dendrotoxin, see Figure 9b) are due to errors in the prediction of which

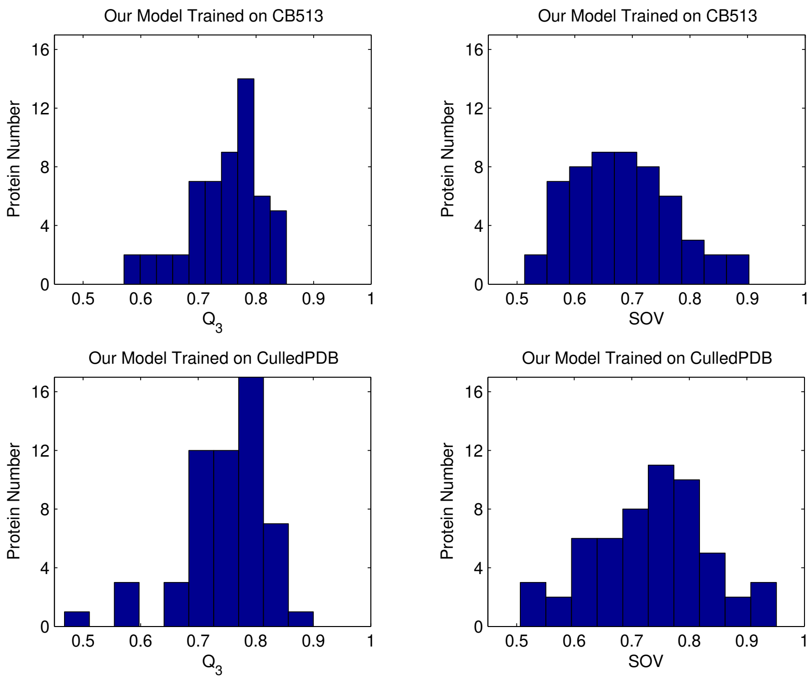


Fig. 6. The histogram of Q_3 and SOV to visualize the marginal posterior mode estimate of our models on CASP5 data. One model was trained on CB513 dataset and another was trained on CulledPDB dataset. The vertical axes are indexed by the number of proteins falling in the bins.

residues are in the β-strands.

To assess the overall prediction accuracy, we have also computed the area under the ROC curve (AUC) [42] for β-sheet contact prediction. The average AUC over these protein chains is 0.90 ± 0.10. The average ROC curves categorized by SCOP classes are presented in Figure 10. The averaged AUC of 44 β proteins is 0.87 ± 0.07, the averaged AUC of 64 α/β proteins is 0.93 ± 0.06 and the averaged AUC of 37 $\alpha + \beta$ proteins is 0.90 ± 0.10. Based on these ROC curves, we find that this algorithm performs better on the α/β class.

VII. Conclusion

In this paper, we have described a novel parametric Bayesian segmental semi-Markov model for proteins which incorporates the information in multiple sequence alignment profiles. Long range interaction information in β-sheets can be directly incorporated. The numerical results show that the generalization performance of this generative model is similar to other contemporary

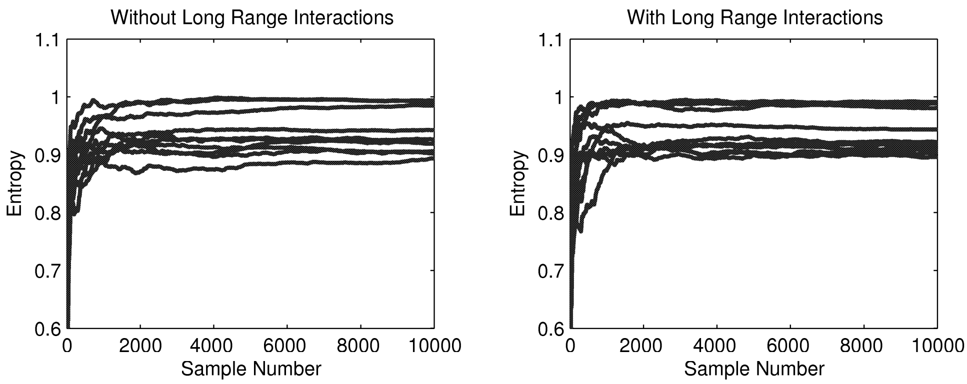


Fig. 7. The trace plot of ten protein chains in MCMC sampling. "Entropy" denoted the average entropy of the posterior distribution on the chain, i.e. $-\frac{1}{n}\sum_{i=1}^{n}\mathcal{P}(T_{O_i}|O)\log\mathcal{P}(T_{O_i}|O)$ where $\mathcal{P}(T_{O_i}|O)$ is the predictive probability of the segment type on the i-th amino acid of the protein chain. The sampling results of the cases with/without long range interaction are presented respectively.

methods. However, contact map prediction can also be carried out in the Bayesian segmental framework, which represents a considerable advantage over discriminative methods. Moreover, with the inclusion of potential functions with dihedral angle information in the joint sequence-structure probability distribution, this probabilistic model also has the potential for tertiary structure prediction, and this is the focus of our current work.

Appendix

A. Sampling in β-sheet Space

Given a specific segmentation, i.e. a set of $\{m, S, T\}$, there is a corresponding β-sheet space defined by a set of interaction variables $\mathcal{I}$ that specifies the interactions within these β-strands. The total number of β-strands is known, denoted as k. The distribution of the β-sheet space, $\mathcal{P}(\mathcal{I}|m, e, T)$, is defined as in (11). There are four steps to collect a sample in $\mathcal{P}(\mathcal{I}|m, e, T)$:

1) generate a sample of r in $\mathcal{P}(r|k)$.
2) collect a valid combination of r pairs from the k β-strands. The valid combination requires that each β-strand should be used at least once and at most twice.
3) for each pair $\{S_j, S_{j'}\}$, generate the alignment direction by $\mathcal{P}(d_{jj'}|S_j, S_{j'})$.
4) for each pair $\{S_j, S_{j'}\}$, generate the alignment position by $\mathcal{P}(a_{jj'}|S_j, S_{j'})$.

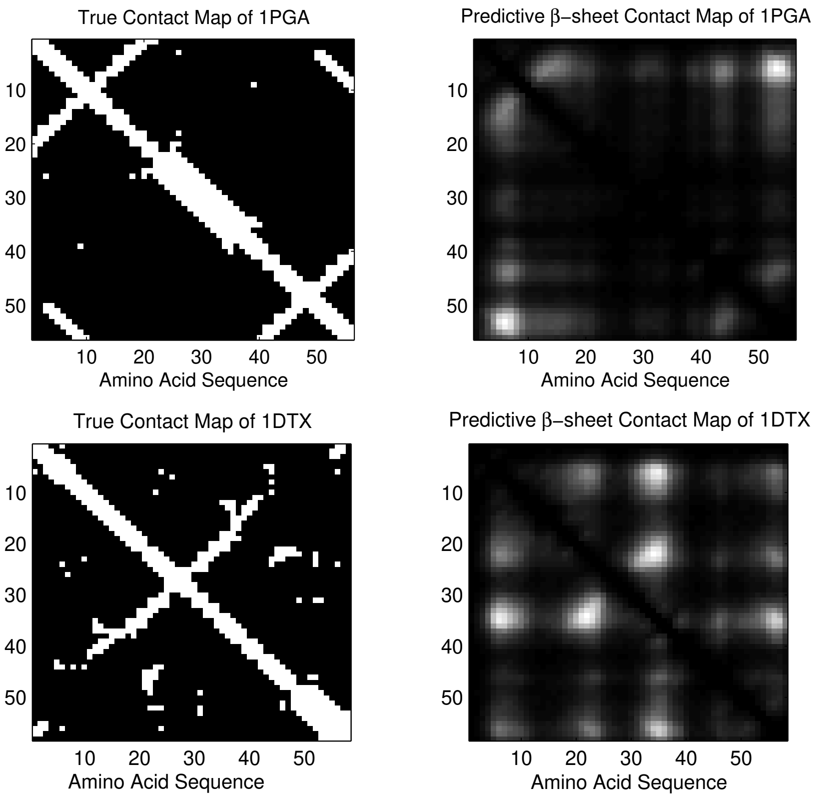


Fig. 8. True β-sheet contact maps and predicted maps for protein chains 1PGA and 1DTX. The true contact maps were produced with a threshold of 7 Å. The grey scale indicates the probability $\mathcal{P}(\mathcal{C}^{ij} = 1|O)$.

References

[1] N. Qian and T. J. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models," *Journal of Molecular Biology*, vol. 202, pp. 865–884, 1988.

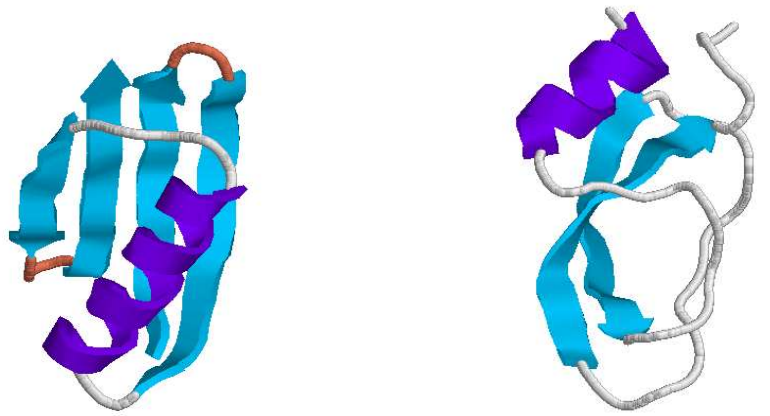
[2] B. Rost and C. Sander, "Prediction of protein secondary structure at better than 70% accuracy," *Journal of Molecular Biology*, vol. 232, pp. 584–599, 1993.

[3] S. F. Altschul, T. L. Madden, A. A. Schaeffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, pp. 3389–3402, 1997.

[4] D. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *Journal of Molecular Biology*, vol. 292, pp. 195–202, 1999.

[5] J. A. Cuff and G. J. Barton, "Application of multiple sequence alignment profiles to improve protein secondary structure prediction," *Proteins: Structure, Function and Genetics*, vol. 40, pp. 502–511, 2000.

[6] A. L. Delcher, S. Kasif, H. R. Goldberg, and W. H. Hsu, "Protein secondary structure modelling with probabilistic networks," in *Proc. of Int. Conf. on Intelligent Systems and Molecular Biology*, 1993, pp. 109–117.



(a) 1PGA (b) 1DTX

Fig. 9. Structural cartoons of protein 1PGA (Protein G) and 1DTX (α-dendrotoxin).

[7] C. Burge and S. Karlin, "Prediction of complete gene structures in human genomic DNA," *Journal of Molecular Biology*, vol. 268, no. 1, pp. 78–94, 1997.

[8] R. F. Yel, L. P. Lim, and C. B. Burge, "Computational inference of homologous gene structures in the human genome," *Genome Res.*, vol. 11, no. 5, pp. 803–816, 2001.

[9] L. Zhang, V. Pavlovic, C. R. Cantor, and S. Kasif, "Human-mouse gene identification by comparative evidence integration and evolutionary analysis," *Genome Res.*, vol. 13, pp. 1190–1202, 2003.

[10] I. Korf, P. Flicek, D. Duan, and M. R. Brent, "Integrating genomic homology into gene structure prediction," *Bioinformatics*, vol. 17 Suppl 1, pp. S140–S148, 2001.

[11] C. S. Schmidler, "Statistical models and monte carlo methods for protein structure prediction," Ph.D. Thesis, Stanford University, May 2002.

[12] M. Ostendorf, V. Digalakis, and O. Kimball, "From HMM to segment models: a unified view of stochastic modelling for speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 5, pp. 360–378, 1996.

[13] C. S. Schmidler, J. S. Liu, and D. L. Brutlag, "Bayesian segmentation of protein secondary structure," *Journal of Computational Biology*, vol. 7, no. 1/2, pp. 233–248, 2000.

[14] ——, *Case Studies in Bayesian Statistics*, ser. Lecture Notes in Statistics. Springer, 2002, vol. 5, ch. Bayesian Protein Structure Prediction, pp. 363–378.

[15] G. E. Hinton, "Products of experts," in *Proceedings of the Ninth International Conference on Artificial Neural Networks*, 1999, pp. 1–6.

[16] K. Simons, C. Kooperberg, E. Huang, and D. Baker, "Assembly of protein tertiary structures from fragments with similar local sequences using simulated anealing and bayesian scoring functions," *Journal of Molecular Biology*, vol. 268, pp. 209–225, 1997.

[17] Y. Ye, L. Jaroszewski, W. Li, and A. Godzik, "A segment alignment approach to protein comparison," *Bioinformatics*, vol. 19, pp. 742–749, 2003.

[18] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensistivity of progressive multiple

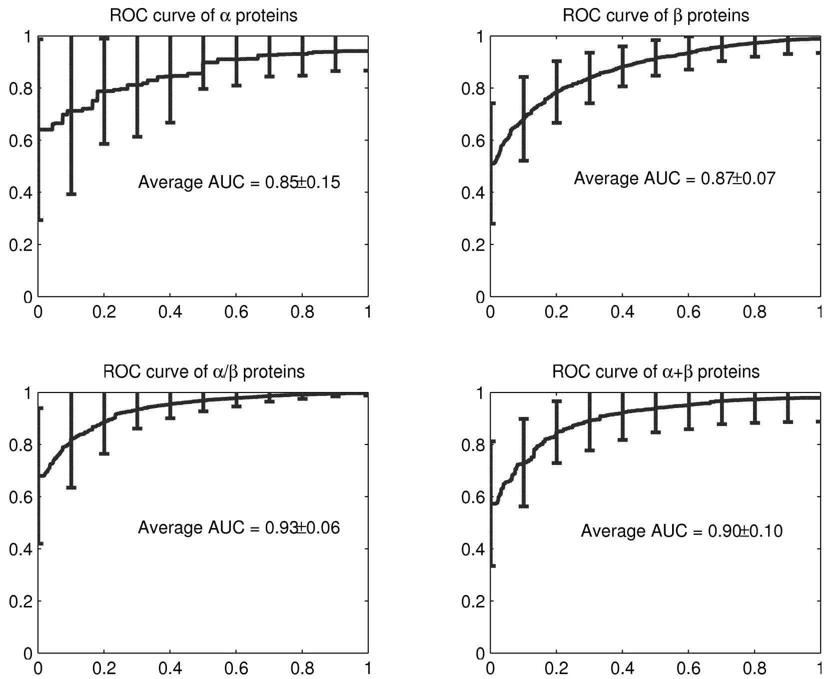


Fig. 10. The ROC curves of the proteins categorized by SCOP. The vertical lines bounded by bars in these graphs indicate the standard deviation at those positions. For these four graphs, the horizontal axes are indexed by $1.0 - specificity$ evaluated by $\frac{\text{Number of False Positive}}{\text{Number of Negative Samples}}$, and the vertical axes are of $sensitivity$ evaluated by $\frac{\text{Number of True Positive}}{\text{Number of Positive Samples}}$. For each of the structural classes, the average value and its standard deviation of AUC (the area under the ROC curve) are given by text in the corresponding graph.

sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Res.*, vol. 22, pp. 4673–4680, 1994.

[19] R. Aurora and G. D. Rose, "Helix capping," *Protein Science*, vol. 7, pp. 21–38, 1998.

[20] D. Eisenberg, R. M. Weiss, and T. C. Terwilliger, "The hydrophobic moment detects periodicity in protein hydrophobicity," *Proceedings of the National Academy of Sciences, USA*, vol. 81, pp. 140–144, 1984.

[21] Z. Aydin, Y. Altunbaşak, and M. Borodovsky, "Protein secondary structure prediction with semi markov hmms," in *Proc. IEEE Int. Conf. on Acoustics Speech and Signal Processing*, Montreal, Canada, 2004.

[22] W. Chu, Z. Ghahramani, and D. Wild, "Protein secondary structure prediction using sigmoid belief networks to parameterize segmental semi-markov models," in *Proc. European Symposium on Artificial Neural Networks*, 2004.

[23] N. C. Fitzkee and G. D. Rose, "Steric restrictions in protein folding: An α-helix cannot be followed by a contiguous β-strand," *Protein Science*, Feb. 2004.

[24] T. M. Klingler and D. L. Brutlag, "Protein science," *Discovering structural correlations in α-helices*, vol. 3, pp. 1847–1857, 1994.

[25] K. Sjölander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I. S. Mian, and D. Haussler, "Dirichlet mixtures: A method

for improving detection of weak but significant protein sequence homology," *Computing Applications in the Biosciences*, vol. 12, no. 4, pp. 327–345, 1996.

[26] O. Winther and A. Krogh, "Teaching computers to fold proteins," *Phys. Rev. E*, 2004, to appear.

[27] R. L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of The IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[28] G. Pollastri and P. Baldi, "Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners," *Bioinformatics*, vol. 18 Suppl 1, pp. S62–S70, 2002.

[29] G. E. Crooks and S. E. Brenner, "Protein secondary structure: Entropy, correlations and prediction," *Bioinformatics*, vol. 20, pp. 1603–1611, 2004.

[30] P. Burman, "A comparative study of ordinary cross validation, v-fold cross validation and the repeated learning-testing methods," *Biomatrika*, vol. 76, no. 3, pp. 503–514, 1989.

[31] M. Stone, "Cross-validatory choice and assessment of statistical predictions (with discussion)," *Journal of the Royal Statistical Society B*, vol. 36, pp. 111–147, 1974.

[32] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*. Springer, 2001.

[33] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochem. Biophys*, vol. 405, pp. 442–451, 1975.

[34] A. Zemla, C. Venclovas, K. Fidelis, and B. Rost, "A modified definition of sov, a segment-based measure for protein secondary prediction assessment," *Proteins: Structure, Function, and Genetics*, vol. 34, pp. 220–223, 1999.

[35] J. A. Cuff and G. J. Barton, "Evaluation and improvement of multiple sequence methods for protein secondary structure prediction," *Proteins: Structure, Function and Genetics*, vol. 34, pp. 508–519, 1999.

[36] W. Kabsch and C. Sander, "A dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, pp. 2577–2637, 1983.

[37] B. Rost and V. Eyrich, "Eva: large-scale analysis of second-ary structure prediction," *Proteins*, vol. 45, Suppl. 5, pp. 192–199, 2001.

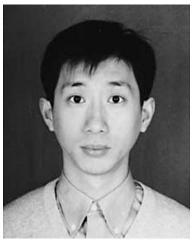
[38] D. Przybylski and B. Rost, "Alignments grow, secondary structure prediction improves," *Proteins*, vol. 46, pp. 197–205, 2002.

[39] A. Murzin, S. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *J. Mol. Biol.*, vol. 247, pp. 536–540, 1995.

[40] M. S. Cline, K. Karplus, R. Lathrop, T. Smith, R. Rogers Jr., and D. Haussler, "Information-theoretic dissection of pairwise contact potentials," *Proteins: Structure, Function, and Bioinformatics*, vol. 49, pp. 7–14, 2002.

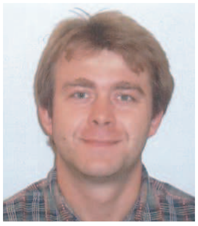
[41] G. E. Crooks, J. Wolfe, and S. E. Brenner, "Measurements of protein sequence-structure correlations," *Proteins: Structure, Function, and Bioinformatics*, vol. 57, pp. 804–810, 2004.

[42] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.



Wei Chu received the B.E. degree in automatic control from Harbin Engineering University, China, in 1995, the M.E. degree in inertial navigation systems from the 3rd Research Academy of the China Aerospace Cooperation, in 1998, and the Ph.D. degree from the Control Division of the Department of Mechanical Engineering, National University of Singapore, in 2003. He is currently a research fellow at the Gatsby Computational Neuroscience Unit, University College London, UK. His research interests include machine learning and bioinformatics.

Zoubin Ghahramani obtained BA and BSEng degrees from the University of Pennsylvania in Cognitive Science and Computer Science, and a PhD from MIT supervised by Michael Jordan. From 1995 to 1998 he was a Postdoctoral Fellow in Computer Science working with Geoffrey Hinton at the University of Toronto. He then obtained a faculty position at the Gatsby Computational Neuroscience Unit, University College London where he worked from 1998 to 2005. In 2006, he is joining the University of Cambridge as Professor of Information Engineering. He also maintains an appointment as Associate Research Professor in the School of Computer Science at Carnegie Mellon University. Zoubin's current research focuses on using Bayesian statistics as a framework for artificial intelligence and machine learning. Zoubin also works on applications of Bayesian machine learning to problems in bioinformatics, information retrieval, and pattern recognition. He serves on the editorial boards of several machine learning, statistics and pattern recognition journals, and on the programme committees of numerous conferences, but prefers to spend his time doing research.



Alexei Podtelezhnikov received his M.S. in Applied Physics and Mathematics from the Moscow Institute of Physics and Technology in 1994 and earned his Ph.D in Biomolecular Chemistry from New York University in 2000. His is currently a postdoctoral fellow at the Keck Graduate Institute. His research interests include but are not limited to bioinformatics, structural biology, and computational chemistry.

David L. Wild received a B.A. in physics from the University of York and a D.Phil. in molecular biophysics from the University of Oxford and has extensive experience in structural and computational molecular biology. He has worked at the Salk Institute's Structural Biology Laboratory, the European Molecular Biology Laboratory, and in industry with Allelix Biopharmaceuticals, Oxford Molecular and GlaxoWellcome. He is currently Associate Professor and Director of Computing at the Keck Graduate Institute of Life Sciences in Claremont, California. His research interests encompass bioinformatics, systems and structural biology.