A Graphical Model for Protein Secondary Structure Prediction

Wei Chu CHUWEI@GATSBY.UCL.AC.UK
Zoubin Ghahramani ZOUBIN@GATSBY.UCL.AC.UK
Gatsby Computational Neuroscience Unit, University College London, London, WC1N 3AR, UK

David L. Wild DAVID_WILD@KGI.EDU
Keck Graduate Institute of Applied Life Sciences, Claremont, CA 91171, USA

Abstract

In this paper, we present a graphical model for protein secondary structure prediction. This model extends segmental semi-Markov models (SSMM) to exploit multiple sequence alignment profiles which contain information from evolutionarily related sequences. A novel parameterized model is proposed as the likelihood function for the SSMM to capture the segmental conformation. By incorporating the information from long range interactions in β-sheets, this model is capable of carrying out inference on contact maps. The numerical results on benchmark data sets show that incorporating the profiles results in substantial improvements and the generalization performance is promising.

1. Introduction

Protein secondary structure prediction remains an important step on the way to full tertiary structure prediction in computational biology. 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 (1988), many of these methods have utilized neural networks. A major improvement in the prediction accuracy of these methods was made by Rost and Sander (1993), 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 type of alignment profile, position-specific scoring matrices (PSSM) derived by the iterative search procedure PSI-BLAST (Altschul et al., 1997), has been used in neural network prediction methods to achieve further improvements (Jones, 1999; Cuff & Barton, 2000).

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. (1993). Generalized HMMs with explicit state duration, also known as segmental semi-Markov models (SSMMs), have been widely applied in the field of gene identification (Burge & Karlin, 1997; Yel et al., 2001; Zhang et al., 2003; Korf et al., 2001). Recently, Schmidler (2002) presented a particular SSMM for protein structure prediction, which is an interesting statistical generative model for sequence-structure relationships. One advantage of the 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 (2002) still falls short of the best contemporary discriminative methods. Incorporation of the profiles from multiple sequence alignments into the model might be a plausible way to improve the performance. In this paper, we propose a novel parameterized model as the likelihood function for the SSMM to exploit the information provided by the profiles. Moreover, it is straightforward to incorporate long range interaction information in β-sheets into the modelling. We describe a Markov Chain Monte Carlo sampling scheme to perform inference in this model, and demonstrate the capability of the parametric SSMM to carry out inference on β-sheet contact maps in the Bayesian segmental framework. This ability to infer contact maps

Appearing in *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004.

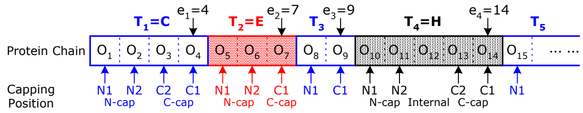

Figure 1. Presentation of the secondary structure of a protein chain in terms of segments. The square blocks denote our observations on these amino acid residues, which is a realization of a multinomial random variable. The rectangular blocks with solid borders denote the segments. The graph represents 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. 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.

represents one of the advantages of the 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 with details in section 2. In section 3 we extend the model to incorporate long range interactions. In section 4 we discuss the issue of parameter estimation. In section 5 a sampling scheme for inference is given, and we point out the capability to infer contact maps in section 6. In section 7 we present the results of numerical experiments, and conclude in section 8.

2. Model Description

The key idea in our model for secondary structure prediction is the alignment profile derived by multiple sequence alignment[1] or PSI-BLAST[2]. For a sequence of n amino acid residues, we can search a sequence database for several other sequences which are similar enough at the sequence level to be evolutionarily related. By aligning these sequences and counting the number of occurrences of each amino acid at each location, we obtain an alignment profile. Formally, the alignment profile $O = [O_1, O_2, \ldots, O_i, \ldots, O_n]$ is a sequence of 20×1 vectors, where O_i contains the occurrence counts for the 20 amino acids at location i which can be regarded as a realization of a multinomial random variable. 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 $T_i \in \mathcal{T}\ \forall i$. It is worth noting the existence of helical capping signals within segments (Aurora & Rose, 1998), which 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. In Figure 1, we present an illustration for the specification of the secondary structure of an observed sequence along with the definition of capping positions within segments. Based on the 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.

The segmental semi-Markov model (SSMM) (Ostendorf et al., 1996) is a generalization of hidden Markov models that allows each hidden state to generate a variable length sequence of the observations. In segment modelling, the segment types are regarded as the 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. The secondary structure prediction problem consists of computing $\mathcal{P}(m, e, T|O)$ for an observed sequence O. For this purpose we need to define the prior $\mathcal{P}(m, e, T)$ and the likelihood $\mathcal{P}(O|m, e, T)$. This Bayesian framework is described with more details in the following.

2.1. Prior Distribution

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

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

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

[1] The techniques of pairwise sequence comparison are employed to search a non-redundant protein sequence database for homologs of query sequence. These are then aligned using standard multiple alignment techniques (Thompson et al., 1994). Ideally, a column of aligned residues occupy similar structural positions and all diverge from a common ancestral residue.

[2] PSI-BLAST (Altschul et al., 1997) 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 this sequence in a position-specific score matrix. The position-specific score matrix can be mapped as relatively occurrence counting (Jones, 1999).

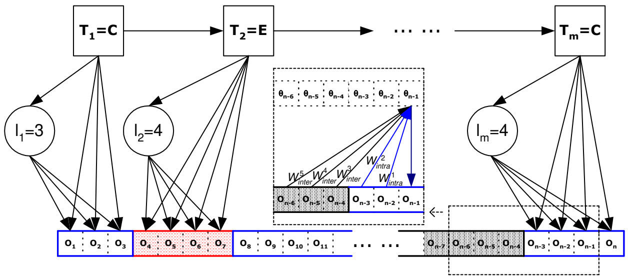

Figure 2. The segmental semi-Markov model illustrated as generative processes. A variable-length segment of observations associated with random length l_i is generated by the state T_i. The observations within a segment need not be fully correlated, while there might be dependencies between the residues 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 O_{n-1}, while θ_{n-1} is assumed to be dependent on $O_{n-6}, \ldots, O_{n-2}$ and the capping position of O_{n-1}.

$\mathcal{P}(l_i|T_i)$ is the segmental length distribution of the type T_i, where $l_i = e_i - e_{i-1}$ with $e_0 = 0$. 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.

2.2. 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 $\{m, e, T\}$:

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

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 written as a product of the conditional probabilities of individual observations

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

where O_k is the 20×1 count vector obtained from the alignment profile at the k-th residue. The likelihood function $\mathcal{P}(O_k|O_{[1:k-1]}, T_i)$ for each residue should be capable of capturing the core features of the segmental composition, such as segmental dependency (Eisenberg et al., 1984) and helical capping signals (Aurora & Rose, 1998). Schmidler et al. (2000) proposed a helical segment model with lookup tables to capture helical capping signals and the hydrophobicity dependency of segmental residues. However, this method is designed to use the residue sequence only, and its secondary structure prediction accuracy falls short of the best contemporary methods. Incorporation of the alignment profiles into the model might be a plausible way to improve the performance. Here, we propose a new parameterization for the likelihood function to exploit the information in the profile.

2.2.1. Multinomial Distribution

We assume that O_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 O_k is a 20×1 vector counting the occurrence of these outcomes. Thus, the probability of getting O_k can be evaluated by

$$\mathcal{P}(O_k|\theta_k, T_i) = \frac{(\sum_a O_k^a)!}{\prod_a O_k^a!} \prod_{a \in \mathcal{A}} (\theta_k^a)^{O_k^a} \qquad (4)$$

where $\mathcal{A}$ is the set of 20 amino acids, O_k^a is the element in O_k for the amino acid a, and θ_k^a denotes the probability of the outcome a with the constraint $\sum_a \theta_k^a = 1$.

2.2.2. Dirichlet Distribution

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 (refer to Figure 1).

The underlying causal impact on the current multinomial distribution, where we observed O_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 (Sjölander et al., 1996). In our case, this can be explicitly parameterized by weight matrices with *positive elements* as follows:

$$\mathcal{P}(\theta_k|O_{[1:k-1]}, T_i) = \frac{\Gamma(\sum_a \gamma_k^a)}{\prod_a \Gamma(\gamma_k^a)} \prod_{a \in \mathcal{A}} (\theta_k^a)^{\gamma_k^a - 1} \qquad (5)$$

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

$$\boldsymbol{\gamma}_k = W_{cap} + \sum_{j=1}^{\ell_k} W_{intra}^j \cdot O_{k-j} + \sum_{j=\ell_k+1}^{\ell} W_{inter}^j \cdot O_{k-j} \qquad (6)$$

with ℓ is the length of dependency window,[3] $\ell_k = \min(k - e_{i-1} - 1, \ell)$, and weight vectors W_{cap} of size 20×1 are used to capture capping signals at the capping position *cap* of O_k. Weight matrices W_{intra} and W_{inter} of size 20×20 are used to capture both

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

intra-segmental and inter-segmental dependency respectively, where the superscript denotes the residue interval. $\Gamma(\cdot)$ is the Gamma function defined as $\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t)\, dt$. The constraint $\gamma_k^a > 0\ \forall a$ is guaranteed by constraining the weight variables to have positive values. Note that we have used two sets of positioning indices for each residue: a sequential number k where $1 \leq k \leq n$, and a capping position *cap* where $cap \in \{\mathrm{N1}, \mathrm{N2}, \ldots, \mathrm{Internal}, \ldots, \mathrm{C2}, \mathrm{C1}\}$. 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, $\boldsymbol{W}_\tau = \{W_{\mathrm{N1}}, \ldots, W_{\mathrm{C1}}, W^1_{intra}, \ldots, W^\ell_{intra}, W^1_{inter}, \ldots, W^\ell_{inter}\}$.

2.2.3. Dirichlet-Multinomial Distribution

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

$$\begin{aligned} &\mathcal{P}(O_k|O_{[1:k-1]}, T_i) \\ &= \int_{\theta_k} \mathcal{P}(O_k|\theta_k, T_i)\mathcal{P}(\theta_k|O_{[1:k-1]}, T_i)\, d\theta_k \\ &= \frac{\Gamma(\sum_a \gamma_k^a) \cdot \prod_a \Gamma(\gamma_k^a + O_k^a)}{\Gamma\left(\sum_a(\gamma_k^a + O_k^a)\right) \cdot \prod_a \Gamma(\gamma_k^a)} \cdot \frac{(\sum_a O_k^a)!}{\prod_a O_k^a!} \end{aligned} \quad (7)$$

where $\Gamma(\cdot)$ denotes the Gamma function, and $\boldsymbol{\gamma}_k$ is defined as in (6).

2.3. 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,

$$\mathcal{P}(m, e, T|O) = \frac{\mathcal{P}(O|m,e,T)\mathcal{P}(m,e,T)}{\mathcal{P}(O)} \quad (8)$$

where $\mathcal{P}(O) = \sum_{\{m,e,T\}} \mathcal{P}(O|m, e, T)\mathcal{P}(m, e, T)$ as the normalizing factor. In this framework, we consider some important measures of the segmental variables for an observed sequence, such as

• 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 marginal posterior mode estimate is defined as $\arg\max\limits_{T} \mathcal{P}(T_{O_i}|O)$, where T_{O_i} denotes the segment type at the i-th observation.

The Viterbi and forward-backward algorithms for SSMM (Rabiner, 1989) can be employed for the MAP and marginal posterior mode estimate respectively.

3. Long Range Interactions in β-sheets

We have set up a Bayesian framework to predict the secondary structure. However, the secondary structure is affected not only by local sequence information,

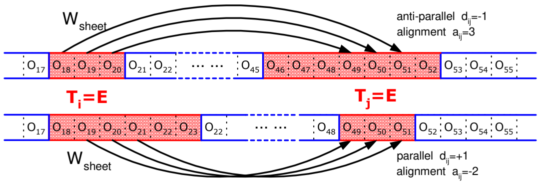

Figure 3. Anti-parallel (top), and parallel (bottom), pairs of interacting segments, S_i and S_j. d_{ij} is the binary variable for alignment direction, and a_{ij} is the integer variable for alignment position. A weight matrix W_{sheet} is introduced to capture the distal residue interactions.

but also by long range interactions with distal regions of the amino acid sequence. An important example is a β sheet which is 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 direction of the polypeptide chain of β-strands is identical, a *parallel* β-sheet, or the strand alignment is in an 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. An 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. $e_{j-1} + 1$.[4] The challenge for a predictive approach is how to introduce these long range interactions into the model. In this section, we extend the parametric model to incorporate the information of long range interactions in β-sheets.

3.1. Prior Specification for Distal Interactions

A set of random variables is introduced to describe the long range interactions, collected as $\mathcal{I} = \left\{\{S_j, S_{j'}, d_{jj'}, a_{jj'}\}_{j=1}^r\right\}$, where r is the number of interacting pairs and $\{S_j, S_{j'}, d_{jj'}, a_{jj'}\}$ 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 (1) and the conditional probability $\mathcal{P}(\mathcal{I}|m, e, T)$ can be further factored as

$$\begin{aligned} \mathcal{P}(\mathcal{I}|m, e, T) = \mathcal{P}(r|k)\mathcal{P}(\{S_j, S_{j'}\}_{j=1}^r)\cdot \\ \textstyle\prod_{j=1}^r \mathcal{P}(d_{jj'}|S_j, S_{j'})\mathcal{P}(a_{jj'}|S_j, S_{j'}, d_{jj'}) \end{aligned} \quad (9)$$

where r is the number of interacting pairs, k is the number of β-strands, and $\{S_j, S_{j'}\}_{j=1}^r$ denotes a com-

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

bination for β-strands to form r interacting pairs. Various specifications for these distributions in (9) are applicable provided that they satisfy $\sum_{\mathcal{I}} \mathcal{P}(\mathcal{I}|m, e, T) = 1$. In the present work, we assume a uniform distribution, $\mathcal{P}(\{S_j, S_{j'}\}_{j=1}^r) = \frac{1}{c(r,k)}$ if the combination is valid, where $c(r, k)$ is the total number of valid combinations,[5] otherwise $\mathcal{P}(\{S_j, S_{j'}\}_{j=1}^r) = 0$. $\mathcal{P}(r|k)$, $\mathcal{P}(d_{jj'}|S_j, S_{j'})$ and $\mathcal{P}(a_{jj'}|S_j, S_{j'}, d_{jj'})$ are discrete distributions. We may specify them according to our prior knowledge or learn them from training data.

3.2. Joint Segmental Likelihood

It is straightforward to extend the parametric model (7) 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

$$\begin{aligned}&\mathcal{P}(S_i|T_i = E, S_{-i}, \mathcal{I})\\ &= \prod_{k=e_{i-1}+1}^{e_i} \frac{\Gamma(\sum_a \tilde{\gamma}_k^a) \cdot \prod_a \Gamma(\tilde{\gamma}_k^a + O_k^a)}{\Gamma(\sum_a (\tilde{\gamma}_k^a + O_k^a)) \cdot \prod_a \Gamma(\tilde{\gamma}_k^a)} \cdot \frac{(\sum_a O_k^a)!}{\prod_a O_k^a!}\end{aligned} \tag{10}$$

with $\tilde{\boldsymbol{\gamma}}_k = \boldsymbol{\gamma}_k + \sum_{\{k^*\}} W_{sheet} \cdot O_{k^*}$ where $\boldsymbol{\gamma}_k$ is defined as in (6) and $\{k^*\}$ denotes the set of interacting residues of O_k that can be determined by $\mathcal{I}$.

4. Parameter Estimates

The probabilistic model we describe above has two classes of free parameters: a) the parameters that specify discrete distributions, which include the state transition probabilities for $\mathcal{P}(T_i|T_{i-1})$ in (1),[6] the segmental length distributions $\mathcal{P}(e_i|e_{i-1}, T_i)$ in (1); b) the weights in the segmental likelihood (7) and (10), which consist of three sets for different segmental types, i.e. $\{\boldsymbol{W}_\tau\}$ for $\tau \in \mathcal{T}$.

The parameters that specify discrete distributions can be directly estimated by their relative frequency of occurrence in the training data set.[7] For a segment type τ, a Maximum A Posteriori estimate of its associated weights $\boldsymbol{W}_\tau$ can be obtained as

$$\arg\max_{\boldsymbol{W}_\tau} \mathcal{P}(\{O, m, e, T, \mathcal{I}\}|\boldsymbol{W}_\tau)\mathcal{P}(\boldsymbol{W}_\tau) \tag{11}$$

under the condition of positive elements, where $\mathcal{P}(\boldsymbol{W}_\tau)$ is the prior probability usually specified by $\mathcal{P}(\boldsymbol{W}_\tau) \propto \exp(-\frac{C_\tau}{2}\|\boldsymbol{W}_\tau\|_2^2)$ with $C_\tau \geq 0$. The optimal $\boldsymbol{W}_\tau$ is therefore the minimizer of the negative logarithm of (11), which can be obtained by

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

subject to $w > 0, \forall w \in \boldsymbol{W}_\tau$, where $\sum_{\{O\}}$ means the sum over all the sequences, $\sum_{\{\tau\}}$ denotes the sum over all the segments of type τ, and $\mathcal{P}(S_i|S_{-i}, \tau)$ is defined as in (3). $\mathcal{L}(\boldsymbol{W}_\tau)$ is a regularized functional, and the optimal regularization factor C_τ can be determined by cross validation.[8] A set of auxiliary variables $\mu = \ln w$ can be introduced to convert the constrained optimization problem into an unconstrained problem, and then standard gradient-based optimization methods are employed to minimize $\mathcal{L}(\boldsymbol{W}_\tau)$.

5. Sampling Scheme for Inference

Generally, the introduction of long range interactions into the graphical model makes exact calculation of posterior probabilities intractable. Markov Chain Monte Carlo (MCMC) algorithms can be applied here to obtain approximate inference. A series of samples will be collected according to the joint distribution in the Markov chain simulation. As the dimension of the variable space varies in this process, the Metropolis-Hasting scheme can be applied with a reversible-jump approach (Green, 1995), which ensures that jumps between models of differing dimension are reversible.

What we are interested in here is the posterior distribution $\mathcal{P}(m, e, T|O)$ which is proportional to the joint distribution $\mathcal{P}(m, e, T, O)$. The joint distribution can be evaluated as

$$\begin{aligned}\mathcal{P}(m, e, T, O) = \mathcal{P}(m, e, T) &\textstyle\prod_{S_i \notin \mathcal{I}} \mathcal{P}(S_i|S_{-i}, T_i)\\ &\cdot \textstyle\sum_{\mathcal{I}} \mathcal{P}(\mathcal{I}|m, e, T) \prod_{S_i \in \mathcal{I}} \mathcal{P}(S_i|S_{-i}, T_i)\end{aligned} \tag{12}$$

where $\mathcal{P}(m, e, T)$ is defined as in (1), and only the segments of β-strands are in the interaction set $\mathcal{I}$. Schmidler (2002) proposed an MCMC algorithm by sampling in the posterior distribution $\mathcal{P}(m, e, T, \mathcal{I}|O)$, in which the dependency between (m, e, T) and $\mathcal{I}$ makes the sampling scheme complicated. However the main idea of the reversible jump approach is still applicable here. The following set of Metropolis proposals are defined for the construction of a Markov chain on the space of segmentations, denoted as $\mathcal{V} = (m, e, T)$:

- *Segment split*: propose $\mathcal{V}^* = (m^*, e^*, T^*)$ with $m^* =$

[5] 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.

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

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

[8] It is also possible to carry out approximate Bayesian inference on weight variables.

$m + 1$ 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 $T_{k^*+1} \sim \text{Uniform}[H, E, L]$.[9]

• *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 (12). *Segment split* and *Segment merge* jumps between segmentations of different dimension are accepted or rejected according to 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

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

where $|\mathcal{T}| = 3$ denotes the number of segment types. Due to the factorizations in (12), only the changed parts require evaluation. Once the interacting set $\mathcal{I}$ has been changed, the joint segmental likelihood has to be calculated again, which is a sum $\sum_{\mathcal{I}} \mathcal{P}(\mathcal{I}|m, e, T) \prod_{S_i \in \mathcal{I}} \mathcal{P}(S_i|T_i, S_{-i})$. Although the set $\mathcal{I}$ is composed of finite elements, it might be too expensive to traverse all of them. 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 (9).

6. Inference on Contact Maps

Contact maps represent the pairwise, inter-residue contacts as a symmetrical, square, boolean matrix. Pollastri and Baldi (2002) 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. 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

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

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

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

where the indicator function $\mathcal{C}^{ij}(\mathcal{I})$ is defined as in (13). Using the samples we have collected in the distributions $\mathcal{P}(m, e, T|O)$ and $\mathcal{P}(\mathcal{I}|m, e, T)$ (refer to Section 5), (14) can be estimated by

$$\begin{aligned} &\mathcal{P}(\mathcal{C}^{ij} = 1|O) = \textstyle\sum_{m,e,T} \sum_{\mathcal{I}} \mathcal{C}^{ij}(\mathcal{I})\, \mathcal{P}(m, e, T, \mathcal{I}|O) \\ &\approx \tfrac{1}{\mathcal{N}} \textstyle\sum_{\{m,e,T\}} \sum_{\{\mathcal{I}\}} \mathcal{C}^{ij}(\mathcal{I}) \frac{\mathcal{P}(O|m,e,T,\mathcal{I})}{\sum_{\{\mathcal{I}\}} \mathcal{P}(O|m,e,T,\mathcal{I})} \end{aligned}$$

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 $\mathcal{P}(m, e, T|O)$.

7. Numerical Experiments

We implemented the proposed algorithm in ANSI C.[10] 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 factors $C_\tau = 0.01$ $\forall \tau$ were chose to estimate the optimal weights.[11]

7.1. 7-fold Cross Validation

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 (Cuff & Barton, 2000).[12] We used 3-state PDB definitions of secondary structure. We removed the proteins that are shorter than 30 residues, or longer than 550 residues, following Cuff and Barton (2000), to leave 480 proteins for cross validation training. Seven partitions were created randomly, and cross validation was carried out on these partitions. We used two kinds of alignment profiles: the multiple sequence alignment profiles (MSAP)

[9]Here $\sim \text{Uniform}[H, E, L]$ denotes uniformly sampling in the set $\{H, E, L\}$, while $[1 : m]$ means from 1 to m.

[10]The source code in ANSI C can be accessed at http://www.gatsby.ucl.ac.uk/∼chuwei/code/bsmpssp.tar.gz.

[11]These model parameters were determined by cross validation, but we also found that there is a large region around these settings where the model performs stably.

[12]The data set and the multiple sequence alignments profiles they generated can be accessed at http://www.compbio.dundee.ac.uk/∼www-jpred/data/.

Table 1. 7-fold cross validation results for secondary structure prediction on 480 protein sequences from CB513. "Sequence Only" denotes the algorithm of Schmidler et al. (2000); MSAP denotes our algorithm using multiple sequence alignment profiles; PSSM denotes our algorithm using position-specific score matrices. Q_3 denotes the overall accuracy. $Q^{obs} = \frac{TruePositive}{TruePositive+FalseNegative}$ and $Q^{pred} = \frac{TruePositive}{TruePositive+FalsePositive}$. 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	59.2%	65.1%	68.8%	71.5%	63.9%	72.8%
Q_H^{obs}	66.3%	66.7%	77.6%	78.7%	67.6%	74.0%
Q_E^{obs}	20.7%	46.3%	44.2%	58.9%	29.5%	56.8%
Q_C^{obs}	72.8%	73.2%	73.9%	71.9%	78.3%	79.8%
Q_H^{pred}	61.9%	68.6%	71.8%	74.0%	69.5%	78.8%
Q_E^{pred}	56.5%	58.9%	69.7%	67.2%	73.5%	71.9%
Q_C^{pred}	57.8%	64.7%	66.1%	71.2%	59.1%	69.0%

Table 2. The results of 7-fold cross validation on 480 proteins of CB513 reported by Cuff and Barton (2000), along with our results.

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 Cuff and Barton (2000)	71.5%
Our algorithm with PSIBLAST PSSM	72.8%

used by Cuff and Barton (2000), and position-specific score matrices (PSSM) as in Jones (1999). For comparison purposes, we also implemented the algorithm proposed by Schmidler et al. (2000), which uses the sequence information only.[13] The validation results are recorded in Table 1. We also cite the results reported by Cuff and Barton (2000) in Table 2 for reference. The results obtained from our model show a great improvement over those of Schmidler et al. (2000) on all evaluation criteria. Compared with the performance of the neural network methods with various alignment profiles as shown in Table 2, the prediction accuracy of our model is also competitive.[14] We observed that the marginal posterior mode is more accurate than the MAP estimate, which shows that averaging over all the possible segmentations helps.

7.2. Prediction of Contact Maps

We prepared a dataset with long range interaction information specified by the data files of Protein Data Bank (PDB). The dataset, a subset of CB513, is composed of 152 protein chains along with β-sheet definitions. This reduction was caused by the incompleteness in the long range interaction information in many of the original PDB files. We carried out 30-fold cross validation on this subset. In MCMC sampling, we collected 9000 samples. We have not yet observed significant improvement on secondary structure prediction accuracy in the sampling results over exact inference without interactions. This suggests that in our current model, the main determinants of β-sheet structure are local contributions rather than distal hydrogen-bonding information. The small size of training data might be another factor. However it is interesting that we can infer β-sheet contacts. We present two predictive contact maps in Figure 4 as examples. We have also computed the area under the ROC curve (AUC) for β-sheet contact prediction. The average AUC over these protein chains is 0.899 ± 0.086.

Table 3. Predictive results of our algorithm using PSSM on the protein data of CASP.

	CASP2 (20 chains)	CASP3 (36 chains)	CASP4 (40 chains)	CASP5 (56 chains)
Q_3	73.40%	71.12%	74.32%	74.03%
Q_H^{obs}	76.62%	73.12%	80.22%	80.43%
Q_E^{obs}	61.29%	56.35%	57.81%	59.52%
Q_C^{obs}	77.73%	78.88%	78.00%	76.81%
Q_H^{pred}	79.71%	74.91%	81.33%	76.95%
Q_E^{pred}	76.48%	78.39%	76.19%	78.10%
Q_C^{pred}	67.36%	65.99%	67.28%	69.88%

7.3. Test on CASP

The meetings of Critical Assessment of techniques for protein Structure Prediction (CASP) facilitate large-scale experiments to assess protein structure prediction methods. We extracted protein chains of the latest four meetings from the public web page of the Protein Structure Prediction Center http://predictioncenter.llnl.gov/.[15] Using the 480 chains from CB513 and their PSSM profiles as training data, we built up our model, and then carried out prediction on these CASP proteins. The predictive results of our algorithm are reported in Table 3 indexed by the meetings. These results indicate that our algorithm gives a performance that is very similar to that given by other contemporary methods.

8. Conclusion

In this paper, we propose a graphical model with a novel parametric likelihood function to exploit the information in alignment profiles. Long range interaction information in β-sheets can be directly incor-

[13] The source code in ANSI C can be accessed at http://www.gatsby.ucl.ac.uk/∼chuwei/code/bspss.tar.gz.

[14] It is also possible to further improve performance by constructing smoothers over current predictive outputs as Cuff and Barton (2000) did in their Jury networks.

[15] On this web site, we can find the predictive results produced by other contemporary methods.

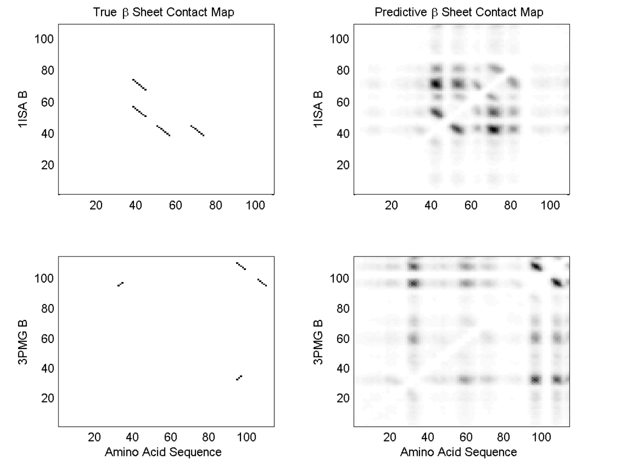

Figure 4. True β-sheet contact maps versus predictive maps on protein chains 1ISA_B and 3PMG_B. Gray scale indicates the probability $\mathcal{P}(\mathcal{C}^{ij} = 1|O)$.

porated into the model. The numerical results show that the generalization performance of this graphical model is competitive with other contemporary methods. Inference on contact maps can also be carried out in the Bayesian segmental framework. Moreover, with the inclusion of dihedral angle information in the joint sequence-structure probability distribution, this graphical model also has the potential for tertiary structure prediction.

Acknowledgments

This work was supported by the National Institutes of Health and its National Institute of General Medical Sciences division under Grant Number 1 P01 GM63208. We gratefully appreciate the reviewers' thoughtful comments.

References

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

Aurora, R., & Rose, G. D. (1998). Helix capping. *Protein Science*, *7*, 21–38.

Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, *268*, 78–94.

Cuff, J. A., & Barton, G. J. (2000). Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins: Structure, Function and Genetics*, *40*, 502–511.

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

Eisenberg, D., Weiss, R. M., & Terwilliger, T. C. (1984). The hydrophobic moment detects periodicity in protein hydrophobicity. *Proceedings of the National Academy of Sciences, USA*, *81*, 140–144.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*, 711–732.

Jones, D. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, *292*, 195–202.

Korf, I., Flicek, P., Duan, D., & Brent, M. R. (2001). Integrating genomic homology into gene structure prediction. *Bioinformatics*, *17 Suppl 1*, S140–S148.

Ostendorf, M., Digalakis, V., & Kimball, O. (1996). From HMM to segment models: a unified view of stochastic modelling for speech recognition. *IEEE Trans. on Speech and Audio Processing*, *4*, 360–378.

Pollastri, G., & Baldi, P. (2002). Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics*, *18 Suppl 1*, S62–S70.

Qian, N., & Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. *Journal of Mol. Biol.*, *202*, 865–884.

Rabiner, R. L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of The IEEE*, *77*, 257–286.

Rost, B., & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology*, *232*, 584–599.

Schmidler, C. S. (2002). *Statistical models and monte carlo methods for protein structure prediction.* Ph.D. thesis, Stanford University.

Schmidler, C. S., Liu, J. S., & Brutlag, D. L. (2000). Bayesian segmentation of protein secondary structure. *Journal of Computational Biology*, *7*, 233–248.

Sjölander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I. S., & Haussler, D. (1996). Dirichlet mixtures: A method for improving detection of weak but significant protein sequence homology. *Computing Applications in the Biosciences*, *12*, 327–345.

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensistivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, *22*, 4673–4680.

Yel, R. F., Lim, L. P., & Burge, C. B. (2001). Computational inference of homologous gene structures in the human genome. *Genome Res.*, *11*, 803–816.

Zhang, L., Pavlovic, V., Cantor, C. R., & Kasif, S. (2003). Human-mouse gene identification by comparative evidence integration and evolutionary analysis. *Genome Res.*, *13*, 1190–1202.