ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 81-86

Protein Secondary Structure Prediction Using Sigmoid Belief Networks to Parameterize Segmental Semi-Markov Models

Wei Chu *, Zoubin Ghahramani
Gatsby Computational Neuroscience Unit, University College London,
London, WC1N 3AR, UK
David Wild
Keck Graduate Institute of Applied Life Sciences,
Claremont, CA 91171, USA

Abstract. In this paper, we merge the parametric structure of neural networks into a segmental semi-Markov model to set up a Bayesian framework for protein structure prediction. The parametric model, which can also be regarded as an extension of a sigmoid belief network, captures the underlying dependency in residue sequences. The results of numerical experiments indicate the usefulness of this approach.

1 Introduction

A variety of approaches have been proposed to derive the secondary structure of a protein from its amino acid sequence. Beginning with the seminal work of Qian and Sejnowski [4], many of these methods have utilized neural networks. A major improvement in the prediction accuracy of these methods was made by Rost and Sander [5], 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 a profile derived from multiple sequence alignments instead of training the networks on single sequences. Recently, Schmidler [6] presented an interesting statistical model for protein structure prediction, which is a segmental semi-Markov model (SSMM) for sequence-structure relationships. In the probabilistic framework, it is advantageous to incorporate varied sources of sequence information using a joint

This work was supported by the National Institutes of Health (NIH) and its National Institute of General Medical Sciences (NIGMS) division under Grant Number 1 P01 GM63208 (NIH/NIGMS grant title: Tools and Data Resources in Support of Structural Genomics).

*All the correspondence should be addressed to this author.

sequence-structure probability distribution based on structural segments; structure prediction can then be formulated as a general Bayesian inference problem. However, the potential capability of this model has not been fully exploited so far. In this paper, we propose a parametric likelihood function for the SSMM to capture the inter-residue dependency in protein sequences. The parametric model is a natural extension of sigmoid belief networks [2], a type of neural network. The key contribution of this work is to combine the structure of neural networks with the SSMM model, which results in a flexible parametric model with enhanced generalization capability.

The paper is organized as follows. In section 2 we build up the Bayesian framework for SSMM, and propose the parametric model for the likelihood. In section 3 we present the results of numerical experiments. We conclude in section 4.

2 Bayesian Framework

For a sequence of n amino acid residues, denoted as $R = [R_1, R_2, \ldots, R_n]$ with $R_i \in \mathcal{A}$ where $1 \leq i \leq n$ and $\mathcal{A}$ is the set of 20 kinds of amino acids, its associated secondary structure can be fully specified in the terms of segment locations and segment types. The segment location can be identified by the position of the last residue of the segment, denoted as $e = [e_1, e_2, \ldots, e_m]$ where m is the number of segments, and the sequence of segment types can be denoted as $T = [T_1, T_2, \ldots, T_m]$ with $T_j \in \mathcal{T}$ where $\mathcal{T}$ is the set of secondary structural types. We use three segment types. H is used for α-helix, E for β-strand and C for Coil. In Figure 1, we present a part of the primary sequence of the protein 2BRZ and its associated secondary structure as an illustration.

The segmental semi-Markov model (SSMM) [3] is a generalization of hidden Markov models that allows each hidden state to generate a variable length sequence of the observations. Now we follow the standard SSMM [6] [3] to set up an explicit probabilistic model for sequence-structure relationships. Given a residue sequence R, its associated secondary structures can be completely defined by the set of random variables $\{m, e, T\}$. In segment modelling, the segment types are regarded as the set of hidden discrete states. Each of the segment types possesses an underlying generator, which generates a variable-length sequence of the residues, i.e. the segment. A schematic depiction of the SSMM is presented in Figure 2 from the perspective of generative models, while a Bayesian framework will be described with more details in the following.

2.1 Prior Distribution

Let us specify a prior distribution $\mathcal{P}(m, e, T)$ for the variable set of secondary structures. Usually $\mathcal{P}(m, e, T)$ is factored as

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

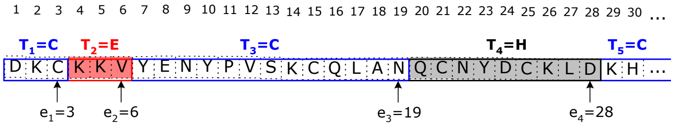

Figure 1: Presentation of the secondary structure of the protein 2BRZ in terms of segments. The square blocks denote the amino acid residues, and the rectangular blocks with solid borders denote the segments. The graph represents the residue sequence $R = [D, K, C, K, K, V, Y, \dots]$, the segment types $T = [C, E, C, H, C, \dots]$ and the segment endpoints $e = [3, 6, 19, 28, \dots]$.

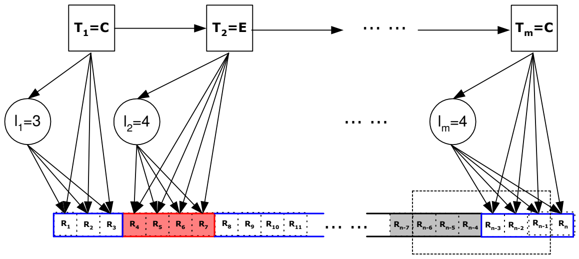
T1=C
T2=E
......
Tm=C
l1=3
l2=4
......
lm=4
R1 R2 R3 R4 R5 R6 R7 R8 R9 R10 R11
......
Rn-7 Rn-6 Rn-5 Rn-4 Rn-3 Rn-2 Rn-1 Rn

Figure 2: The segmental semi-Markov model illustrated as generative processes. A variable-length segment of observations is generated by the state T_i associated with random length l_i. The dotted rectangle denotes the dependency window for the residue R_{n-1}. The residues within a segment need not be fully correlated, while there might be dependencies between the residues in adjacent segments.

where the segment type only depends on the nearest previous neighbour in the sequence.[1] The state transition probabilities $\mathcal{P}(T_i|T_{i-1})$ 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 length distribution of each segment type. An improper uniform prior can be assigned for $\mathcal{P}(m)$.

2.2 Likelihood

Holding the assumption of segmental independence, the probability of our observations can be simply evaluated by

$$\mathcal{P}(R|m, e, T) = \prod_{i=1}^{m} \mathcal{P}(R_{[e_{i-1}+1:e_i]}|e_{i-1}, e_i, T_i) = \prod_{i=1}^{m} \mathcal{P}(S_i|T_i) \qquad (2)$$

where $S_i = R_{[e_{i-1}+1:e_i]} = [R_{e_{i-1}+1}, R_{e_{i-1}+2}, \dots, R_{e_i}]$ denotes the i-th segment. More generally, we can allow for dependency between residues in adjacent segments, and then the likelihood of the residues becomes

$$\mathcal{P}(R|m, e, T) = \prod_{i=1}^{m} \mathcal{P}(R_{[e_{i-1}+1:e_i]}|e_{i-1}, e_i, T_i, R_{[1:e_{i-1}]}) = \prod_{i=1}^{m} \mathcal{P}(S_i|T_i, S_{-i}) \quad (3)$$

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

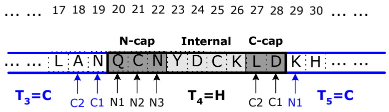

Figure 3: The graph of helical capping signals. The grey residues are of the N- and C-terminal positions for the α-helical segment, while the light-grey parts are internal.

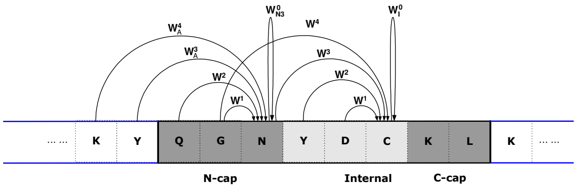

Figure 4: The graph of the parametric model for helical capping and segmental dependency with the window length $\ell = 4$. For the residue N, W^0_{N3} is used for the local contribution to capture the information at helical capping position N3, W^1 and W^2 are used for the contributions from intra-segmental dependency, while W^3_A and W^4_A are used for the inter-segmental contributions.

where $S_{-i} = [S_1, S_2, \ldots, S_{i-1}]$. The specific formulation of the segment likelihood $\mathcal{P}(S_i|T_i, S_{-i})$ should capture the core aspects of protein secondary structure, such as hydrophobicity patterns[2], helical capping signals[3] etc.

Schmidler et al. [6] proposed a helical segment model to capture position-specific preferences and dependency of intra-segmental residues, which used a lookup table with $3\times 3^\ell$ free parameters where ℓ is the length of the dependency window. We note that the lookup table is somewhat inadequate to generalize the dependency between residues, and the number of free parameters is exponential with the window-length ℓ.

Motivated by the structure of belief networks [2], we propose a parametric model for segmental likelihood evaluation. Weight matrices are introduced to represent the statistical relations between residues. The position-specific distributions are evaluated as local contributions W^0_p, a column vector with 20 elements, where p denotes capping or internal positions. Weight matrices of size 20×20 are proposed to capture both intra-segmental dependency, $W^1, \ldots, W^\ell$, and inter-segmental dependency, $W^1_A, \ldots, W^\ell_A$, where ℓ is the length of dependency window,[4] as shown in Figure 4. The residue R_k is denoted as a column vector with 20 elements in which only one element is 1, indicating the amino acid type, while others are zero. The segmental likelihood function in (3) can

[2] The 20 amino acids have varied physico-chemical properties. According to their chemical properties, we may group the 20 amino acids into roughly three classes: hydrophilic, neutral and hydrophobic. An α-helix exhibits periodicity in sequence hydrophobicity.

[3] Helical capping signals 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 (see Figure 3).

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

be explicitly given as

$$\mathcal{P}(S_i|T_i,S_{-i}) = \prod_{k=e_{i-1}+1}^{e_i} \mathcal{P}(R_k|T_i,R_{[1:k-1]}) = \prod_{k=e_{i-1}+1}^{e_i} \frac{\exp(-R_k^{\mathrm{T}}\cdot S(W))}{\sum_{R'_k}\exp(-{R'_k}^{\mathrm{T}}\cdot S(W))} \tag{4}$$

where W is the set of weight matrices associated with the segment type T_i, $S(W) = W^0_{p_k} + \sum_{j=1}^{\ell_k} W^j \cdot R_{k-j} + \sum_{j=\ell_k+1}^{\ell} W^j_A \cdot R_{k-j}$ with $\ell_k = \min(k - e_{i-1} - 1, \ell)$, p_k denotes the capping position for the residue R_k, and $\sum_{R'_k}$ denotes the sum over all the 20 possible residues. Note the linear relationship between the number of free parameters and the length of dependency window. Moreover, it is possible to further reduce the size of weight matrices to 3×3 by making use of the hydrophobicity class.

2.3 Posterior Distribution

Using Bayes' theorem, the posterior probability can be written as

$$\mathcal{P}(m,e,T|R) = \frac{\mathcal{P}(R|m,e,T)\mathcal{P}(m,e,T)}{\mathcal{P}(R)} \tag{5}$$

where the normalizing factor $\mathcal{P}(R) = \sum_{(m,e,T)} \mathcal{P}(R|m,e,T)\mathcal{P}(m,e,T)$. In this framework, we may consider some important measures of the segmental variables for an amino acid sequence [6], such as: 1. the distribution of the segment type at each residue: $\mathcal{P}(T_{R_i}|R)$ where we denote T_{R_i} as the segment type at the i-th residue, known as marginal posterior mode estimate; 2. the most probable segmental variables: $\arg\max_{m,e,T} \mathcal{P}(m,e,T|R)$, known as the MAP estimate. The forward-backward and Viterbi algorithms for SSMM [3] can be employed for the marginal posterior mode and MAP estimate respectively.

The parameters that specify discrete distributions can be directly estimated by their relative frequency of occurrence in the training data set. The optimal values of the weights in segmental likelihood can be estimated by maximum likelihood.

3 Numerical Experiments

In this section, we report the results of numerical experiments. For a fair comparison against the algorithm of Schmidler et al. [6], we used the reduced 3×3 weight matrices in the implementation of our approach. The length of the dependency window is fixed at 5 for all segment types. We used 635 proteins from the protein list generated by the PDB_SELECT algorithm [1] that only contains sequences with less than 25% sequence similarity,[5] and obtained the definition of their secondary structures from the data files of Protein Data Bank (PDB). We randomly partitioned the 635 proteins into 30 folds, and recorded

[5] The list of the 635 proteins we used in the numerical experiments can be found at http://www.gatsby.ucl.ac.uk/∼chuwei/biopss/pdbselect.id.

Table 1: Validation results of Schmidler et al.'s algorithm and our approach for secondary structure prediction. $Q^{obs} = \frac{TruePositive}{TruePositive+FalseNegative}$ and $Q^{pred} = \frac{TruePositive}{TruePositive+FalsePositive}$. Q_3 denotes the overall accuracy.

	Schmidler et al.		$3 \times 3 \times 5$	
	MODE	MAP	MODE	MAP
Q_3	66.89%	61.70%	**67.49%**	**62.05%**
Q_H^{obs}	70.13%	70.77%	**72.28%**	**71.70%**
Q_E^{obs}	46.51%	**24.69%**	**46.86%**	23.96%
Q_C^{obs}	**73.29%**	70.52%	72.64%	**70.86%**
Q_H^{pred}	69.42%	63.02%	**69.69%**	**63.36%**
Q_E^{pred}	60.11%	60.84%	**60.70%**	**61.81%**
Q_C^{pred}	67.02%	60.72%	**67.85%**	**60.97%**

the validation results in Table 1. The results obtained from our model show a modest improvement over those of Schmidler et al. [6] on all evaluation criteria.

4 Conclusion

In this paper, we proposed a parametric likelihood function for Bayesian segmental semi-Markov models to capture the inter-residue dependency in protein sequences. The results of numerical experiments in secondary structure prediction verify the feasibility of this approach. With the infusion of multiple sequence alignment profile or position-specific scoring matrices, the parametric model we have proposed would result in improved prediction accuracy.

References

[1] U. Holbohm and C. Sander. Enlarged representative set of protein structures. *Protein Science*, 3:522–524, 1994.

[2] R. M. Neal. Connectionist learning of belief networks. *Artificial Intelligence*, 56:71–113, 1992.

[3] 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*, 4(5):360–378, 1996.

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

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

[6] C. S. Schmidler, J. S. Liu, and D. L. Brutlag. Bayesian segmentation of protein secondary structure. *Journal of Computational Biology*, 7(1/2):233–248, 2000.