A Bayesian Model for Protein Secondary Structure Prediction

David B. Dahl[14], Qiwei Li[2], Marina Vannucci[2], Hyun Joo[3], and Jerry W. Tsai[3]
[1]Brigham Young University, Provo, Utah, United States
[2]Rice University, Houston, Texas, United States
[3]University of the Pacific, Stockton, California, United States
[4]Corresponding author: David B. Dahl, e-mail: dahl@stat.byu.edu

Abstract

This paper proposes a Bayesian model for secondary structure prediction given the primary structure. The method considers the packing influence of residues on secondary structure determination, including those packed close in space but distant in sequence. This modeling allows insights into the rules governing packing, filling a substantial gap in the current understanding of protein structure.

Key Words: Protein tertiary structure, Residue packing structure

1 Introduction

Advances in genomic sequencing technologies have made obtaining the primary structure (the linear sequence of amino acid) of a protein relatively cheap, accurate, and fast. For protein sequences of unknown biological function and/or structure, one standard and quite insightful analysis is a prediction of the protein sequences secondary structure. Current secondary prediction methods are based solely on similarity to sequences with known structure yet produce a prediction accuracy of around 80% (Rost 2001). To improve upon this accuracy, this paper develops a secondary structure prediction method that considers higher order information about the structure of protein residue packing provided by the knob-socket motif (Joo et al. 2012). A novel representation of packing structure is used where information about a residues secondary structure state is gained not only from local sequence but also from residues packed distant in sequence. The overall modeling allows insights into how packing governs secondary structure, filling a substantial gap in the current understanding of protein structure.

2 Proposed Model

2.1 Notation

Consider a protein whose primary structure is its observed amino acid sequence $\boldsymbol{a} = (a_1, \ldots, a_k)$, where a_i is a one-letter code denoting one of the 20 proteinogenic amino acids and k is the protein length. The secondary structure of a protein is the general form of its local segments, which we refer to as "block types." Kabsch and Sander (1983) proposed the Dictionary of Protein Secondary Structure (DSSP) for protein secondary structure with single letter codes. We consider the following 4 block types (in italics) from the original 8 structures defined in DSSP (in parentheses): 1. Helix "H": 3_{10} helices (G), α-helices (H), or π-helices (I), 2. Strand "E": extended strands in parallel and/or anti-parallel β-sheets (E), 3. Turn "T": hydrogen bonded turns of length 3, 4, or 5 amino acids (T), and 4. Coil "C": β-bridge residues (B), bends (S), or random coils (C). Let $\mathcal{S} = \{H, E, T, C\}$ denote the set of block types.

We introduce two equivalent parameterizations of the secondary structure. The *linear sequence* notation encodes the secondary structure using a vector $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_k)$, where $\rho_i \in \mathcal{S}$, indicating the secondary structure at each of the k positions. Equivalently, the secondary structure can be encoded in *block* notation using a vector $(\boldsymbol{\eta}, \boldsymbol{\lambda}) = ((\eta_1, \lambda_1), \ldots, (\eta_m, \lambda_m))$, where $\eta_j \in \mathcal{S}$ gives the secondary structures form repeated consecutively λ_j times in the j^{th} block. Note that $\lambda_j \in \{1, \ldots, k\}$ and $\sum_{j=1}^{m} \lambda_j = k$. For example, the following are equivalent secondary structures:

$$\boldsymbol{\rho} = (H, H, H, H, H, T, T, T, H, H, H, H, H) \iff (\boldsymbol{\eta}, \boldsymbol{\lambda}) = ((H, 5), (T, 3), (H, 5))$$

2.2 Sampling Model

We start by considering the joint distribution of the data $\boldsymbol{a} = (a_1, \ldots, a_k)$ given the latent secondary structure $(\boldsymbol{\eta}, \boldsymbol{\lambda})$. Assume the joint probability mass function (p.m.f.) $p(\boldsymbol{a}|\boldsymbol{\eta}, \boldsymbol{\lambda})$ is a product over blocks:

$$p(\boldsymbol{a}|\boldsymbol{\rho}) = p(\boldsymbol{a}|\boldsymbol{\eta}, \boldsymbol{\lambda}) = \prod_{j=1}^{m} p_{\eta_j}(a_{l_j}, \ldots, a_{u_j}),$$

where $l_j = 1 + \sum_{j' < j} \lambda_{j'}$, $u_j = \sum_{j' \leq j} \lambda_{j'}$, and p_{η_j} is one of p_H, p_E, p_T, and p_C based on the value of $\eta_j \in \mathcal{S} = \{H, E, T, C\}$, as described below.

2.2.1 Sampling Model for Helices

We propose that the sampling model for a helical block, with joint p.m.f. $p_H(a_l, \ldots, a_u)$, is defined by a product of three simpler p.m.f.'s p_{H1}, p_{H2}, p_{H3}, as follows:

$$\begin{aligned} p_H(a_l, \ldots, a_u) = \; & p_{H1}(a_l) \times \\ & p_{H2}(a_{l+1}|a_l)\, p_{H2}(a_{l+2}|a_{l+1})\, p_{H2}(a_{l+3}|a_{l+2}) \times \\ & p_{H3}(a_{l+4}|a_l, a_{l+1}, a_{l+3}) \;\cdots\; p_{H3}(a_u|a_{u-4}, a_{u-3}, a_{u-1}), \end{aligned}$$

where p_{H1} is a multinomial distribution with a category for each of the 20 amino acids, p_{H2} is a 20-dimensional multinomial distribution conditioned on the value of the antecedent amino acid, and p_{H3} is a 20-dimensional multinomial distribution conditioned on the values of the previous amino acid, the amino acid three positions back, and the amino acid four positions back. This formulation for $p_H(a_l, \ldots, a_u)$ is tractable, yet still respects the biochemistry of helices.

The 20-dimensional probability vector for p_{H1} is taken to be the posterior mean from a Bayesian model assuming a multinomial sampling model and a noninformative Dirichlet prior with all hyperparameters equal to 1. The data for this estimation is obtained from the PDB by counting the number of helical blocks that start with each of the 20 amino acids. This estimation is performed "offline," that is, this estimated probability vector is fixed and assumed known when evaluating the likelihood for a helical block.

Since there are 20 amino acids on which to condition, there are 20 p.m.f.'s of type p_{H2}. Likewise, since there are $20 \times 20 \times 20 = 8{,}000$ combinations of three amino acids, there are 8,000 p.m.f.'s of type p_{H3}. Again, these probability vectors are estimated from the PDB and assumed to be known when evaluating the helical likelihood.

2.2.2 Sampling Model for Strands

We propose that the sampling model for strands, with joint p.m.f. $p_E(a_l, \ldots, a_u)$, is defined by a product of three simpler p.m.f.'s p_{E1}, p_{E2}, p_{E3}, as follows:

$$\begin{aligned} p_E(a_l, \ldots, a_u) = \; & p_{E1}(a_l) \times \\ & p_{E2}(a_{l+1}|a_l) \times \\ & p_{E3}(a_{l+2}|a_l, a_{l+1})\, p_{E3}(a_{l+3}|a_{l+1}, a_{l+2}) \; \cdots \; p_{E3}(a_u|a_{u-2}, a_{u-1}), \end{aligned}$$

where p_{E1} is a multinomial distribution with a category for each of the 20 amino acids, p_{E2} is a 20-dimensional multinomial distribution conditioned on the value of the antecedent amino acid, and p_{E3} is a 20-dimensional multinomial distribution conditioned on the values of the previous two amino acids Again, this formulation $p_H(a_l, \ldots, a_u)$ is tractable, yet still respects the biochemistry of strands. Note that $p_{E1} \neq p_{H1}$ despite the fact that both are marginal multinomial distributions. Likewise, $p_{E2} \neq p_{H2}$ despite the fact that both are conditional multinomial distributions given an amino acid. In particular, p_{E1}, p_{E2}, and p_{E3} are estimated from PBD data involving strands, whereas p_{H1}, p_{H2}, and p_{H3} are estimated from PBD data involving helices. Still the estimation strategy for the probability vectors is the same.

2.2.3 Sampling Model for Turn

We propose that the sampling model for turns, with joint p.m.f. $p_T(a_l, \ldots, a_u)$, is defined by a product of simpler p.m.f.'s, as follows:

$$\begin{aligned} & p_T(a_l, \ldots, a_u) \\ & \quad = p_{T31}(a_l)\, p_{T32}(a_{l+2}|a_l)\, p_{T33}(a_{l+1}|a_l, a_{l+2}) && \text{if } u - l = 2 \\ & \quad = p_{T41}(a_l)\, p_{T42}(a_{l+3}|a_l)\, p_{T43}(a_{l+1}|a_l, a_{l+3})\, p_{T43}(a_{l+2}|a_l, a_{l+3}) && \text{if } u - l = 3 \\ & \quad = p_{T51}(a_l)\, p_{T52}(a_{l+4}|a_l)\, p_{T53}(a_{l+1}|a_l, a_{l+4})\, p_{T53}(a_{l+3}|a_l, a_{l+4}) \times \\ & \qquad p_{T54}(a_{l+2}|a_{l+1}, a_{l+3}) && \text{if } u - l = 4, \end{aligned}$$

where each condition in the equation above is estimated based on the PDB data using hydrogen bonded turns of length 3, 4, and 5 amino acids, respectively. If the length is larger than 5, we assume the first five units sample from the turn model and the other units sample from the coil model, which is described below.

2.2.4 Sampling Model for Coil

We propose that the sampling model for coils, with joint p.m.f. $p_C(a_l, \ldots, a_u)$, is defined by a product of simpler p.m.f.'s, as follows:

$$p_C(a_l, \ldots, a_u) = p_{C1}(a_l)\, p_{C2}(a_{l+1}|a_l)\, p_{C2}(a_{l+2}|a_{l+1}) \; \cdots \; p_{C2}(a_u|a_{u-1})$$

whose component distributions are again estimated from the PDB data.

2.3 Prior Distribution

The model is completed by specifying the prior distribution, with p.m.f. $p(\boldsymbol{\rho}) = p(\boldsymbol{\eta}, \boldsymbol{\lambda})$. Let m denote the number of blocks in $\boldsymbol{\rho}$. We consider a prior of the form:

$$p(\boldsymbol{\rho}) = p(\boldsymbol{\eta}, \boldsymbol{\lambda}) = p(m)p(\boldsymbol{\eta}|m)p(\boldsymbol{\lambda}|\boldsymbol{\eta}, m),$$

but this p.m.f. equals zero if, for $j = 1, \ldots, m$, any of the following conditions are met: 1. $\eta_1 \neq C$, 2. $\eta_m \neq C$, 3. $\lambda_j < 5$ and $\eta_j = H$, 4. $\lambda_j < 3$ and $\eta_j = E$, or 5. $\lambda_j < 3$ and

$\eta_j = T$. All of those conditions violate the biochemistry inherent in secondary structure. We specify the components of this hierarchical prior using $16,675$ amino acid sequences from the PDB, together with their corresponding secondary structures. The total number of analyzed blocks is $471,361$.

Based on a simple linear regression of the number of blocks m on the known length k of an amino acid sequence $\boldsymbol{a}$, we let the prior on the number of blocks be $m \sim$ Normal$($ $2.347253 + 0.154154k, (5.526382/d)^2)$, where d is a hyperparameter limiting the spread of the prior distribution on the number of blocks m.

For all the $16,675$ secondary structure $\boldsymbol{\rho}$'s, the secondary structure of the first and last block is coil. Therefore, we assume $p(\boldsymbol{\rho}) = 0$ if $\eta_1 \neq C$ or $\eta_K \neq C$. For all the other positions, we assume the number of the blocks of each type follow a multinomial distribution:

$$(m_H, m_E, m_T, m_C)|m \sim \text{Multinomial}(m-2, \boldsymbol{\theta}),$$

where $\boldsymbol{\theta} = (\theta_H, \theta_E, \theta_T, \theta_C) = (0.170207, 0.224021, 0.178348, 0.427424)$ is obtained from all the $471,361$ blocks in the PDB, except the first and last ones of each structure. So we can write the conditional prior of $\boldsymbol{\eta}$ given m as:

$$p(\boldsymbol{\eta}|m) = \frac{m!}{m_H! m_E! m_T! m_C!} \theta_H^{m_H} \theta_E^{m_E} \theta_T^{m_T} \theta_C^{m_C}.$$

Lastly, we consider the distribution of block length for each block type. The minimum block length for helix, strand, turn, and coil is 5, 3, 3, and 1, respectively. Because of the large variance, we use the negative binomial distribution to model the block length for each block form. Reading all the $471,361$ blocks in the PDB, the parameters are estimated by the maximum likelihood estimation (MLE) method and given as:

$$\lambda_j|\eta_j \sim \begin{cases} 5 + \text{Negative Binomial}(1.885880, 6.953392) & \text{if } \eta_j = H \\ 3 + \text{Negative Binomial}(2.521091, 2.899121) & \text{if } \eta_j = E \\ 3 + \text{Negative Binomial}(0.839557, 0.728294) & \text{if } \eta_j = T \\ 1 + \text{Negative Binomial}(0.990796, 3.725501) & \text{if } \eta_j = C. \end{cases}$$

3 MCMC Algorithms

Our goal is to make inference on the secondary structure $\boldsymbol{\rho}$ given the amino acid sequence $\boldsymbol{a}$. We use Markov chain Monte Carlo (MCMC) methods described below to sample from the posterior distribution $p(\boldsymbol{\rho}|\boldsymbol{a}) \propto p(\boldsymbol{a}|\boldsymbol{\rho})p(\boldsymbol{\rho})$. The Metropolis Hastings (MH) ratio can be written as:

$$r = \frac{p(\boldsymbol{\rho}^*|\boldsymbol{a})}{p(\boldsymbol{\rho}^{(t-1)}|\boldsymbol{a})} \frac{q(\boldsymbol{\rho}^{(t-1)}; \boldsymbol{\rho}^*)}{q(\boldsymbol{\rho}^*; \boldsymbol{\rho}^{(t-1)})},$$

where $q(\boldsymbol{\rho}^*; \boldsymbol{\rho}^{(t-1)})$ is the proposal density, the density for proposing a move to $\boldsymbol{\rho}^*$ given the previous state $\boldsymbol{\rho}^{(t-1)}$ and $q(\boldsymbol{\rho}^{(t-1)}; \boldsymbol{\rho}^*)$ is the reverse case. The move is accepted $\boldsymbol{\rho}^{(t)} = \boldsymbol{\rho}^*$ with the probability $\min(1, r)$, otherwise, the move is rejected and $\boldsymbol{\rho}^{(t)} = \boldsymbol{\rho}^{(t-1)}$.

In order to make the Markov chain ergodic and efficient, we design five types of moves: forward addition, backward addition, forward shift, backward shift, and sub-block replacement. We denote $Q_\kappa, \kappa = 1, 2, \cdots, 5$, as the probability to propose the corresponding type of move on each MCMC step.

The first two moves are to add a unit for j^{th} block from the following block $j+1$ (forward) or the previous block $j-1$ (backward) while keeping $\boldsymbol{\eta}$ unchanged. The ratio

of proposal densities is Q_1/Q_2 for forward case and Q_2/Q_1 for backward case, so the MH ratio is the ratio of the posterior probabilities of $\boldsymbol{\rho}^*$ and $\boldsymbol{\rho}^{(t-1)}$ multiplies Q_1/Q_2 or Q_2/Q_1.

To make our algorithms more efficient, we also design shift moves, that is, to shift forward or backward a randomly selected block j. The ratio of proposal densities is Q_3/Q_4 for forward case and Q_4/Q_3 for backward case, so the MH ratio is the ratio of the posterior probabilities for $\boldsymbol{\rho}^*$ and $\boldsymbol{\rho}^{(t-1)}$ multiplies Q_3/Q_4 or Q_4/Q_3. Notice that the only parameter we change in the above four types of moves is $\boldsymbol{\lambda}$ and the number of blocks m remains the same.

We also consider a Gibbs sampling algorithm for a sub-block replacement move. In this step, we randomly choose a block $j, j = 2, 3, \cdots, m$, except the first and last one. We cut out a section of this block, that is, start from a random starting point within the block and end up with a possible random length. Then, we calculate the conditional probability of generating this sub-block by each secondary structure form, H, E, T, and C. These four unnormalized conditional probabilities are assigned as the sampling weight to each form. Notice that we may change the whole block form η_j or split the block into two or three sections in this move. So the number of blocks m may change.

4 Results

We evaluated the performance of our proposed method using data from the PDB which contains not only the data (amino acid sequence $\boldsymbol{a}$) but also the true secondary structure $\boldsymbol{\rho}$ for thousands of proteins. We can then compare our estimates for $\boldsymbol{\rho}$ with the true value for many proteins. All computations were conducted in R on a Mac laptop with 2.3 GHz Core i7 CPU and 16 GB memory.

The settings of our algorithms are $d = 4$, $Q_1 = Q_2 = Q_3 = Q_4 = 0.1$, and $Q_5 = 0.6$. We randomly select 200 amino acid sequences out of $16,675$ from the PDB, whose lengths range from 27 to $1,114$ and implement the MCMC algorithms with $10,000$ iterations per sequence.

We report the following observations. First, the algorithm is fast. For example, $10,000$ iterations on a 100-long sequence only takes 30 seconds (CPU time) and 400 seconds (wall time). The runtime increases approximately linearly in length. Second, our algorithm has satisfactory convergence property. We run multiple independent chains starting from different randomly chosen secondary structure states. Trace plots of the posterior probability indicate convergence and good mixing, as shown in Figure 1 (Left). Third, our Metropolis-Hastings sampling scheme is very efficient. Discarding the first half iterations as burn-in period, the averaged acceptance rate for addition and shift moves are 40% and 29%, respectively.

We considered two ways to summarize the posterior distribution of $\boldsymbol{\rho}$ to yield a point estimator: 1. Choosing a particular $\boldsymbol{\rho}$ that maximize the posterior probability $p(\boldsymbol{\rho}|\boldsymbol{a})$ and 2. Selecting the most likely block form for each position. We name them the maximum *a posteriori* method (MAP) and the marginal probability method (MP), respectively. The accuracy is defined as the number of actual forms that are correctly predicted, divided by the number of all actual forms. Figure 1 (Right) shows an example to examine the accuracy and Table 1 lists the accuracy summary. The accuracy achieved by MAP and MP methods are almost the same. Among the four secondary structure forms, the helical sampling model performs better than others, with an accuracy of about 90% by the MP method.

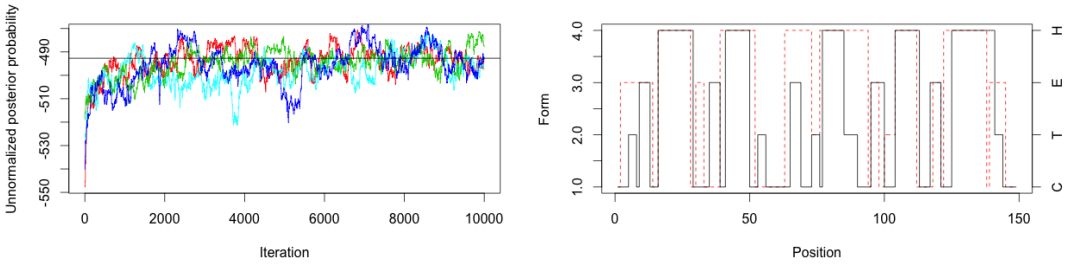

Figure 1: Consider the amino acid k66a as an example. Left: the trace plot of the unnormalized log posterior probability $p(\boldsymbol{\rho}|\boldsymbol{a})$ of multiple chains with the horizontal truth line; Right: the true (red dash line) structure and the predicted (black solid line) structure by MAP.

Table 1: Accuracy of maximum *a posteriori* (MAP) and marginal probability (MP) methods

	MAP					MP				
	Overall	Helix	Strand	Turn	Coil	Overall	Helix	Strand	Turn	Coil
Median	0.474	0.813	0.468	0.267	0.258	0.483	0.890	0.500	0.250	0.177
Mean	0.475	0.820	0.467	0.284	0.281	0.483	0.870	0.503	0.273	0.220
SD	0.115	0.122	0.165	0.202	0.146	0.132	0.136	0.164	0.200	0.142

5 Conclusion

The predictive performance of our method in its current form is inconsistent, but can be improved within the proposed framework. The predictive performance for helical structure is quite good, so we plan on reformulating the sampling models for coils, strands, and turns. We also hypothesize that we can improve the predictive accuracy by more accurate estimation of the probability vectors in the sampling model for secondary structure. We currently compute them using a multinomial sampling model and a noninformative Dirichlet prior. Take, for instance, the case of the $8,000$ p.m.f.'s of type p_{H3}. We hypothesize that many of the $8,000$ probability vectors of type p_{H3} will be similar and, as such, we can borrow strength in estimating the $8,000$ probability vectors using Bayesian nonparametric techniques. The first idea is to borrow strength using a Dirichlet process mixture model. We eventually propose to use a novel random partition distribution index by pairwise distances in which items that are "close" to each other (in terms of their biochemistry) cluster with higher probability than distant ones.

References

Joo, H., Chavan, A., Phan, J., Day, R., and Tsai, J. (2012), "An amino acid packing code for -helical structure and protein design," *Journal of Structural Biology*, 134, 234–54.

Kabsch, W. and Sander, C. (1983), "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, 22, 2577–2637.

Rost, B. (2001), "Review: protein secondary structure prediction continues to rise," *Journal of Structural Biology*, 134, 204–18.