PROTEIN SECONDARY STRUCTURE PREDICTION WITH SEMI MARKOV HMMS

Zafer Aydın[*], *Yücel Altunbaşak*[*], *Mark Borodovsky*[**]

[*]Center for Signal and Image Processing
Georgia Institute of Technology
Atlanta, GA 30332-0250
E-mail:{aydinz,yucel}@ece.gatech.edu

[**]Schools of Biology and Biomedical Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0280
E-mail: mark@amber.biology.gatech.edu

ABSTRACT

Secondary structure prediction has been an essential task in determining the structure and function of the proteins. Prediction accuracy is improving every year towards the 88% estimated theoretical limit [1]. There are two approaches for the secondary structure prediction. The first one, *ab initio* (single sequence) prediction does not use any homology information. The evolutionary information, if available, is used by the second approach to improve the prediction accuracy by a few percentages [2]. In this paper, we address the problem of single sequence prediction by developing a semi Markov HMM, similar to the one proposed by Schmidler *et al.* [2]. We introduce a better dependency model by considering the statistically significant amino acid correlation patterns at segment borders. Also, we propose an internal dependency model considering right to left dependencies without modifying the left to right HMM topology. In addition, we propose an iterative training method to better estimate the HMM parameters. Putting all these together, we obtained 1.5% improvement in three-state-per-residue accuracy.

1. INTRODUCTION

A protein is a biomolecule constructed from amino acid units. The adjacent amino acids of 20 different types are connected by a peptide bond. A protein chain could be represented by a string of amino acid sequence as illustrated in Fig. 1. Protein sequence analysis is an important area where the goal is to predict the structure and function of the newly identified proteins.

It has been shown that all the structural information about the protein is embedded in its amino acid sequence. There are several levels at which protein structure prediction can be performed. In secondary structure prediction, one is mainly concerned with the assignment of secondary structure elements to each amino acid residue as shown in Fig. 1. In tertiary structure estimation (*i.e.,* protein folding), the goal is to predict the conformation assumed by protein molecule in 3D space.

The three major secondary structure elements are α-helix {H}, β-strand {E} and loop {L}. α-helices are strengthened by hydrogen bonds between every fourth amino acid so that the protein backbone adopts a helical configuration. In β-strands the hydrogen bonding is non-local. They adopt a parallel or anti-parallel sheet configuration. Other structural elements such as bends and turns are classified as loops. Therefore a secondary structure prediction assigns for each amino acid a structural state from a 3-letter alphabet {H, E, L}, as depicted in Fig. 1. The secondary structure prediction is an important problem in protein sequence analysis. Accurate predictions provide insights into the molecular structure and function of a protein.

G	K	C	.....	N	T	F	V	←	Amino Acid Sequence
|	|	|		|	|	|	|		
L	L	E	.....	H	H	H	H	←	Secondary Structure Labels

Figure 1: Secondary Structure Prediction

To date, secondary structure prediction has benefited mostly from Machine Learning tools where Artificial Intelligence, Neural Networks and Hidden Markov Models played a central role. There are essential steps in the development of a machine learning based predictor. The first step is to perform a statistical analysis in order to explore the most informative correlations and patterns. This allows us to choose a model that represents the dependency behavior of various structure elements. The statistical analysis is followed by the the training phase where a training set is compiled and the model parameters are derived. Finally, in the testing phase, the performance is evaluated by making predictions for new test samples.

There are two aspects of secondary structure prediction. In *ab initio* or single sequence prediction, the test sequence does not exhibit significant similarity to any of the training sequences at the sequence level. This is a limiting factor for the prediction accuracy. On the other hand, if there are closely related sequences, this generally implies their structural similarity, and the predictions are improved by considering multiple alignments.

In this paper, we addressed the problem of single sequence predictions. First of all, we performed a statistical analysis to explore the most informative correlations for different secondary structures. Then, we chose a semi Markov HMM, which is similar to the model developed by Schmidler *et al.* [2]. In this model, we specifically considered correlations at terminal positions of structural segments and dependencies to forward [1] residues within the segments. Finally, we implemented an iterative estimation of the HMM parameters.

2. CORRELATION ANALYSIS

The first step in building a statistical model is to perform a statistical analysis in order to explore the dependency structure. We performed a χ^2-test to identify the most informative correlations

[1]Dependencies to the right positions in a left-to-right-topology

between amino acid pairs in different types of secondary structure segments and positions. The χ^2 test computes the joint distribution of amino acid pair, and compares it with the product of marginal distributions. The 20×20 contingency table shows the frequencies of possible amino acid pairs observed in different states and positions[2]. Throughout the analysis, we worked with 8100 proteins and their secondary structures collected from Protein Data Bank (PDB). DSSP assignments [3] for secondary structure were used to reduce eight structural states to three states where {H, G, I}→{H}, {E,B}→{E}, {S,T,' '}→{L}.

We first considered the correlations between amino acid pairs at various separation distances. Table 1 shows the results of the χ^2-test for the three secondary structure types. We found that in α-helix segments, a residue[3] at position i is highly correlated with residues at positions $i-2$, $i-3$ and $i-4$, where i denotes the position of the amino acid within a segment. Similarly, a β-strand residue had highest correlations with residues at positions $i-1$, $i-2$, and a loop residue had its most significant correlations with those at positions $i-1$, $i-2$ and $i-3$.

Next, we considered position specific correlations. The terminal positions, which are typically the first and last four positions of a secondary structure segment as illustrated in Table 2, are known to have different amino acid frequency distribution from the internal positions. Especially, α-helices are characterized by capping boxes where the hydrogen bonding patterns and side-chain interactions are different from the internal positions [4].

The results for terminal positions are given in Table 3 for the α-helix segments. This shows that there are statistically significant correlations between residues in terminal positions and the residues that are outside the segment. This can be explained by the fact that such terminal residues form their hydrogen bonds with residues that are outside the segment [4]. Another observation is that there exist significant correlations with the forward residues. Also, the degree of correlation for the forward residues might be different from those of backward, which indicates an asymmetrical dependency behavior for forward and backward residues. Internal positions also exhibit similar correlation pattern. For instance, for α-helix segments, the i^{th} residue in an internal position is highly correlated with $i-2^{nd}$, $i-3^{rd}$, $i-4^{th}$, $i+2^{nd}$ and $i+4^{th}$ residues. The degree of correlation between i^{th} and $i-2^{nd}$ residues is different from the degree of correlation between i^{th} and $i+2^{nd}$ residues.

3. THE SEMI MARKOV HMM

3.1. Derivation of the Model

In a typical HMM, there is a finite number of distinct hidden states. Hidden states in our case are structural states {H, E, L }. Each state generates an observation in the form of amino acid segment. Starting from an initial state, transitions occur from one state to the other, following a transition probability distribution. At each state an amino acid segment is generated according to the observation frequency distribution. For a thorough review on HMMs, see [5].

A secondary structure of a protein is defined by a vector $(m, \mathbf{S}, \mathbf{T})$, where m denotes the total number of segments, $\mathbf{S}$ represents the segment end positions and $\mathbf{T}$ represents the structural state of each segment (α-helix, β-strand or loop) A graphical representation for the case where $\mathbf{T}$ = (L, H, L, E, L,...) and $\mathbf{S}$ = (3, 8, 12, 15, ...) is depicted in Fig. 2.

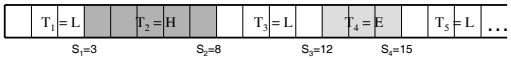

Figure 2: Representation of the secondary structure of a protein in terms of structural segments

The state prediction could be re-stated as a posterior maximization problem. That is, given the observation sequence of amino acids, denoted by $\mathbf{R}$, find the vector $(m, \mathbf{S}, \mathbf{T})$ with maximum posterior probability $P(m, \mathbf{S}, \mathbf{T} \mid \mathbf{R})$. Using Bayes rule, this probability could be expressed as follows:

$$P(m, \mathbf{S}, \mathbf{T} \mid \mathbf{R}) = \frac{P(\mathbf{R} \mid m, \mathbf{S}, \mathbf{T}) P(m, \mathbf{S}, \mathbf{T})}{P(\mathbf{R})}, \quad (1)$$

where $P(\mathbf{R} \mid m, \mathbf{S}, \mathbf{T})$ denotes the sequence likelihood and $P(m, \mathbf{S}, \mathbf{T})$ represents the apriori distribution. Maximizing $P(m, \mathbf{S}, \mathbf{T} \mid \mathbf{R})$ with respect to the state variables is equivalent to maximizing the product $P(\mathbf{R} \mid m, \mathbf{S}, \mathbf{T}) P(m, \mathbf{S}, \mathbf{T})$. Next, we will model each of these two probability terms.

We modeled the apriori distribution $P(m, \mathbf{S}, \mathbf{T})$ as follows:

$$P(m, \mathbf{S}, \mathbf{T}) = P(m) \prod_{j=1}^{m} P(T_j \mid T_{j-1}) P(S_j \mid S_{j-1}, T_j) \quad (2)$$

Here $P(m)$ is the probability of observing m secondary structure segments, and it is assumed to be independent from the other state variables. $P(T_j \mid T_{j-1})$ represents the state transition probability (among different secondary structure types), and $P(S_j \mid S_{j-1}, T_j)$ allows us to model the length distribution of secondary structure segments with the following assumption:

$$P(S_j \mid S_{j-1}, T_j) = P(S_j - S_{j-1} \mid T_j). \quad (3)$$

Next, the likelihood term $P(\mathbf{R} \mid m, \mathbf{S}, \mathbf{T})$ is modeled as

$$\begin{aligned} P(\mathbf{R} \mid m, \mathbf{S}, \mathbf{T}) &= \prod_{j=1}^{m} P(\mathbf{R}_{[S_{j-1}+1:S_j]} \mid \mathbf{S}, \mathbf{T}) \quad (4) \\ &= \prod_{j=1}^{m} P(\mathbf{R}_{[S_{j-1}+1:S_j]} \mid S_{j-1}, S_j, T_j) \end{aligned}$$

Here we assume the independence of segment likelihood terms. $\mathbf{R}_{p:q}$ denotes the sequence of residues with indices from p to q. $P(\mathbf{R}_{[S_j+1:S_j]} \mid \mathbf{S}, \mathbf{T})$ represents the probability of observing a particular amino acid segment given all state variables. It is equal to $P(\mathbf{R}_{[S_{j-1}+1:S_j]} \mid S_{j-1}, S_j, T_j)$ because in a HMM the symbol observation probability depends only on its generator state.

Although the observation probability of amino acids at different secondary structure states is assumed to be independent, the amino acids within the segments are allowed to depend on neighboring residues. To reflect this dependency, $P(\mathbf{R}_{[S_j+1:S_j]} \mid \mathbf{S}, \mathbf{T})$

[2]The threshold was computed as 404.6 for a statistical significance level of 0.05.

[3]Residue refers to the amino acid.

	Helix		Strand		Loop	
Separation	χ^2	# of pairs	χ^2	# of pairs	χ^2	# of pairs
1	3708.74	192,055	8700.16	136,631	6760.65	271,097
2	11408.08	166,037	2812.36	210,150	2812.36	210,150
3	4123.08	140,019	5836.95	69,335	2716.70	162,491
4	8168.01	119,352	3996.21	45,407	1601.23	124,711
5	3340.79	102,328	2631.21	27,829	1382.10	96,342
6	2160.77	86,913	2117.23	15,876	1116.70	75,664
7	3458.79	73,010	——	——	937.24	60,104
8	1085.31	60,654	——	——	919.26	48,127

Table 1: Correlations of amino acids

N'''	N''	N'	N1	N2	N3	N4	Internal	C4	C3	C2	C1	C'	C''	C'''

Table 2: Capping Positions

is modeled (for an α-helix segment) as

$$
\begin{aligned}
P(\mathbf{R}_{[S_j+1:S_j]}|\mathbf{S},\mathbf{T}) &= P(\mathbf{R}_{[S_j+1:S_j]}|S_{j-1},S_j,T_j=H) \\
&= \prod_{i=S_{j-1}+1}^{S_{j-1}+l_N^H} P^H_{N_{i-S_{j-1}}}(R_i|R_{[S_{j-1}+1:i-1]}) \\
&\times \prod_{i=S_{j-1}+l_N^H+1}^{S_j-l_C^H} P^H_I(R_i|R_{[S_{j-1}+1:i-1]}) \\
&\times \prod_{i=S_j-l_C^H+1}^{S_j} P^H_{C_{S_j-i+1}}(R_i|R_{[S_{j-1}+1:i-1]})
\end{aligned}
$$

Here the first product term represents the observation probability of amino acids at the N terminal positions of length l_N^H, the second product term represents the observation probability at the internal positions, and finally, the third product expression denotes the observation probability at the C terminal residues of length l_C^H. We also have similar expressions for the strands and the loops.

Unfortunately, at this time, the number of sequences in the PDB is not sufficient to reliably estimate the conditional probabilities given in Eq. 5. Therefore, to reduce the number of dependency parameters, amino acids were grouped into three hydrophobicity classes. The dependency patterns shown in Table 4 were discovered by our statistical analysis. N and C refer to terminal positions. $h_i \in \{\text{hydrophobic, hydrophilic, neutral}\}$ denotes the hydrophobicity class of the residue R_i at the position $i \in [1,n]$, where n is the length of the segment. For instance, the probability of observing a particular amino acid R in an α-helix segment at first terminal position depends on the hydrophobicity class of amino acids at positions $i-1$, $i+2$ and $i+4$. For a particular secondary structure segment, the segment likelihood term $P(\mathbf{R}_{[S_j+1:S_j]} \mid \mathbf{S},\mathbf{T})$ was computed by the multiplication of conditional probability terms given in Table 4 for $i = 1,..,n$.

Helix						
N1	R_i	|	h_{i-1},	h_{i+2}		
N2	R_i	|	h_{i-2},	h_{i+1}		
C1	R_i	|	h_{i-2},	h_{i-4}		
C2	R_i	|	h_{i-2},	h_{i-4}		
Int	R_i	|	h_{i-2},	h_{i-3},	h_{i-4},	h_{i+2}
Strand						
N1	R_i	|	h_{i-1},	h_{i-2}		
C1	R_i	|	h_{i-2},	h_{i-3}		
Int	R_i	|	h_{i-1},	h_{i-2},	h_{i+1},	h_{i+2}
Loop						
N1	R_i	|	h_{i-1},	h_{i-2}		
N2	R_i	|	h_{i-1},	h_{i-2}		
C1	R_i	|	h_{i-1},	h_{i-3}		
C2	R_i	|	h_{i-1},	h_{i-3}		
Int	R_i	|	h_{i-1},	h_{i-2}		

Table 4: Dependencies within segments

3.2. Computational Methods

Given an amino acid sequence $\mathbf{R}$, the vector $(m,\mathbf{S},\mathbf{T})$ that maximizes the posterior probability $P(m,\mathbf{S},\mathbf{T} \mid \mathbf{R})$ is determined as the predicted secondary structure. This could be found using a forward-backward algorithm generalized for the semi Markov HMM [5]. For each position, a posterior probability of being either an α-helix, a β-strand or a loop is computed considering all possible segmentations. The predicted state is chosen as the secondary structure state with maximum posterior probability.

4. ITERATIVE TRAINING METHOD

After predicting the secondary structure for a particular sequence, it is useful to iteratively re-adjust HMM parameters using proteins that have close secondary structure composition, and repeat the prediction step. That is, once we obtain the prediction result for a test sequence, we compute the α-helix, β-strand, and loop composition. We then remove the sequences from the training set that do not have close secondary structure content and re-estimate the HMM parameters. This is then followed by the prediction of the secondary structure using the newly estimated parameters. Although close structural composition is not a rigorous definition of structural similarity, using this measure allows us to reduce the training dataset to proteins from closely related SCOP families [6]. Therefore, a prediction of a structure from all-α class is likely to be followed by a training using proteins having high α-helix content.

χ^2	$i-5$	$i-4$	$i-3$	$i-2$	$i-1$	$i+1$	$i+2$	$i+3$	$i+4$	$i+5$
N1	1176.4	1264.1	1277.1	1324.3	2018.8	1902.9	2351.0	1591.5	2178.4	937.0
N2	907.8	1217.7	1473.0	2308.0	1902.9	2373.3	2146.9	1355.1	1645.2	1044.4
N3	1073.8	1301.1	4053.3	2105.3	1761.0	1438.5	1802.9	1458.3	1230.7	1245.5
N4	1036.0	941.8	1449.1	1866.1	1272.9	1042.4	1754.6	1071.1	1705.8	945.8
C4	923.9	1098.7	746.5	1225.3	741.2	730.3	1223.2	776.8	1342.6	850.6
C3	819.4	1198.7	995.3	1252.6	732.0	777.6	1297.7	856.0	1013.4	776.5
C2	841.0	1293.1	815.1	1150.1	755.6	636.5	1085.2	789.3	733.5	625.4
C1	785.1	1069.0	750.3	1104.6	630.1	846.0	711.1	684.9	666.0	592.7

Table 3: Position Specific Correlations in Helix Terminal Positions

5. RESULTS

In our simulations, we worked with the single-sequence set derived from the latest version of PDB [7]. Then, to match the constraints described in the paper by Schmidler *et al.* [2], we filtered out the sequences that have less than 50 and more than 900 residues. There remained about 1800 proteins.

We followed the same adjustments proposed by Frishman and Argos [8] to restrict the minimum β-strand length to 3 and minimum α-helix length to 5. Results of the cross validation experiments are provided in Table 5 [4]. From these results, we see that there is a 1.5% increase in overall 3-state prediction accuracy in comparison with BSPSS method [2]. The accuracy measure is defined as:

$$Q_3 = \sum_{i=1}^{N} \frac{\text{\# of correct predictions}}{\text{\# of amino acids}}$$

The prediction accuracies for α-helices and β-strands also increased. When the non-reduced dependency structure is used, we expect the accuracy results to be even higher.

	Q_3	Q_α	Q_β	Q_L
BSPSS	67.70	63.45	42.31	79.86
PSS-IC	69.20	67.46	43.51	79.28

Table 5: Cross Validation Results

6. ACKNOWLEDGEMENTS

YA was supported by grant CCR-0105654 from the NSF-SPS and MB was supported in part by grant H600783 from the NIH.

7. CONCLUSIONS

For a typical machine learning predictor, the basic improvement in prediction accuracy would come from developing elegant models to better capture correlations and implementation of better training methods. In this work, we performed a statistical analysis to identify the correlations between amino acids in various secondary structure segments. Then, we implemented forward-backward algorithm for a semi Markov HMM similar to the model proposed by Schmidler *et al.* [2]. We introduced a better dependency model by considering statistically significant correlations at structural segment borders. We also extended the internal dependency model that includes correlations to forward residues. In order to have better training of the model, we proposed a training method that iteratively adjusts of HMM parameters using the sequences that have close secondary structure composition. Because of the restrictions in the dataset, we reduced the dependency structure resulting from statistical analysis and obtained a 1.5% increase in the overall three-state-prediction accuracy. We believe that as more data becomes available it would be possible to implement and evaluate even higher order dependency models.

Typically protein secondary structure prediction methods suffer from low accuracy in β-strand predictions where non-local correlations have a significant role. In this work, we did not specifically address this particular problem, but showed that improvements are possible when higher order dependency models are used and significant correlations outside the segments are considered. To achieve higher improvements in prediction accuracy, one needs to develop better models that capture non-local amino acid correlations especially for β-strands.

[4] BSPSS: Bayesian Segmentation of Protein Secondary Structure
PSS-IC Protein Secondary Structure Iterations Correlations

8. REFERENCES

[1] Rost. B., "Rising accuracy of protein secondary structure prediction," http://cubic.bioc.columbia.edu/papers/2002_rev_dekker/paper.html.

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

[3] Kabsch W. Sander C., "Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features," *Biopolymers*, vol. 22, pp. 2577–2637, 1983.

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

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

[6] Chothia C. Brenner M. A. G. Hubbard S. E., T., "Scop: a structural classification of proteins database for the investigation of sequences and structures," *Journal of Molecular Biology*, vol. 247, pp. 536–540, 1995.

[7] EVA, "List of sequence-unique pdb files," http://maple.bioc.columbia.edu/eva/doc/ftp.html.

[8] Frishman D. Argos P., "Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence," *Protein Eng.*, vol. 9, no. 2, pp. 133–142, 1996.