Journal of
Data Mining in Genomics & Proteomics

Liang and Wang J Data Mining in Genom Proteomics 2011, 2:1
DOI: 10.4172/2153-0602.1000107

Research Article Open Access

Protein Secondary Structure Prediction using Deterministic Sequential Sampling

Kuo-ching Liang and Xiaodong Wang*

Department of Electrical Engineering, Columbia University, New York, NY 10027, USA

Abstract

The prediction of the secondary structure of a protein from its amino acid sequence is an important step towards the prediction of its three-dimensional structure. While many of the existing algorithms utilize the similarity and homology to proteins with known secondary structures in the Protein Data Bank, other proteins with low similarity measures require a single sequence approach to the discovery of their secondary structure. In this paper we propose an algorithm based on the deterministic sequential sampling method and hidden Markov model for the single-sequence protein secondary structure prediction. The predictions are made based on windowed observations and by the weighted average over possible conformations within the observation window. The proposed algorithm is shown to achieve better performance on real dataset compared to the existing single-sequence algorithm.

Keywords: Protein secondary structure; Single sequence prediction; Deterministic sequential sampling; Bayesian analysis

Introduction

In living organisms, numerous proteins are utilized to carry out important cellular functions. Proteins make up the physical structure of the cell to support and maintain the cell shape, and are involved in cellular signaling and transduction, where they can transmit both intra- and extra- cellular information. Other proteins act as catalyst for inert bio-molecules, to facilitate essential biochemical reactions, such as DNA repair and replication. Moreover, antibodies, or immunoglobulin, are a type of protein that is responsible for the defense of the organism against foreign substances. Proteins also play an important role in the synthesis of other proteins, forming a complex regulatory network that controls the cellular machinery [1].

A protein is basically a linear chain of amino acids folded into a three-dimensional (3D) structure. The 3D conformation of the protein plays an important role in determining the functions of the protein. For example, a protein that acts as a transcription factor is shaped so that it will only recognize a specific pattern of DNA sequence on which it will bind itself to initiate the transcription of RNA. The zinc fingers are one such family of proteins that often act as transcription factors, which recognize specific patterns of DNA sequences through the different amino acids found on the finger-like structure [2]. The interactions between antibodies and antigens have also been shown to be based on the shapes of the antibody and antigen involved, rather than their chemical properties [3]. Therefore, methods for discovering the 3D protein structures and the understanding of the structure-function relationship are instrumental to functional prediction of newly discovered proteins, and the design of novel proteins for specific tasks.

The structures of proteins are often described in four categories with increasing complexity:

- *Primary Structure*: The amino acid sequence of the protein.
- *Secondary Structure*: Local folding patterns that are formed due to hydrogen bonding between the N-H and C=O groups. Commonly observed patterns include α helix and β sheet.
- *Tertiary Structure*: Global folding structure of the amino acid chain. Each protein typically has a native conformation, which is the structure that the protein is typically found in, but can also take other folding structures depending on the condition. Many different forces come together to influence the tertiary structure of the protein, but one important observation is that the hydrophobic parts of the protein is typically hidden in the core of the secondary structure.
- *Quaternary Structure*: A superstructure formed by the interaction of multiple proteins.

Typically, the 3D shape of the protein can be determined by X-ray crystallography, and more recently, nuclear magnetic resonance (NMR) spectroscopy, which also determines the local secondary structures manifested within the 3D structure [4]. However, both methods have proven to be expensive and time-consuming to perform. Therefore, efficient methods for preliminary prediction of the protein structure are needed before either X-ray crystallography or NMR is used. The shape of a protein, as it turns out, is determined by its unique amino acid sequence, as can be seen from the denatured protein sequence quickly returning to its native state when the denaturing agents have been removed. Thus, constructing computational models based on the amino acid sequences of proteins to predict their 3D structures, and furthermore, their functions, has become an important subject of research. Additionally, it is well known that protein molecules can take on multiple secondary and tertiary structures under different conditions, rather than a single structure [5,6]. Therefore, it is also of interest to discover not only the most likely conformation, but also a set of possible conformations for the given amino acid sequence [7,8].

Existing works

However, even when considering only the most likely 3D structure of a protein, the problem of predicting it from the amino acid sequence

***Corresponding author:** Xiaodong Wang, Department of Electrical Engineering, Columbia University, New York, NY 10027, USA, E-mail: wangxg@ee.columbia.edu

Received January 11, 2011; **Accepted** February 19, 2011; **Published** February 22, 2011

Citation: Liang K, Wang X (2011) Protein Secondary Structure Prediction using Deterministic Sequential Sampling. J Data Mining in Genom Proteomics 2:107. doi:10.4172/2153-0602.1000107

is still a difficult one. Proteins are complex molecules that contain hundreds of thousands of atoms, and many different forces come together to determine the final shapes of proteins. Methods such as molecular simulation would also prove to be time consuming, and are typically limited to proteins of certain amino acid length, even when simulated on super computers. One approach is by divide-and-conquer, where the local problem of secondary structure prediction is tackled first, and then using the predicted secondary structure information to assist in the prediction of the tertiary structure [9]. In [10], one of the earliest secondary structure prediction algorithms was proposed. The Chou-Fasman Algorithm was empirically-based, utilizing the observed frequencies of occurrence of the different amino acids for the different types of secondary structures. The probabilities are then used to evaluate a given amino acid being favoring, breaking, or indifferent to the different secondary structures. Other algorithms such as [11,12] also employ point statistics for prediction, mostly due to the limited amount of known secondary structure data.

The second generation of secondary structure prediction methods was able to achieve greater accuracy by drawing from a larger set of proteins in the databases with known secondary structures, discovered through X-ray crystallography and NMR spectroscopy. Moreover, these algorithms make use of segment statistics, where a window of amino acids is used to make the prediction, instead of only point statistics as in the first generation algorithms. In [13], the secondary structure of an amino acid is predicted by computing the mutual information between the secondary structure, and a window of amino acids flanking the given amino acid. Other second generation algorithms can be found based on statistical models [14,15], sequence patterns [16,17], neural networks [18,19], and graph theory [20].

Currently, there are two main categories of protein secondary structure prediction methods: multiple-sequence and single-sequence methods. Multiple-sequence methods make use of multiple sequence alignment with homologous sequences having known secondary structures found in the databases. These methods are based on the notion that if the amino acid sequences are close evolutionarily, their secondary structures should also bear close similarities [21]. The *PHD* algorithm [22], first introduced this idea, and is still by far one of the most accurate algorithms available.

The *PHD* algorithm first performs a database search for possible homologous proteins, then aligns and filters the sequences to decide on the most likely homologues, and finally feeds the sequences and alignment profile to a feed-forward neural network for secondary structure prediction. Other multiple-sequence algorithms include [23-25].

While the multiple-sequence approach to protein secondary structure prediction has achieved the highest prediction accuracy, not all proteins are suitable for this method. Even with the number of proteins with known secondary structure steadily increasing, there are still many proteins with no close homologues in the databases. Although one can simply choose the closest possible proteins in the database, there is no guarantee that any meaningful result can be obtained. In this case, using the single-sequence approach becomes necessary. Some recent single-sequence methods include *BSPSS* [26] and *IPSSP* [27]. In [28], the N-best algorithm is used to obtain a set of top scoring secondary structure predictions. In each of the above algorithms, the lengths of the secondary structure segments are also considered. The secondary structure segment, or simply segment, in [26-28] and in this paper is defined as a sequence of consecutive secondary structures of the same type. The conformation of a protein can be broken down into many segments. The hidden semi-Markov model is used to model the uncertainty related to the length of each segment.

In the existing approaches, the amino acid sequence is processed sequentially from left to right, predicting whether the secondary structure of the current amino acid extends the most recent segment, or begins a new secondary structure of a different type. In these approaches, the score of a conformation up to the current amino acid position is given by the combinations of secondary structure segments and their length in this particular conformation. The contribution of the current amino acid to the scores of the different possible conformations generated will depend on the length and type of the final segment, i.e., the current observed amino acid is assumed to be the ending position or the starting position of the final segment. However, this assumption is not entirely correct, since the secondary structure of the current amino acid can also be located within a longer secondary structure segment, and the former assumption can cause a possible conformation to be prematurely discarded due to a mismatched segment length. In this work, we use a deterministic sequential sampling approach to obtain a set of possible conformations. The corresponding secondary structure of an amino acid is determined by a window of flanking amino acids. Furthermore, when processing the current amino acid, instead of terminating the latest segment at the current position, we will enumerate all possible conformations within the given window, and take a weighted average to determine the most likely secondary structure assignment. We will show through numerical experiments that the proposed approach can obtain better prediction accuracy on a set of proteins chosen based on their low similarity to one another.

The remainder of the paper is organized as follows. In Section 2, we present the signal model for the single-sequence secondary structure prediction problem. In Section 3, we derive the proposed sequential sampling algorithm for solving the problem. In Section 4, we present numerical results using proteins with low homology scores to known proteins in the databases. Section 5 concludes the paper.

Signal Model

Let us denote the amino acid sequence of length T as $r \triangleq \{r_1, r_2, \ldots r_T\}$ g, where r_t is the t-th amino acid in the sequence, and takes value from the 20 proteinogenic amino acids used to construct proteins. Let the sequence of secondary structure types associated with the amino acid sequence be denoted as $s \triangleq \{s_1, s_2, \ldots s_T\}$, where s_t is the secondary structure type of the amino acid r_t.

The types of secondary structures are defined by the hydrogen bonds formed between the amino acids [1]. According to the Dictionary of Protein Secondary Structure (DSSP), the major secondary structure types include: G (3-turn helix), H (4-turn helix), I (5-turn helix), E (beta sheet), B (beta bridge), and S (bend). For simplicity of modeling, these structures are often grouped together to form larger class assignments. In our work, we take the convention of 3-class assignments, i.e., H (α-helix), E (β-strand), and L (loop), i.e., $s_t \in \{H,E,L\}$.

Due to the way the atoms are bonded in a given secondary structure, each of the three structures also appear in an amino acid sequence with minimum consecutive length requirements. For an ®-helix segment, the minimum length is 5 consecutive amino acids; for a β-strand segment, the minimum length is 3; and for a loop, the minimum length is 1. For example, the secondary structure sequence $s = \{H,H,H,H,H,H,L,L,E,E,E\}$ is a valid secondary structure with 3 segments (1 segment of H with length 6, 1 segment of L with length 2, and 1 segment of E with length 3) for a protein of 11 amino acids long;

whereas $s = \{H,H,H,H,L,L,L,L,E,E,E\}$ is not a valid sequence since the first H segment is only of length 4.

In this work, we model the relationship between the amino acid sequence and the corresponding sequence of secondary structure assignments using a state space model. In particular, the state sequence is a sequence of secondary structures, and the corresponding observations are windows of amino acids taken from the given amino acid sequence. Given the known amino acid sequence r, to predict the corresponding secondary structure sequence s, we would like to compute the probability $p(s|Y)$, where $Y \triangleq \{\overline{y}_1, \overline{y}_2, \cdots, \overline{y}_T\}$, and $\overline{y}_t$ is a window of amino acids flanking the t-th amino acid, r_t. For a chosen window length of $2w + 1$, the observation for the state s_t is $\overline{y}_t = \{r_{t-w}, r_{t-w}+1, \cdots, r_{t-w}-1, r_{t+w}\}$.

As we have discussed previously, to predict the secondary structure of an amino acid at each step, instead of assuming that the latest secondary structure segment with the current amino acid, we would like to consider all the possible segment lengths for the latest secondary structure segment, and all possible subsequent configurations that make up the rest of the window. For example, suppose that we have a window of amino acids $\overline{y}_7 = \{r_4, r_5, \cdots, r_{10}\}$, based on which we are trying to predict the secondary structure type for the 7th amino acid in r, or, the 4th amino acid within this window, and that we have already predicted the secondary structures of the first 6 amino acids, s_1 to s_6. Suppose further that all of the first 6 amino acids have the β-strand structure, and that we want to find the probability of s_7 also being E given the observation $\overline{y}_7$. In both *BSPSS* [26] and *IPSSP* [27], this predicted segment of β-strands would be assumed to have a total length of 7, and the distribution of amino acids in a β -strand segment of length 7 would be used to compute the score of this particular extension. However, the length of this segment of β -strand could in fact be longer than 7 amino acids long, and the distribution of amino acids in a β -strand segment of length 7 is different from the distribution of amino acids in β -strand segment of any other lengths. Furthermore, since the window in this case extends past the current amino acid by 3 positions, and the β -strand has already reached its required minimum length, a segment of different secondary structure type can also begin in s_8, s_9, or s_{10}. To account for all of these different possibilities, we would like to take a weighted average over the probabilities of all the allowed configurations of secondary structures in the window.

Let us denote $s_u \triangleq \{s_1, s_2, \cdots, s_u\}$ and $s_{u:v} \triangleq \{s_u, s_u+1, \cdots, s_v\}$. Figure 1 shows some possible configurations in $s_{8:10}$ where $2w + 1 = 7$ and $s_6 = \{E,E,E,E,E,E\}$. In Figure 1 (a)-(c), three possible configurations of $s_{8:10}$ are shown for the case of $s_7 = E$. In Figure 1(a), the β-strand segment extends past the range of the window until s_{11}, therefore, we have $s_{8:10} = \{E,E,E\}$. For Figure 1(b), we also have $s_{8:10} = \{E,E,E\}$. However, as we can see from the possible conformations outside of the current window, the β-strand segment in (b) is longer than the segment in (a) by 1 amino acid. When computing the scores for these two different possible conformations, we would like to use the amino acid distribution in a β-strand segment of length 11 for (a), and β-strand segment of length 12 for (b); whereas in previous works, both of these possibilities would use the segment length of 11 when computing the scores. To differentiate between (a) and (b), we need indicator variables to denote the location of a secondary structure within its secondary structure segment, and the length of that segment, and we denote these as ρ_i and μ_i, respectively. Thus, for Figure 1(a), we have $\rho_{10} = 10$ and $\mu_{10} = 11$, and for Figure 1(b), we have $\rho_{10} = 10$ and $\mu_{10} = 12$.

In Figure 1(c), the β-strand segment terminates at $s8$, and is followed by an α-helix segment of length 5, which extends beyond the boundary of the current window. For this example, $s_{8:10} = \{E,H,H\}$, $\rho_{10} = 2$ and $\mu_{10} = 5$. Note from these examples, we do not need a length indicator for the segment that s_{t-w} belongs to, since we have already estimated s_{t-1}. If its segment terminates within $s_{t-w:t+w}$, that information is contained in $\rho_t - w$ and $s_{t-w:t+w}$. If the segment terminates outside of $s_{t-w:t+w}$, then the segment length is given by μ_{t+w}. Also note that since there is no restriction on the length of a segment other than that it should be less than the length of the given amino acid sequence, we typically upper bound it with the maximum length of a segment observed in the training dataset.

Inference problem

From the discussions above, we can summarize the single-sequence protein secondary structure prediction problem as follows. Given the observations $Y \triangleq \{\overline{y}_1, \overline{y}_2, \cdots, \overline{y}_T\}$, which are windows of amino acids obtained from the amino acid sequence of a protein, $r \triangleq \{r_1, r_2, \cdots, r_T\}$, we would like to predict the secondary structure sequence $s \triangleq \{s_1, s_2, \cdots, s_T\}$, of the protein. The proposed algorithm will obtain a set of highest scoring solutions. These solutions can either be used to reach a consensus on the secondary structure, or be used to represent the possible conformations a protein can take under different situations. In the next section, we will derive the algorithm to solve the protein secondary structure prediction problem.

Protein Secondary Structure Prediction Algorithm

In this section, we first give a brief overview of the deterministic sequential sampling method. We then derive the protein secondary structure prediction algorithm to obtain a set of conformations with high scores.

Deterministic sequential sampling

Let us consider the following dynamic model

initial state model: $p(s_1)$, (1)

state transition model: $p(s_t \mid s_t - 1)$, $\forall t \geq 1$, (2)

measurement model: $p(\overline{y}_t \mid s_t)$, $\forall t \geq 1$, (3)

where s_t and $\overline{y}_t$ are the state and the observation at time t, respectively. At time t, we want to make an online inference of the states $s_t = (s_1, \cdots, s_t)$ based on the observation $Y_t = (\overline{y}_1, \cdots, \overline{y}_t)$. Similar to traditional sequential Monte Carlo (SMC) methods [29], we assume that we have at time $t - 1$ a set of particles and their associated weights $\{(s_{t-1}^{(k)}, w_{t-1}^{(k)}), k = 1, \cdots, K\}$ properly weighted with respect to the posterior distribution $p(s_{t-1} \mid Y_{t-1})$. In the protein secondary structure prediction problem, at each time step, the possible secondary structure that the current amino acid can take is limited to a finite set. Specifically, the secondary structure of the current amino acid is from

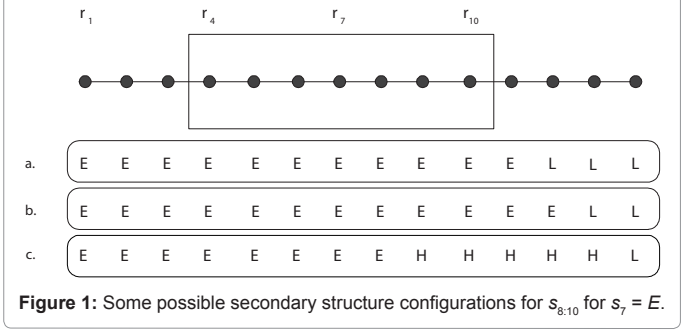

Figure 1: Some possible secondary structure configurations for $s_{8:10}$ for $s_7 = E$.

the set {*H,E,L* }if the most recent segment has satisfied the minimum length requirement, and it is the same type of secondary structure as the most recent segment if the minimum length is not met. Since the set of possible secondary structure is finite, this naturally leads us to consider the deterministic approach where we enumerate all possible extensions for each sample at step *t*.

The deterministic sampling approach was developed in [30,31], and extended to handle Markov state processes. For each particle $s_{t-1}^{(k)}$, k = 1, ⋯, K, we consider all K_{ext} possible extensions, and perform a selection step to keep only K of the $K \times K_{ext}$ particles to avoid an exponential growth of the number of particles. Also, in this context there is no reason to keep particles that represent the same path, thus in this regard; the deterministic sequential sampling method is also different from the traditional resampling scheme. There exist various selection schemes, and in this work, we adopt the simple scheme of keeping only the K particles with the highest weights while discarding the remaining particles.

Given a set of particles and associated weights $\{(s_{t-1}^{(k)}, w_{t-1}^{(k)}), k = 1, \cdots, K\}$ that does not contain duplicate paths; we can obtain the posterior distribution of s_{t-1} as the following,

$$\hat{p}\left(s_{t-1}\middle|Y_{t-1}\right) = \frac{1}{W_{t-1}} \sum_{k=1}^{k} w_{t-1}^{(k)} \mathrm{II}\left(s_{t-1} - s_{t-1}^{(k)}\right), \tag{4}$$

where $W_{t-1} = \sum_{k=1}^{K} W_{t-1}^{(k)}$, and II (.) is the indicator function such that II(x) = 1 for x=0 and II (x)=0 otherwise. From Bayes' theorem we have

$$p\left(s_t\middle|Y_t\right) \alpha p(\overline{y_t}\middle|s_t, Y_{t-1})\, p(s_t\middle|Y_{t-1})$$
$$\alpha p\left(\overline{y}_t\middle|s_t, Y_{t-1}\right) p(\overline{y}\middle|s_{t-1}, Y_{t-1})\, p(s_{t-1}\middle|Y_{t-1}). \tag{5}$$

From this relationship, we can approximate the posterior distribution of s_t as

$$\hat{p}^{ext}(s_t\middle|Y_t) = \frac{1}{W_t^{ext}} \sum_{k=1}^{K} \sum_{i=1}^{K_{ext}} w_t^{(k,t)} \mathrm{II}\left(s_t - [s_{t-1}^{(k)}, \theta_i]\right), \tag{6}$$

where $\left[s_{t-1}^{(k)}, \theta_i\right]$ represents the vector obtained by appending the element θ_i to the vector $s_{t-1}^{(k)}$, and

$W_t^{ext} = \sum_{k,i} w_t^{(k,i)}$ with

$$w_t^{(k,i)} \alpha w_{t-1}^{(k)} p(\overline{y}_t\middle|s_t = \theta_i, s_{t-1}^{k}, Y_{t-1}) p(s_t = \theta_i\middle|s_{t-1}^{k}, Y_{t-1}). \tag{7}$$

Note that during the initialization steps, if the total number of particles obtained after enumerating all possible extensions from all particles from $t-1$ is less than the maximum number allowed K, all enumerated particles are retained with weights computed as stated above.

Deterministic sequential sampling protein secondary structure prediction algorithm

For system states up to the t-th amino acid, s_t, corresponding observations Y_t, where $Y_t \triangleq \{\overline{y}_1, \overline{y}_2, \cdots, \overline{y}_t\}$, we have the following according to (5)

$$p(s_t\middle|Y_t) \;\alpha\; p(\overline{y}_t\middle|s_t, Y_{t-1}) \;p\; (s_t\middle|Y_{t-1})$$
$$\alpha\; p(\overline{y}_t\middle|s_t, Y_{t-1}) \;p\; (s_t\middle|s_{t-1}, Y_{t-1})\, p(s_{t-1}\middle|Y_{t-1})$$
$$\alpha\; p(\overline{y}_t\middle|s_t)\, p(s_t\middle|s_{t-1}, \overline{y}_{t-1})\, p(s_{t-1}\middle|Y_{t-1}). \tag{8}$$

Thus at each step, the recursion in (8) involves the computation of the probabilities $p(\overline{y}_t\middle|s_t)$ and $p(s_t\middle|s_{t-1}, \overline{y}_{t-1})$.

The distribution

$$p(\overline{y}_t\middle|s_{t-1}, s_t) = \sum_{(s_{t+1:t+w}, \rho_{t+w}, \mu_{t+w}) \in \Lambda} p(\overline{y}_t\middle|s_{t-w:t}, s_{t+1:t+w}, \rho_{t-w}, \rho_{t+w}, \mu_{t+w}) \times$$
$$p(s_{t+1:t+w}, \rho_{t+w}, \mu_{t+w}\middle|s_{t-w:t}, \rho_{t-w})$$
$$= \sum_{(s_{t+1:t+w}, \rho_{t+w}, \mu_{t+w}) \in \Lambda} p(s_{t+1:t+w}, \rho_{t+w}, \mu_{t+w}\middle|s_{t-w:t}, \rho_{t-w}) \times$$
$$\prod_{i=t-w}^{t+w} p_{s_i, \rho_i \mu_i}(r_i), \tag{9}$$

where Λ is the set of all allowed conformations of length w in the right-half of the window given the left-half window $s_{t-w:t}$ and ρ_{t-w}, and $p_{s_i, \rho_i, \mu_i}(r_i)$ is the frequency of observing the amino acid r_i at position $\rho_i$in an s_i type secondary structure segment of length μ_i.

Similar to the approaches taken in [26] and [28], the distribution of the amino acids are computed by training from a set of proteins with known secondary structure, which are chosen to have low sequence similarities from one another. The segments in these proteins are grouped according to their secondary structure type and the segment length to compute the amino acid frequencies. The probability $p(\overline{y}\middle|s_{t-w:t}, s_{t+1:t+w}, \rho_{t-w}, \rho_{t+w}, \mu_{t+w})$ is then constructed by selecting the appropriate columns according to the parameters $s_{t-w:t}, s_{t+1:t+w}$, and μ_{t+w}. The transition probability $p(s_{t+1:t+w}, \rho_{t+w}, \mu_{t+w}\middle|s_{t-w:t}, \rho_{t-w})$ can be computed similarly, by shifting a window of size $2w+1$ acrossthe proteins in the training dataset, and counting the number of occurrences of the different secondary structure transitions.

To evaluate $p(s_t\middle|s_{t-1}, \overline{y}_{t-1})$, we can use the Bayes' rule to obtain the following relationship:

$$p(s_t\middle|s_{t-1}, \overline{y}_{t-1}) = \frac{p(s_t, \overline{y}_{t-1}\middle|s_{t-1})}{p(\overline{y}_{t-1})}$$
$$\alpha\; p(\overline{y}_{t-1}\middle|s_t, s_{t-1}) p(s_t\middle|s_{t-1}). \tag{10}$$

The distribution $p(\overline{y}_{t-1}\middle|s_t, s_{t-1})$ can be computed similarly to (9), with the exception that it is averaged over possible right-half conformations of length $w-1$, instead of w, and is given as follows:

$$p(\overline{y}_{t-1}\middle|s_t, s_{t-1}) = \sum_{(s_{t+1:t+w-1}, \rho_{t+w}, \mu_{t+w-1}) \in \Lambda}$$
$$p(s_{t+1:t+w-1}, \rho_{t+w-1}, \mu_{t+w-1}\middle|s_{t-w-1:t}, \rho_{t-w-1}) \times$$
$$\prod_{i=t-w-1}^{t+w-1} p_{s_i, p_i \mu_i}(r_i), \tag{11}$$

For $p(s_t\middle|s_{t-1})$, and window length of $2w+1$, the distribution can be evaluated by the $w-$th order Markov chain

$$p(s_t\middle|s_{t-1}) = p(s_t\middle|s_{t-w:t-1}). \tag{12}$$

Once again, this distribution can be evaluated using the counting process from the training dataset.

Deterministic Sequential Sampling Estimator: We will now outline the deterministic sequential sampling algorithm for protein secondary structure prediction. Suppose at time t, we have a set of

weighted samples $\left\{\left(s_{t-1}^{(k)}, w_{t-1}^{(k)}\right), k=1,\cdots,K\right\}$, properly weighted with respect to $p(s_{t-1}|Y_{t-1})$, then, as in (4), $p(s_t|Y_t)$ can be approximated by

$$\hat{p}^{ext}(s_t|Y_t)=\frac{1}{W_t^{ext}}\sum_{k=1}^{K}\sum_{\theta^{(k)}\in\Theta^{(k)}} w_t^{(k,\theta)}\mathrm{II}(s_t-[s_{t-1}^{(k)},\theta^{(k)}]), \tag{13}$$

Where $\Theta^{(k)}$ is the set of all possible secondary structures that the amino acid, r_t, can assume. If the latest segment of secondary structure has exceeded the minimum length required, then $\Theta^{(k)}=\{H,E,L\}$. If the minimum length has not been reached, $\Theta^{(k)}$ can take only the same secondary structure type as the latest segment. The weight update formula is then given by

$$w_t^{(k,\theta)} \alpha\, w_{t-1}^{(k)} p(\overline{y}_t|s_{t-1}^{(k)}, s_t=\theta^{(k)})\, p(s_t=\theta^{(k)}|s_{t-1}^{(k)}, \overline{y}_{t-1}), \tag{14}$$

Where $p(\overline{y}_t|s_{t-1}^{(k)}, s_t=\theta^{(k)})$ is evaluated from (9), and $p(s_t=\theta^{(k)}|s_{t-1}^{(k)}, \overline{y}_{t-1})$ from (10). In the selection step, only the K particles with the highest weights are retained to avoid the exponential explosion of particles.

We now give the deterministic sequential sampling protein secondary structure inference algorithm for predicting a set of K most likely conformations and their scores:

Algorithm 1 [Deterministic sequential sampling protein secondary structure prediction algorithm]

- *Initialization: Use the first γ amino acids to enumerate all possible particles, where γ is the largest number such that the total number of particles enumerated from the γ amino acids does not exceed K, and compute their weights.*
- *Update: For $t=\gamma+1,\gamma+2,\cdots$*
  - *For $k=1,2,\cdots$*

**Find the set $\Theta^{(k)}$, the possible secondary structure extensions for particle k.*

**Enumerate all possible particle extensions, $s_t^{(k,\theta^{(k)})}=[s_{t-1}^{(k)},\theta^{(k)}], \theta^{(k)}\in\Theta^{(k)}$.*

** $\forall\,\theta^{(k)}\in\Theta^{(k)}$, compute the weights $w_t^{(k,\theta^{(k)})}$ according to* (14).

– Select and preserve K distinct sample streams $\{s_t^{(k)}, k=1,\cdots,K\}$ with the highest importance weights $\{w_t^{(k)}, k=1,\cdots,K\}$ from the set $\{s_t^{(k,\theta^{(k)})}, w_t^{k,\theta^{(k)}}, k=1,\cdots,K\theta^{(k)}\in\Theta^{(k)}\}$.

Viterbi algorithm for most probable conformation

In the previous section, we have presented the deterministic sequential sampling algorithm to obtain a set of suboptimal secondary structure conformations with high scores to describe the possible conformations that a protein can take under various circumstances. However, typically, we are also interested in the most likely secondary structure given the amino acid sequence. As can be seen from Section 2, the system model is basically on of a hidden Markov model (HMM), and the most likely secondary structure is simply the most probable state sequence of the state model given the amino acid sequence r. For an HMM with no unknown parameters, the optimal path can be obtained through the Viterbi algorithm. Therefore, we would like to consider the optimal path found by using the Viterbi algorithm, and compare it with those found using the deterministic sequential sampling algorithm with weighted majority voting. For each secondary structure type s_t at position t, using (8), the Viterbi algorithm computes the following:

$$\begin{aligned} f_{t-w:t-1,t}(s_{t-w:t-1}, s_t) &= \delta(s_{t-w-1:t-1})\, p(\overline{y}_t|s_{t-w:t})\, p(s_t|s_{t-w:t-1}, \overline{y}_{t-1}) \\ \delta(s_{t-w:t-1}) &= \max_{t-w:t-1} f_{t-w:t}(s_{t-w:t-1}, s_t). \end{aligned} \tag{15}$$

Experimental Results

We have implemented the proposed protein secondary structure prediction algorithm and evaluated its performance on real data. The real dataset we used in our experiment is a set of proteins with low sequence homology chosen from the Protein Data Bank (PDB). The proteins in this dataset are chosen so that the length-dependent threshold between any pair of proteins does not exceed the length-dependent threshold as given in [32], and the dataset consists of a total of 2810 proteins.

In the experiments, similar to what is done in [28], we first removed from the dataset proteins that have lengths shorter than 35 amino acids long. Furthermore, we removed proteins that contain secondary structure segments over 35 amino acids long. Such pruning is to limit the number of possible conformations in $s_{t+w-1:t+w}$ that we have to average over when evaluating (9). The low frequencies in which these segments occur also make the accurate estimation of their distribution difficult. The resulting size of the dataset after these filters have been applied is 2661.

Since the secondary structures for these proteins obtained from PDB come in 8 DSSP secondary structure types, we follow the "CK" mapping used in [28] and [33]. In the "CK" mapping, *H* is mapped to *H*, *E* is mapped to *E*, and all other secondary structure types are mapped to *L*. Also *H* segments shorter than 5 and *E* segments shorter than 3 amino acids are mapped to *L* as well.

The metric used to evaluate the performance is the three-state-per-residue accuracy, or the Q_3 metric [34]. The Q_3 metric is computed by dividing the total number of correctly predicted secondary structure in the dataset, by the total number of amino acids present in the dataset, i.e.,

$$Q_3 \overset{\Delta}{=} \frac{\sum_{l=1}^{L}\psi_l}{\sum_{l=1}^{L}N_l}, \tag{16}$$

where N_l is the number of amino acids in the l-th protein, and ψ_l is the number of correctly predicted secondary structure in the l-th protein.

To evaluate the performance of the algorithms, we use the leave-one-out approach. Each time we will remove one protein from the dataset, and the remaining proteins are used to compute the distributions used in (9) and (10). Using the estimated parameters in the deterministic sequential sampling algorithm, we predict the secondary structures of the protein that was left out, and compare the result with the known secondary structures obtained from PDB. This process is repeated for all other proteins in the dataset, and the aggregate results are used to compute the Q_3measure.

Performance results on PDB dataset: The leave-one-out analysis is performed on the PDB dataset using the deterministic sequential sampling algorithm proposed in this paper, and the modified stack decoder proposed in [28]. For the deterministic sequential sampling algorithm, the window size used was $2w+1=11$, and for both algorithms a total of $N=500$ particles/size of N list is kept at each iteration, and the highest scoring $M=100$ candidates are combined using weighted majority voting with their scores to obtain a single secondary structure prediction for each algorithm. The weighted majority voting is done

	Q_3(%)	Q_α(%)	Q_β(%)	Q_L(%)
Sequential sampling: highest score	58.00	60.63	25.01	70.48
Sequential sampling: weighted majority voting	58.57	61.25	25.33	71.12
Modified Stack Decoder	58.35	62.15	24.15	70.51
Viterbi Algorithm	57.65	60.18	26.34	69.46

Table 1: Overall $Q_3$and individual secondary structure accuracy comparisons.

Window size ($2w$ + 1)	Q_3 (%)	Q_α (%)	Q_β (%)	Q_L (%)
7	53.97	50.48	23.41	69.82
11	58.57	61.25	25.33	71.12
15	57.92	61.15	24.58	70.11

Table 2: Overall Q_3 and individual secondary structure accuracy comparisons for different window sizes.

N	M	Q_3 (%)	Q_α (%)	Q_β (%)	Q_L (%)
500	100	58.57	61.25	25.33	71.12
750	150	59.33	61.57	25.42	72.48
1000	200	60.09	62.37	26.22	73.21

Table 3: Overall Q_3 and individual secondary structure accuracy comparisons for different number of particles and different number of top candidates used in weighted majority voting.

at each amino acid location, by adding up the weights for the same secondary structure type, and choosing the secondary structure type with the highest total weight as the consensus. The Q_3 measures are shown in Table 1, with Q_α, Q_β, and Q_L being the individual secondary structure type prediction accuracy for α- helix, β- strand, and loop, respectively.

As we can see from Table 1, both the deterministic sequential sampling algorithm and the modified stack decoder were able to achieve better performance through weighted majority voting, than by simply considering the candidate conformation with the highest weight. This indicates that the most probable path does not necessarily result in the most accurate secondary structure prediction. These predictions are made based on algorithms which are trained using datasets that consist of proteins with low similarity scores. Since these training captures the overall averaged statistical behavior, profiles of individual proteins may still deviate greatly from these averaged distributions. As these experiments show, averaging over a set of generated secondary structures can improve the actual performance compared to the most likely structure.

Moreover, the deterministic sequential sampling algorithm is shown to have better performance than that of the modified stack decoder. Upon closer inspection, we can see that both algorithms generated different lists of likely secondary structures. During the secondary structure extension process, while computing the scores for each possible extension, the proposed algorithm does not terminate the latest segment with the current amino acid. This helped the deterministic sequential sampling algorithm to keep more of the better candidates during the pruning process. Also, while the deterministic sequential sampling algorithm has poorer prediction accuracy for β-strands, the overall accuracy is compensated by the higher accuracy for loops, which accounts for about 46% of the total secondary structures.

Window size and particle number: In the following experiments, we compare the prediction accuracy of the proposed deterministic sequential sampling algorithm under different window sizes and different particle sizes. In Table 2, we compared the Q_3 performance and individual secondary structure type accuracy for window sizes $2w+1 = 7$, $2w+1 = 11$, and $2w+1 = 15$. For each case we use $N = 500$ particles, and the final $M = 100$ particles with the highest weights are used to obtain the consensus secondary structure through weighted majority voting.

From Table 2, we can see that the deterministic sequential sampling algorithm with window size $2w+1 = 11$ achieves the best performance, while at $2w+1 = 7$ the performance is much worse than the other two window sizes. The poor prediction performance of the smaller window size is due to the the inadequately captured statistical features of the secondary structure by the small window size. However, larger window sizes also pose problems as shown by the worse performance with window size of 15 compared to window size of 11. This is caused by the inaccurate predictions of the two state transition probabilities in (9) and (11). When large window sizes are chosen, the number of unique transitions also increases, which can lead to many transitions without sufficient number of observations, thus resulting in poor estimation of the transition probabilities.

Next, we perform three experiments; each has window length of $2w+1 = 11$, but using different total number of particles and the number of top candidates used for weighted majority voting. The results in Table 3 show that the performance of the deterministic sequential sampling algorithm can be improved by increasing the number of particles. However, the improvement does not seem to be very significant, as the overall Q_3performance is only improved by 2% when the number of particles was doubled from 500 to 1000, showing that the deterministic sequential sampling algorithm is still robust when the number of particles used in the experiment is relatively small.

Conclusions

In this work, we have proposed a single-sequence, deterministic sequential sampling-based algorithm to find the most likely secondary structure conformation of a protein, and a set of suboptimal conformations to simulate the changing protein structure under different environments. The algorithm is based on a windowed-observation hidden Markov model. While the enumeration of the states may seem more complex, we have improved upon the way each conformation is scored, by taking an average over the possible conformations within a window. One should take note that in our work, we have not included the effects from the in sequences of non-local amino acids in the sequence. To model the non-local in sequences, more available data than what is presently available is needed in order to accurately describe the interactions between amino acids that are far apart in terms of sequence distance. With more available data, a joint secondary and tertiary structure prediction may become a more feasible approach to the prediction of protein structure, and the HMM-based approach presented in this work should lend itself very well with extensions to treat additional tertiary structure information.

References

1. Alberts B (2003) Essential Cell Biology. New York, NY: Garland Science.
2. Nakata K (1995) Prediction of zinc finger DNA binding protein. Bioinformatics 11: 125-131.
3. Pauling L, Campbell DH, Pressman D (1943) The nature of the forces between antigen and antibody and of the precipitation reaction. Physiol Reviews 23: 203-219.
4. Barnwal RP, Chary KVR, (2008) An efficient method for secondary structure determination in polypeptides by NMR. Current Science 94: 1302-1306.
5. Meiler J, Baker D (2003) Coupled prediction of protein secondary and tertiary structure. Proc Natl Acad Sci U S A 100: 12105-12110.
6. Lindorff-Larsen K, Best RB, Depristo MA, Dobson CM, Vendruscolo M (2005) Simultaneous determination of protein structure and dynamics. Nature 433: 128-132.

7. Furnham N, Blundell TL, DePristo MA, Terwilliger TC (2006) Correspondence: is one solution good enough. Nature Struct and Mol Biol 13: 184-185.
8. Rother D, Sapiro G, Pande V (2008) Statistical characterization of protein ensembles. IEEE/ACM Trans Comput Biol Bioinf 5: 42-55.
9. Petersen TN, Lundegaard C, Nielsen M, Bohr H, Bohr J (2000) Prediction of protein secondary structure at 80% accuracy. Proteins 41: 17-20.
10. Chou PY, Fasman GD (1974) Prediction of protein conformation. Biochem 13: 251-276.
11. Lim VI (1974) Structural principles of the globular organization of protein chains: A stereo-chemical theory of globular protein secondary structure. J Mol Biol 88: 857-872.
12. Rose GD (1978) Prediction of chain turns in globular proteins on a hydrophobic basis. Nature 272: 586-590.
13. Garnier J, Osguthorpe DJ, Robson B (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 120: 97-120.
14. Maxfield FR, Scheraga HA (1976) Status of empirical methods for the prediction of protein backbone topography. Biochem 15: 5138-5153.
15. Robson B, Suzuki E (1976) Conformational properties of amino acid residues in globular proteins. J Mol Biol 107: 327-356.
16. Taylor WR, Thornton JM (1983) Prediction of super-secondary structure in proteins. Nature 301: 540-542.
17. Rooman MJ, Kocher JP, Wodak SJ (1991) Prediction of protein backbone conformation based on seven structure assignments: influence of local interactions. J Mol Biol 221: 961-979.
18. Kneller DG, Cohen FE, Langridge R (1990) Improvements in protein secondary structure prediction by an enhanced neural network. J Mol Biol 214: 171-182.
19. Zhang X, Mesirov JP, Waltz DL (1992) Hybrid system for protein secondary structure Prediction. J Mol Biol 225: 1049-1063.
20. Mitchell EM, Artymiuk PJ, Rice DW, Willett P (1992) Use of techniques derived from graph theory to compare secondary structure motifs in proteins. J Mol Biol 212: 151-166.
21. Goldman N, Thorne JL, Jones DT (1996) Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. J Mol Biol 263: 196-208.
22. Rost B, Sander C, Schneider R (1994) PHD: an automatic mail server for protein secondary structure prediction. Comput Appl Biosci 10: 53-60.
23. Levin JM, Pascarella S, Argos P, Garnier J (1993) Quantification of secondary structure prediction improvement using multiple alignment. Prot Engin 6: 849-854.
24. King RD, Sternberg MJ (1996) Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. Prot Sci 5: 2298-2310.
25. Salamov AA, Solovyev VV (1997) Protein secondary structure prediction using local alignments. J Mol Biol 268: 31-36.
26. Schmidler SC, Liu JS, Brutlag DL (2000) Bayesian segmentation of protein secondary structure. J Comp Biol 7: 233-248.
27. Aydin Z, Altunbasak Y, Borodovsky M (2006) Protein secondary structure prediction for a single sequence using hidden semi-Markov models. BMC Bioinform 7: 178.
28. Aydin Z, Altunbasak Y, Erdogan H (2007) Bayesian protein secondary structure prediction with near-optimal segmentations. IEEE Trans Sig Proc 7: 3512-3525.
29. Doucet A, Wang X, (2005) Monte Carlo methods for signal processing: a review in the statistical signal processing context. IEEE Sig Proc Mag 22: 152-170.
30. Fearnhead P (1998) Sequential Monte Carlo methods in filter theory. PhD dissertation, University of Oxford.
31. Punskaya E (2003) Sequential Monte Carlo methods for digital communications. PhD dissertation, University of Cambridge.
32. Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12: 85-94.
33. Cuff JA, Barton GJ (1999) Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins 34: 508-519.
34. Rost B, Sander C (1993) Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 232: 584-599.