Conditional Random Fields for Transmembrane Helix Prediction

Lior Lukov[1], Sanjay Chawla[1], and W. Bret Church[2]

[1] University Of Sydney[†]
{lior, chawla}@it.usyd.edu.au
[2] University Of New South Wales[‡]
b.church@unsw.edu.au
[†] School of Information Technologies, Sydney University, NSW 2006, Australia
[‡] Department of Physiology and Pharmacology, University of New South Wales, NSW 2052, Australia

Abstract. It is estimated that 20% of genes in the human genome encode for integral membrane proteins (IMPs) and some estimates are much higher. IMPs control a broad range of events essential to the proper functioning of cells, tissues and organisms and are the most common target of clinically useful drugs [1]. However there is a dearth of high-resolution 3D structural information on the IMPs. Therefore good prediction methods of IMPs structures are to be highly valued. In this paper we apply Conditional Random Fields (CRFs) to build a probabilistic model to solve the membrane protein helix prediction problem. The advantage of CRFs is that it allows seamless and principled integration of biological domain knowledge into the model. Our results show that the CRF model outperforms other well known helix prediction approaches on several important measures.

1 Introduction

A number of high throughput projects have been positioned to assist in the interpretation of the human genome sequence data. Structural determination of integral membrane proteins can be problematic due to difficulties in obtaining sufficient amounts of sample. Protein sequence analysis methods extended by our knowledge of protein structure may be suited to contribute significantly to these aspects of protein structure and function.

In this paper we cast the protein helix prediction task as a binary sequential classification problem and use Conditional Random fields (CRFs) to solve it [2]. Given a set of membrane proteins sequences, each single record in the set contains pair of sequences: The observation sequence, represented by x and the label sequence, represented by y. The protein observation sequence is a sequence of amino acids, represented by 20 different letters. The label sequence is a transmembrane helical/non-helical structure sequence represented by binary labels 0/1 respectively. This data, called the training data, is represented by

T.B. Ho, D. Cheung, and H. Liu (Eds.): PAKDD 2005, LNAI 3518, pp. 155–161, 2005.

$T = (x^{(k)}, y^{(k)})_{k=1}^{N}$, where N is the total number of proteins. Our goal is to predict the helical structure of a target set, which has observation sequences only.

2 The Sequential Classification Problem

The sequential classification problem is well known in many different fields such as computational linguistics, part of speech tagging, computational biology and many more. Given set of observation sequences, goal here is to find corresponding label sequences to these observations. A very common approach is using generative models, such as Hidden Markov Models (HMMs), finding the joint probability distribution $p(X, Y)$ where X and Y are random variables describing the observation and the labelled sequences respectively. This approach suffers from a major drawback that in order to find the joint distribution, a generative model has to calculate all possible observation sequences, which may be not practical [3]. In contrast, the conditional models specify the probability of a label given an observation sequence $p(Y|X)$. Thus, no effort is spent on modelling all possible observation sequences, but only on selecting the labels which maximize the conditional probability [2].

3 Conditional Random Fields (CRFs)

Conditional Random Fields (CRFs) is a probabilistic framework for labelling sequential data. CRFs is a form of undirected graphical state model that defines a log-linear distribution for each state over the label sequence based on the observation sequence [3]. CRFs main advantage over other non-generative finite-state models based on directed graphical models, such as Maximum Entropy Markov Models (MEMMs), is by avoiding a weakness called the label bias problem. The Markovian assumptions in MEMMs and similar state-conditional models separate the decision making at one step from future dependent decisions of consecutive steps, and may be biased towards states with fewer outgoing transitions. In contrast, CRFs have a single exponential model for the joint probability of the entire sequence of labels given the observation sequence [2].

Formally, we define $G = (V, E)$ to be an undirected graph when $v \in V$ corresponding to each of the random variables representing a label sequence Y_v from Y and $e \in E$ corresponding to the transition between a given label to the next one. Even though in theory the structure of graph G may be arbitrary, in our application the graph is a simple chain, where each node corresponds to a label [3].

3.1 Definition

Let $G = (V, E)$ be a graph that $Y = (Y_v)v \in V$. If each random variable Y_v in the graph G obeys the Markov property, then (Y, X) is a conditional random field F in which $p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$, where $w \sim v$ are neighbors in G. A clique c in the graph G is defined as a subset of vertices which are completely connected. In a chain graph the cliques are either from first order (single vertex) or second order neighbors (two neighbor vertices).

From the definition of Gibbs Random Fields (GRFs), a set of random variables f is said to be a Gibbs random field if and only if its configuration obey a Gibbs distribution of the form:

$$P(f) = Z^{-1} \times e^{-\frac{1}{T}U(f)} \tag{1}$$

where Z is a normalizing factor: $Z = \sum_{f \in F} e^{-\frac{1}{T}U(f)}$, T is a constant called the temperature which equals to 1 in the most simple case and $U(f)$ is the energy function. By the The Hammersley-clifford theorem if f obeys the Markov property (and positivity) then the physical topology (chain) coincides with the logical topology and the energy function can be expressed as a sum of the cliques's neighbors order:

$$U(f) = \sum_{\{v\} \in C_1} V_1(f_v) + \sum_{\{v,w\} \in C_2} V_2(f_v, f_w) \tag{2}$$

[4]. Since conditional random fields also hold the conditions of Markov random field, then according to Hammersley-clifford theorem, they have a Gibbs distribution, leading us to the fundamental theorem of random fields:

$$p_\theta(y|x) \propto exp\left(\sum_j \lambda_j f_j(y_{i-1}, y_i, x, i) + \sum_k \mu_k g_k(y_i, x, i)\right) \tag{3}$$

where $f_j(y_{i-1}, y_i, x, i)$ is a transition feature function of the entire observation sequence and the labels at positions i and $i-1$, $g_k(y_i, x, i)$ is a state feature function of the entire observation sequence and the label at position i. λ_j and μ_k are estimated from the training data. We assume that the feature functions f_k and g_k are given and fixed [3].

3.2 Feature Functions and Model Estimation

Each potential function actually represents a constraint on subset of random variables on which it operates. Thus, by satisfying a constraint we actually increase the likelihood of the global configuration. In what follows, we look at the transition function as a general case of the state function by writing $g(y_i, x, i) = g(y_{i-1}, y_i, x, i)$. We also define the sum of a feature over the sequence by $F_j(y, x) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i, x, i)$ where $f_j(y_{i-1}, y_i, x, i)$ refers to either transition or state function [3]. Therefore, the probability of a label sequence y given the observation sequence x is in the form

$$p(y|x, \lambda) = \frac{1}{Z(x)} exp(\sum_j \lambda_j F_j(y, x)) \tag{4}$$

where $Z(x) = \sum_y exp(\sum_j \lambda_j F_j(y, x))$. The parameters (λ_j) are computed by maximizing the log-likelihood with the training data using either iterative scaling or conjugate gradient methods [5, 3]. The most likely label sequence $\hat{y}$ for input sequence x is

$$\hat{y} = arg \max_y p(y|x,\lambda) = arg \max_y \sum_j \lambda_j \cdot F_j(y,x)$$

3.3 Feature Integration with the Model

The most important aspect of specifying the model is selecting the set of features that capture the important relationships among the observation and the label sequences, in our case the protein sequence and the helical structure respectively [6]. In our work we have selected a basic set of features capturing the model's constraints and divided them into several groups:

Start, End and Edge Features. By using these features we capture the probability of starting/ending a sequence with assigning a given label or the transition probability for moving from one state to the consecutive state. For instance, the start unigram feature has the form:

$$u_{start}(x,i) = \begin{cases} 1 \text{ if the Amino Acid at position i is the first in the sequence} \\ 0 \text{ otherwise} \end{cases}$$

The relationship between the observation and a potential helix membrane structure is described in the feature:

$$f_{start_H}(y_i,x,i) = \begin{cases} u_{start}(x,i) & \text{if } y_i = \text{Helix membrane} \\ 0 & \text{otherwise} \end{cases}$$

Similarly, we define another set of features for the relationship with a non-helix membrane structure.

The Edge feature in contrast, is a bigram feature which depends on two consecutive labels:

$$f_{edge_{H-H}}(y_{i-1},y_i,x,i) = \begin{cases} u_{edge}(x,i) & \text{if } y_{i-1}\text{=Helix membrane and } y_i\text{=Helix membrane} \\ 0 & \text{otherwise} \end{cases}$$

Basic Amino Acid Feature. Amino acids have different tendencies to populate one membrane helical structure in preference to another. Since our language contains 20 possible amino acids, we have 20 different unigram features from this type. The unigram feature of amino acid n in position i is:

$$u_n(x,i) = \begin{cases} 1 \text{ if the Amino Acid in sequence } x \text{ at position i is from type } n \\ 0 \text{ otherwise} \end{cases}$$

Amino Acid Property Feature. Amino acids differ one from another in their chemical structure expressed by their side chains, providing them different properties. The fact that amino acids from the same classification group tend to appear in similar locations, motivated us to create special property features. We have adopted the properties classification taken from Sternberg [7] classifying the amino acids into nine groups[1], each group described by a unigram feature. Note that some amino acids may appear in more than one group simultaneously.

The hydrophobicity property for instance, is described in the feature:

$$u_{Hydrophobic}(x,i) = \begin{cases} 1 \text{ if the Amino Acid in } x \text{ at position i} \in \text{(M,I,L,V,A,G,F,W,Y,H,K,C)} \\ 0 \text{ otherwise} \end{cases}$$

4 Experiments, Results and Analysis

We now report on our experiments to test the effectiveness of features proposed in Section 3.3, embedded in a CRF model, to predict the location of membrane helical regions in protein sequences.

4.1 Data Set

The data set consists of a set of 148 transmembrane protein sequences with experimentally confirmed transmembrane regions, which are significantly non-similar, based on pairwise similarity clustering compiled by Möller et al [8]. The data set can be accessed via ftp://ftp.ebi.ac.uk/databases/testsets/trans membrane. We randomly picked 24 sequences out of the 148 and grouped them as a test set, using the remaining 124 sequences as the training set. We repeated this procedure ten times, having a cross validation test of ten independent experiments and calculated the average values of these measurements.

4.2 Results and Analysis

In our experiment we have evaluated the prediction accuracy of the test set with the experimentally confirmed results based on two two main approaches: *per-residue accuracy* and *per-segment accuracy* as described in Chen, Kernytsky and Rost (henceforth referred as CKR) [9]. In per-residue accuracy the predicted label and actual label are compared by residue. In per-segment accuracy we determine how accurately a method correctly predicts the location of a transmembrane helix (referred as TMH) region. We have used two popular methods to score per-segment accuracy. The first method requires a minimal overlap of 3 residues between the two corresponding segments and does not allow the same helix to be counted twice, as used in the paper of Chen et al. [9]. This method we refer as $3R$. The second method requires minimal overlap of 9 residues but does allow counting the same helix twice, indicated by $9R$. For our comparison we will closely follow the CKR paper as it has collated results of several methods for transmembrane helix prediction on a common benchmark data set displayed in the following table:

[1] Aromatic (F,W,Y,H), Hydrophobic (M,I,L,V,A,G,F,W,Y,H,K,C), Positive (H,K,R), Polar (W,Y,C,H,K,R,E,D,S,Q,N,T), Charged (H,K,R,E,D), Negative (E,D), Aliphatic (I,L,V), Small (V,A,G,C,P,S,D,T,N), Tiny (A,G,S).

Per-Residue Accuracy					Per-Segment Accuracy				
Q_2	$Q_{2T}^{\%obs}$	$Q_{2T}^{\%prd}$	$Q_{2N}^{\%obs}$	$Q_{2N}^{\%prd}$	$Q_{ok}^{(3R)}$	$Q_{tmh}^{\%obs(3R)}$	$Q_{ok}^{(9R)}$	$Q_{tmh}^{\%obs(9R)}$	$Q_{tmh}^{\%prd}$
83	67	74	92	88	28	43	44	72	99

In order to compare our results with other available methods, we consider the work of Chen et al. [9] and methods contained within as a reference. In the "Per-Residue Accuracy" results we have achieved high prediction accuracy for both transmembrane and non-transmembrane residues, lower accuracy of transmembrane residues only, and higher accuracy of non-transmembrane residues. In the "Per-Segment Accuracy" results we can see a considerable difference between the $3R$ test and the $9R$ test. The figures in $Q_{tmh}^{\%prd}$ indicate high precision of true prediction among those helices who were detected by the model. When comparing our prediction results with the other methods, our model performed well with high percentage of accuracy on the per-residue test. **The CRFs model achieved the highest score among all 28 other methods in the overall percentage of residues predicted correctly in both transmembrane and non-transmembrane helices (Q_2) with 83% of true prediction.** On the per-segment test, our model achieved high precision but low prediction score compared to the other models. Notice that some methods may have involved use of proteins from the data set as training so their results may be overestimates.

5 Conclusions

In this paper we introduced the Conditional Random Fields (CRFs) technique which has found good application in the solution of sequential mining problems. We used CRFs to segment and label sequence data to solve the membrane protein helix prediction problem. Our results look promising compared to currently available methods, and as such will motivate the future use of CRFs to solve sequential labelling data problems. For more information on this paper please check our website on http://www.it.usyd.edu.au/~chawla/publications/crf1.pdf.

References

1. Chen, C.P., Rost, B.: State-of-the-art in membrane protein prediction. Applied Bioinformatics **1** (2002) 21–35
2. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. 18th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA (2001) 282–289
3. Wallach, H.M.: Conditional random fields: An introduction. Technical Report MS-CIS-04-21, University of Pennsylvania (2004)
4. Li, S.: Markov random field modeling in computer vision. Springer-Verlag New York (1995)
5. Berger, A.: The improved iterative scaling algorithm: A gentle introduction. Technical report, Carnegie Mellon University (1997)

6. Buehler, E.C., Ungar, L.H.: Maximum entropy methods for biological sequence modeling. In: BIOKDD. (2001) 60–64
7. Sternberg, M.J.: Protein Structure Prediction: A Practical Approach. Oxford University Press (1996)
8. Moller, S., Kriventseva, E.V., Apweiler, R.: A collection of well characterized integral membrane proteins. Bioinformatics **16** (2000) 1159–1160
9. Chen, C.P., Kernytsky, A., Rost, B.: Transmembrane helixpredictions revisited. Protein Science **11** (2002) 2774–2791