CONDITIONAL RANDOM FIELDS FOR THE PREDICTION OF SIGNAL PEPTIDE CLEAVAGE SITES

Man-Wai Mak

Dept. of Electronic and Information Engineering
The Hong Kong Polytechnic University,
Hong Kong SAR

Sun-Yuan Kung

Dept. of Electrical Engineering
Princeton University, USA
National Chung-Hsing University, ROC

ABSTRACT

Correct prediction of signal peptide cleavage sites has a significant impact on drug design. State-of-the-art approaches to cleavage site prediction typically use generative models (such as HMMs) to represent the statistics of amino acid sequences or use neural networks to detect the changes in short amino-acid segments along a query sequence. By formulating cleavage site prediction as a sequence labeling problem, this paper demonstrates how conditional random fields (CRFs) can be applied to cleavage site prediction. The paper also demonstrates how amino acid properties can be exploited and incorporated into the CRFs to boost prediction performance. Results show that the performance of CRFs is comparable to that of a state-of-the-art predictor (SignalP V3.0). Further performance improvement was observed when the decisions of SignalP and the CRF-based predictor are fused.

Index Terms— Conditional random fields, discriminative models, signal peptides, cleavage sites, protein sequences.

1. INTRODUCTION

The amino acid sequence of a protein contains information about its organelle destination. The information can be considered as zipcode that directs the transport of a protein, ensuring its delivery to the correct secretory pathway [1]. Typically, the information can be found within a short segment of amino acids. These short segments are generally known as sorting-signal sequences, targeting sequences, or signal peptides. After the protein is translocated across the cell membrane, the signal peptide will be *cleaved* off by an extracellular signal peptidase. The location at which the cleave off occurs is called the cleavage site.

The mechanism by which a cell transports a protein to its target location within or outside the cell is called the protein sorting process. Defects in the sorting process can cause serious diseases. Therefore, identifying signal peptides and their cleavage sites have both scientific and commercial values. For instance, to produce recombinant secreted proteins or receptors, it is important to know the exact cleavage sites of signal peptides. The information of signal peptides also allows pharmaceutical companies to manipulate the secretory pathway of a protein by attaching a specially designed tag to it. This ability has opened up opportunity for the design of better drugs.

Although signal sequences that direct proteins to their target location differ in length and contents, common features that make the sequences to act like signals still exist, as exemplified in Fig. 1. For example, all signal sequences have a long central region (the h-region) that is highly hydrophobic. These properties allow the cleavage sites to be predicted computationally. There are three main approaches to cleavage site prediction: weight matrices, neural networks, and hidden Markov models.

1. *Weight Matrices.* A weight matrix is calculated from the position-specific amino acid frequencies of aligned signal peptides (aligned at the cleavage site) [2]. To predict the cleavage site of an unknown sequence, the matrix is scanned against the sequence to find the position of highest sum of weights. A recent implementation based on this approach is the PrediSi [3]. The weight matrix approach is very efficient, but the performance is inferior to more advanced approaches discussed below.
2. *Neural Networks.* This approach uses a sliding window to scan over an amino acid sequence. For each subsequence within the window, a numerically encoded vector is presented to a neural network for detecting whether the current window contains a cleavage site. SignalP 1.1 [4] is one of the best known examples of this approach. In SignalP, symmetric and asymmetric sliding windows are used and the 20 amino acids are converted to numerical vectors using a distributive (sparse) encoding technique. An advantage of this approach is that a wide range physicochemical properties can be selected as network inputs. However, the prediction accuracy is dependent on encoding methods [5].
3. *Hidden Markov Models (HMMs).* In this approach, an

This work was in part supported by The RGC of Hong Kong SAR (PolyU 5251/08E). The research was conducted in part when S.Y. Kung was on leave with the National Chung-Hsing University as a Chair Professor.

Word	This	has	increased	the	risk	of	the	government
POS	DT	VBZ	VBN	DT	NN	IN	DT	NN
Chunk ID	B-NP	O	O	B-NP	I-NP	O	B-NP	I-NP

Table 1. An example sentence with a part-of-speech (POS) tag and a chunk identifier (in IOB2 format) for each word.

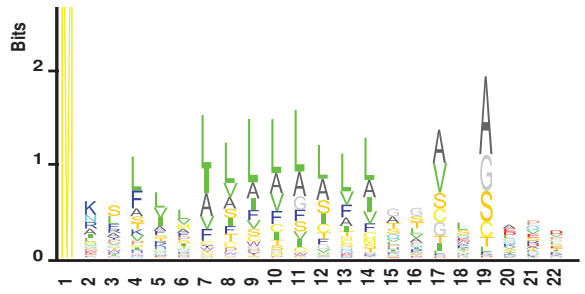

Fig. 1. Logo diagram of 179 signal peptides with cleavage site between Positions 19 and 20. Positions preceding to the cleavage site are rich in hydrophobic (e.g. A and L) and polar (e.g. G and S) residues. The taller the letter, the more often the corresponding amino acid appears in the signal peptides.

amino acid sequence is thought of as generated from a Markov process that emits amino acids according to some probability distributions when transiting probabilistically from state to state. To predict the cleavage site of an unknown sequence, the most likely transition path is found and the amino acid that aligns with the cleavage site node is considered as the cleavage site. One advantage of using HMMs is that biological knowledge can be easily incorporated into the models. For example, in SignalP 2.0 and 3.0 [6,7], the HMM is divided into three regions, each with a length constraint corresponding to the biological constraints of the three regions of signal peptides. Another advantage of HMMs is that symbolic inputs can be naturally accommodated, and therefore numerical encoding as in the neural network approach is not required.

This paper proposes using conditional random fields (CRFs) [8] to predict cleavage site locations. CRFs were originally designed for sequence labelling tasks such as Part-of-Speech (POS) tagging (see Table 1 for an example). Given a sequence of observations, a CRF finds the most likely label for each of the observations. CRFs have a graphical structure consisting of edges and vertices in which an edge represents the dependency between two random variables (e.g., two amino acids in a protein) and a vertex represents a random variable whose distribution is to be inferred. Therefore, CRFs are undirected graphical models, as opposed to directed graphical models such as HMMs. Also, unlike HMMs, the distribution of each vertex in the graph is conditioned on the whole input sequence.

2. CONDITIONAL RANDOM FIELDS

2.1. Formulation

Denote $\mathbf{x} = \{x_1, \ldots, x_T\}$ as an observation sequence and $\mathbf{y} = \{y_1, \ldots, y_T\}$ as the associated sequence of labels. In the case of cleavage site prediction, $\mathbf{x} \in \mathcal{A}$ and $\mathbf{y} \in \mathcal{L} = \{\mathrm{S}, \mathrm{C}, \mathrm{M}\}$, where $\mathcal{A}$ is the set of 20 amino acid letters, and S, C, and M stand for the signal part, cleavage site, and mature part of a protein sequence, respectively. The cleavage site is located at the transition from C to M in $\mathbf{y}$.

Generative models such as HMMs model the joint distribution $p(\mathbf{x}, \mathbf{y})$ and computes the likelihood $p(\mathbf{x}|\mathbf{y})$ by assuming that the state y_t is only responsible for generating the observation x_t. The independence assumption of x_t's restricts HMMs from capturing long-range dependence between $\mathbf{x}$ and $\mathbf{y}$. For example, standard HMMs cannot model explicitly the dependence between x_{t-d} and x_t where $d > 1$ or between x_{t-d} and y_t where $d \neq 0$. Most biological sequences, however, have such long-range dependence [9, 10].

In fact, to predict the labels $\mathbf{y}$ given $\mathbf{x}$, the only distribution needs to be modeled is $p(\mathbf{y}|\mathbf{x})$. CRFs [8] are discriminative models that directly evaluate $p(\mathbf{y}|\mathbf{x})$:

$$p(\mathbf{y}|\mathbf{x}) = \frac{F(\mathbf{x},\mathbf{y})}{Z(\mathbf{x})} = \frac{1}{Z(\mathbf{x})} \prod_{t=1}^{T} \exp\left\{ \sum_{i\in\mathcal{L}} \sum_{j\in\mathcal{L}} \alpha_{ij} f_{ij}(\mathbf{x}, y_{t-1}, y_t) + \sum_{j\in\mathcal{L}} \sum_{k\in\mathcal{P}} \beta_{jk} g_{jk}(\mathbf{x}, y_t) + \sum_{i\in\mathcal{L}} \sum_{j\in\mathcal{L}} \sum_{k\in\mathcal{P}} \gamma_{ijk} h_{ijk}(\mathbf{x}, y_{t-1}, y_t) \right\} \quad (1)$$

where $Z(\mathbf{x}) = \sum_{\mathbf{y}} F(\mathbf{x}, \mathbf{y})$ is a normalization factor, α_{ij}, β_{jk}, and γ_{ijk} are model parameters, f_{ij} are transition-feature functions, g_{jk} and h_{ijk} are state-feature functions, and $\mathcal{P}$ is a set of AA patterns. Therefore, in CRFs, the relationship between adjacent states (y_{t-1}, y_t) is modelled as a Markov random field conditioned on the whole input sequence $\mathbf{x}$. Note that the function $h_{ijk}(\cdot)$ depends on both y_t and y_{t-1}, meaning that the CRF in Eq. 1 is second order.

2.2. Feature Functions

The definitions of feature functions depend on the application. In fact, one advantage of CRFs is the freedom of choosing suitable feature functions for modeling. This allows investigators to incorporate domain knowledge into the model. Typically, the feature functions are boolean functions of the form:

$$f_{ij}(\mathbf{x}, y_{t-1}, y_t) = \begin{cases} 1 & \text{if } y_{t-1} = i \text{ and } y_t = j \\ 0 & \text{Otherwise} \end{cases} \quad (2)$$

$$g_{jk}(\mathbf{x}, y_t) = \begin{cases} 1 & \text{if } y_t = j \text{ and } b(\mathbf{x}, t) = k \\ 0 & \text{Otherwise} \end{cases} \quad (3)$$

$$h_{ijk}(\mathbf{x}, y_{t-1}, y_t) = \begin{cases} 1 & \text{if } y_{t-1} = i, y_t = j \text{ and } b(\mathbf{x}, t) = k \\ 0 & \text{Otherwise} \end{cases} \quad (4)$$

where $i, j \in \mathcal{L}$, $k \in \mathcal{P}$, and $b(\mathbf{x}, t)$ is a function that depends on the amino acids in $\mathbf{x}$ around position t. One possibility is to use n-grams of the amino acid alphabet as $\mathcal{P}$ and the residues near position t as $b(\mathbf{x}, t)$. More formally, we have

$$\mathcal{P} = \text{n-gram}(\mathcal{A}) \text{ and } b(\mathbf{x}, t) = x_{t-d_1} x_{t-d_2} \cdots x_{t-d_n}, \quad (5)$$

where $d_1 > d_2 > \cdots > d_n$. A large d_i enables the CRF to capture the long-range dependence among the amino acids in the input sequence.

2.3. Advantages of CRFs

The CRFs enjoy several advantages over the HMMs.

1. *Avoid computing likelihood*. Because CRFs are discriminative models that compute the conditional probability $p(\mathbf{y}|\mathbf{x})$, it is not necessary to compute the likelihood of the input observation. It is commonly believed that discriminative models are superior to generative models [11].
2. *Model long-range dependence.* CRFs can model long-range dependence between the labels and observations without making the inference problem intractable.
3. *Guarantee global optimal.* The global normalization in Eq. 1 means that the global optimal solution can always be found.
4. *Alleviate label-bias problem.* Many discriminative models, such as the maximum entropy Markov model, are prone to the label-bias problem (preferring states with fewer outgoing transitions) [8]. Because CRFs use global normalization, they possess the advantages of discriminative models but without suffering from the label bias problem.

3. CRF FOR CLEAVAGE SITE PREDICTION

To use CRFs for cleavage site prediction, the prediction problem is formulated as a sequence labelling task. Similar to the part-of-speech tagging task in Table 1 where words are categorized as different types, amino acids of similar properties can be categorized as sub-groups. This paper divides the 20 amino acids according to their hydrophobicity and charge/polarity shown in Table 2. These properties are believed to posses information about cleavage sites because the h-region of signal peptides is rich in hydrophobic residues and the c-region (positions -1 and -3) is dominated by small, non-polar residues [12].

An example amino acid (AA) sequence with the corresponding derived hydrophobicity sequence and charge/polarity sequence is shown below:

AA Sequence ($\mathbf{x}$) :	`T-Q-T-W-A-G-S-H-S`
Hydrophobicity ($\mathbf{x}$) :	`2-1-2-3-3-2-2-2-2`
Charge/Polarity ($\mathbf{x}$) :	`3-3-3-4-4-3-3-2-3`
Labels ($\mathbf{y}$) :	`S-S-S-S-C-M-M-M-M`

Property	Group
Hydrophobicity	H1={D,E,N,Q,R,K} H2={C,S,T,P,G,H,Y} H3={A,M,I,L,V,F,W}
Charge/Polarity	C1={R,K,H} C2={D,E} C3={C,T,S,G,N,Q,Y} C4={A,P,M,L,I,V,F,W}

Table 2. Grouping of amino acids according to their hydrophobicity and charge/polarity [13].

where the numbers in the 2nd and 3rd rows correspond to the hydrophobicity and charge/polarity groups shown in Table 2. Note that either AA, hydrophobicity, charge/polarity, or their combinations can be used as observations to train a CRF.

4. EXPERIMENTS AND RESULTS

4.1. Data and Procedures

Amino acid sequences of eukaryotic proteins with experimentally found cleavage sites were extracted from the flat files of Swissprot Release 56.5 using the programs provided by Menne et al. [14], which results in 1,937 sequences. Ten-fold cross validations were applied to these sequences to obtain the prediction accuracies.

For the 1st-order state features (g_{jk}), the property set $\mathcal{P}$ contains n-grams of amino acids, hydrophobicity groups, and polarity/charge groups, where $n = 1, \ldots, 5$. For the 2nd-order state features (h_{ijk}), only uni-grams and bi-grams were used. CRF++[1] was used to implement the CRFs. The parameters `-c` and `-f` were set to 1.0.

4.2. Results and Discussions

Effectiveness of AA Properties. To investigate the effectiveness of using hydrophobicity and charge/polarity groups as observations, CRFs that use different types of input sequences were trained. The results are shown in Table 3. Evidently, the amino acids provide the most relevant information for the prediction task. Although the hydrophobicity groups or charge/polarity groups by themselves are not very effective, they can help improve the prediction performance when used with the amino acids.

Effectiveness of Feature Functions. To analyze the contribution of different types of feature functions to the prediction accuracy, transition features, 1st-order state features, and 2nd-order states features were progressively added to the CRFs. The results are shown in Table 4. Evidently, using either transition features (f_{ij}) or first-order state features (g_{jk}) exclusively leads to very poor performance. However, once both features were used together, performance improves significantly (from 43.06% to 79.71%). This suggests that amino

[1] http://crfpp.sourceforge.net/

Type of Input Sequence	Accuracy
Amino Acids (AA) only	79.19%
Hydrophobicity only	38.26%
Charge/Polarity only	32.89%
Hydrophobicity + Charge/Polarity	44.76%
AA + Charge/Polarity	78.88%
AA + Hydrophobicity	79.40%
AA + Hydrophobicity + Charge/Polarity	79.92%

Table 3. Prediction accuracies achieved by CRFs using different types of input sequences.

Types of Feature Functions	Accuracy
f_{ij}	10.53%
g_{jk}	43.06%
$f_{ij} + g_{jk}$	79.92%
h_{ijk}	66.60%
$f_{ij} + h_{ijk}$	66.60%
$g_{jk} + h_{ijk}$	78.88%
$f_{ij} + g_{jk} + h_{ijk}$	79.81%

Table 4. Accuracy of cleavage site prediction achieved by CRFs using different types of feature functions. f_{ij}: Transition features; g_{jk}: 1st-order state features; h_{ijk}: 2nd-order state features. See Eqs. 2–4 for their formulation.

acids in a sequence are dependent on each other. Interestingly, transition features become redundant when 2nd-order features are used (comparing h_{ijk} and $f_{ij} + h_{ijk}$), suggesting that the latter can capture the dependence between the amino acids.

Compared with State-of-the-Art Predictors. We compared the performance of the CRF-based predictor with SignalP V3.0 [7] and PrediSi [3]. Table 5 shows that SignalP performs the best, followed by CRF and PrediSi. We noticed from the outputs of SignalP and CRF that for some sequences, when SignalP made a wrong decision, CRF made a correct one. This suggests a potential performance improvement by fusing their decisions. We implemented the fusion as follows: Select the decision of CRF if the Z-norm score of CRF is greater than that of SignalP plus a decision threshold determined from training data; otherwise select the decision of SignalP. Table 5 suggests that fusing the decisions of SignalP and CRF can increase the prediction accuracy. The p-values (based on McNemar's tests [15]) in Table 5 also show that the fusion result is significantly better than that of SignalP and CRF.

5. CONCLUSION AND FUTURE WORK

This paper has demonstrated the application of conditional random fields to signal-peptide cleavage site prediction and shown that CRFs' predictions are complementary to those of SignalP. Possible extensions of this work include replacing the category groups by real-values such as hydrophobicity profiles computed by averaging the hydrophobicity scales of AA residues within a sliding window.

Cleavage Site Predictor	Accuracy	p-value
SignalP [7]	81.88%	–
PrediSi [3]	77.06%	0.0003
CRF	79.92%	0.0181
CRF + SignalP	83.12%	0.0071

Table 5. Accuracy and p-values [15] (with respect to SignalP) of different cleavage site predictors.

6. REFERENCES

[1] L. M. Gierasch, "Signal sequences," *Biochemistry*, vol. 28, pp. 923–930, 1989.

[2] G. von Heijne, "A new method for predicting signal sequence cleavage sites," *Nucleic Acids Research*, vol. 14, no. 11, pp. 4683–4690, 1986.

[3] K. Hiller, A. Grote, M. Scheer, R. Munch, and D. Jahn, "PrediSi: Prediction of signal peptides and their cleavage positions," *Nucleic Acids Research*, vol. 32, pp. 375–379, 2004.

[4] H. Nielsen, J. Engelbrecht, S. Brunak, and G. von Heijne, "A neural network method for identification of prokaryotic and eukaryotic signal perptides and prediction of their cleavage sites," *Int. J. Neural Sys.*, vol. 8, pp. 581–599, 1997.

[5] S. R. Maetschke, M. Towsey, and M. B. Boden, "BLOMAP: An encoding of amino acids which improves signal peptide cleavage site prediction," in *3rd Asia Pacific Bioinformatics Conference*, Y. P. Phoebe Chen and L. Wong, Eds., Singapore, 17-21 Jan 2005, pp. 141–150.

[6] H. Nielsen and A. Krogh, "Prediction of signal peptides and signal anchors by a hidden Markov model," in *Proc. Sixth Int. Conf. on Intelligent Systems for Molecular Biology*, J. Glasgow et al., Ed. 1998, pp. 122–130, AAAI Press.

[7] J. D. Bendtsen, H. Nielsen, G. von Heijne, and S. Brunak, "Improved prediction of signal peptides: Signalp 3.0," *J. Mol. Biol.*, vol. 340, pp. 783–795, 2004.

[8] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. on Machine Learning*, 2001.

[9] O. Weiss and H. Herzel, "Correlations in protein sequences and property codes," *J. theor. Biol*, vol. 190, pp. 341–353, 1998.

[10] C. Hemmerich and S. Kim, "A study of residue correlation within protein sequences and its application to sequence classification," *EURASIP J. Bioinformatics Syst. Biol.*, vol. 2007, no. 1, pp. 9–9, 2007.

[11] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," in *Advances in Neural Information Processing 14*, Cambridge, MA, 2002, MIT Press.

[12] G. von Heijne, "Patterns of amino acids near signal-sequence cleavage sites," *Eur J Biochem.*, vol. 133, no. 1, pp. 17–21, Jun 1983.

[13] C. H. Wu and J. M. McLarty, *Neural Networks and Genome Informatics*, Elsevier, New York, 2000.

[14] K. M. L. Menne, H. Hermjakob, and R. Apweiler, "A comparison of signal sequence prediction methods using a test set of signal peptides," *Bioinformatics*, vol. 16, pp. 741–742, 2000.

[15] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. ICASSP'89*, 1989, pp. 532–535.