FUSION OF CONDITIONAL RANDOM FIELD AND SIGNALP FOR PROTEIN CLEAVAGE SITE PREDICTION

Man-Wai Mak and Wei Wang

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

Prediction of protein cleavage sites is an important step in drug design. Recent research has demonstrated that conditional random fields are capable of predicting the cleavage site locations of signal peptides, and their performance is comparable to that of SignalP—a state-of-the-art predictor based on hidden Markov models and neural networks. This paper investigates the degree of complementarity between CRF-based predictors and SignalP and proposes using the complementary properties to fuse the two predictors. It was found that about 40% of the sequences that are incorrectly predicted by SignalP can be correctly predicted by CRF, and that about 30% of the sequences that are incorrectly predicted by CRF can be correctly predicted by SignalP. This suggests that the two predictors complement each other. The paper also shows that the performance of CRF can be further improved by constructing the state features from spatially dispersed amino acids in the training sequences.

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

Web Service: http://158.132.148.85:8080/CSitePred/faces/Page1.jsp

1. INTRODUCTION

1.1. Signal Peptides and Their Cleavage Sites

A newly created protein will either transported to an organelle of a cell or secreted outside the cell through a secretary pathway [1]. The destination information can be found in a short segment of the amino acid sequence of the protein, which is in some way analogous to the IP address of a TCP/IP packet in data communication or the zipcode of letters. These short segments are generally known as sorting-signals, 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.

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.

1.2. Importance of Cleavage Site Prediction

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.

1.3. Existing Cleavage-Site Prediction Methods

Due to the ever increase in the number of new proteins entering the data banks and the time involved in identifying signal peptides and determining their cleavage sites by experimental means, the development of effective computation tools for cleavage site prediction has become increasingly important. However, because of the great variation in length and degree of conservation of signal peptides in different proteins, finding the cleave sites by computation means is a challenging task.

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.

The earliest approach to cleavage site prediction is to compute a weight matrix based on 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.

Different machine learning techniques have been applied to cleavage site prediction. For example, in SignalP 1.1 [4], a sliding window is applied 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. 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 the encoding methods [5]. In SignalP 2.0 and 3.0 [6, 7], an 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 this approach is that biological knowledge can be easily incorporated into the models. Another advantage is that symbolic inputs can be naturally accommodated, and therefore numerical encoding as in the neural network approach is not required.

1.4. Proposed Method

In our previous investigation [8], we have shown that conditional random fields (CRFs) [9] are capable of predicting cleavage site locations and that the prediction accuracy of CRFs is comparable to that of SignalP. In this paper, we extend our previous work in two fronts: (1) we investigate the degree of complementarity between CRF-based predictors and SignalP and propose a new fusion scheme based on the complementary information; and (2) we attempt to improve the prediction accuracy of CRFs by using spatially dispersed amino acids to construct the state features of the CRFs. Evaluation based on the signal peptides extracted from the Swissprot database shows that SignalP and CRFs posses significant complementary information, leading to better prediction performance when this information is exploited in the fusion process.

2. CONDITIONAL RANDOM FIELDS

CRFs were originally designed for sequence labeling tasks such as Part-of-Speech (POS) tagging, as exemplified in Table 1. 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.

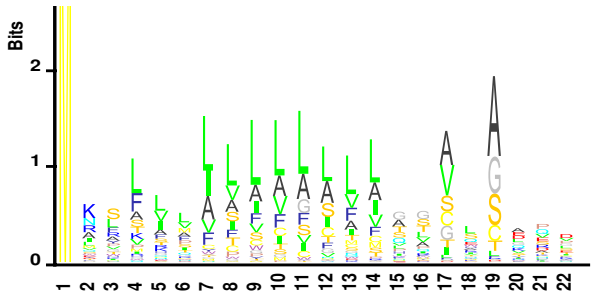


Fig. 1. Logo diagram of 179 signal peptides with cleavage site between Positions 19 and 20. Positions preceding to the cleavage site are rich in hydrophobic (e.g. A and L) and polar (e.g. G and S) residues. The taller the letter, the more often the corresponding amino acid appears in the signal peptides.

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.

2.1. Formulation

Denote

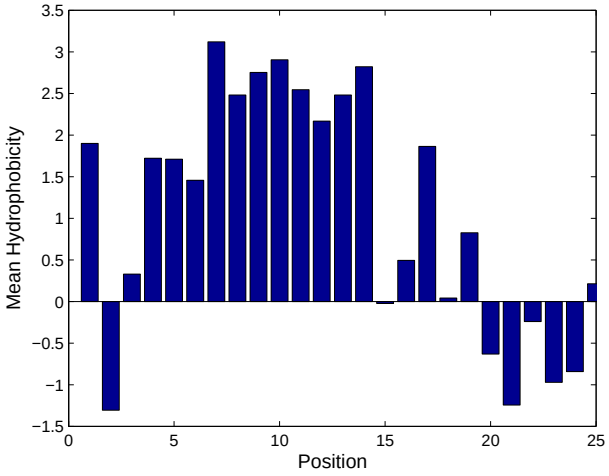
$$\mathbf{x} = \{x_1, \ldots, x_T\} \quad \text{and} \quad \mathbf{y} = \{y_1, \ldots, y_T\}$$

as an observation sequence and the associated sequence of labels, respectively. In the case of cleavage site prediction,

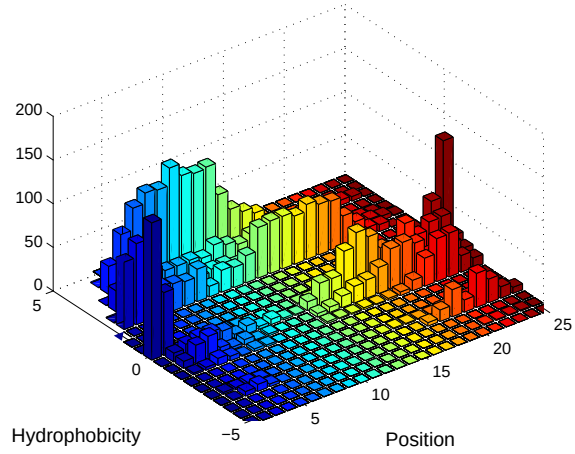
$$\mathbf{x} \in \mathcal{A} \quad \text{and} \quad \mathbf{y} \in \mathcal{L} \equiv \{\mathrm{S}, \mathrm{C}, \mathrm{M}\},$$

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. In other words, when predicting the label at position t, HMMs cannot directly use information other than 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 [10, 11]. Fig. 3 shows the correlation of amino acids at different positions relative to the cleavage site. Evidently, there is significant correlation between amino acids at non-adjacent positions. In particular, the correlation is fairly strong between amino acids at positions -6 and -14, which are 8 positions apart.



(a)



(b)

Fig. 2. (a) The mean and (b) the histograms of hydrophobicity of 179 signal peptides at different sequence positions. The cleavage site of these sequences is between Positions 19 and 20.

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 [9] are discriminative models that directly evaluate $p(\mathbf{y}|\mathbf{x})$:

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

where $Z(\mathbf{x}) = \sum_{\mathbf{y}} F(\mathbf{x},\mathbf{y})$ is a normalization factor, α_{ij} and β_{jk} are model parameters, $f_{ij}(\cdot)$ are transition-feature functions, $g_{jk}(\cdot)$ are state-feature functions, $\mathcal{P}$ is a set of amino acid patterns (see Section 2.2 for an example), and $|\mathcal{L}|$ is the cardinality of the set $\mathcal{L}$. 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 $\mathbf{x}$.

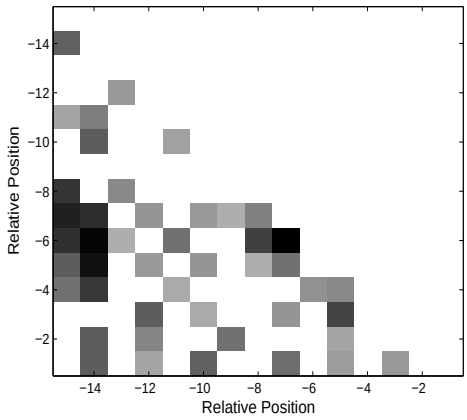


Fig. 3. Correlation of hydrophobicity of 1695 protein sequences at different positions relative to their cleavage site. Entries in gray mean that the correlation between the hydrophobicity at the corresponding relative positions are statistically significant (p-value < 0.05). The grayness is proportional to the degree of correlation. Correlation at identical relative positions, which gives maximum correlation, is not shown for clarity of display.

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.

To facilitate presentation in the sequel, let's denote $\mathcal{L}_i$ as the i-th label in $\mathcal{L}$, e.g., $\mathcal{L}_1 \equiv$ S. A similar notation is also applied to $\mathcal{P}$. The feature functions are typically boolean functions of the form:

$$f_{ij}(y_{t-1}, y_t) = \begin{cases} 1 & \text{if } y_{t-1} = \mathcal{L}_i \text{ and } y_t = \mathcal{L}_j \\ 0 & \text{Otherwise} \end{cases} \tag{2}$$

$$g_{jk}(\mathbf{x}, y_t) = \begin{cases} 1 & \text{if } y_t = \mathcal{L}_j \text{ and } b(\mathbf{x}, t) = \mathcal{P}_k \\ 0 & \text{Otherwise} \end{cases} \tag{3}$$

where $1 \le i, j \le |\mathcal{L}|$, $1 \le k \le |\mathcal{P}|$, 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

$$\mathcal{P} = \text{n-gram}(\mathcal{A}) \;\text{ and }\; b(\mathbf{x}, t) = x_{t-d_1} x_{t-d_2} \cdots x_{t-d_n}, \tag{4}$$

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.

The operation of the feature functions can be explained via a simple example. Consider the amino acid sequence and its labels in Table 2. At $t = 5$, we have $y_4 = \text{S}$ and $y_5 = \text{C}$. Because $\mathcal{L}_1 = \text{S}$, $\mathcal{L}_2 = \text{C}$, and $\mathcal{L}_3 = \text{M}$, we have $f_{1,2}(y_4, y_5) = 1$. Assume that bi-gram is used for generating $\mathcal{P}$, i.e.,

$$\mathcal{P} = \{\texttt{AA}, \texttt{AC}, \ldots, \texttt{WA}, \ldots, \texttt{YY}\},$$

and that $d_1 = 1$ and $d_2 = 0$. Assume further that the amino acid pair WA occupies position k in $\mathcal{P}$, i.e., $\mathcal{P}_k = \texttt{WA}$. Then, we have $b(\mathbf{x}, 5) = \texttt{WA} = \mathcal{P}_k$ and therefore $g_{2,k}(\mathbf{x}, y_5) = 1$.

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 has been shown that discriminative models are usually superior to the generative models [12] because computing the probability of the observation is avoided.
2. *Model long-range dependence.* CRFs can model long-range dependence between the labels and observations without making the inference problem intractable, making it particularly useful for text processing [9] and bioinformatics [13].
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) [9]. 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 POS tagging task [14] in Table 1 where words are categorized as different types, amino acids of similar properties can be categorized as sub-groups.[1] We propose to divide the 20 amino acids according to their hydrophobicity and charge/polarity as shown in Table 3. These properties are used because the h-region of signal peptides is rich in hydrophobic residues and the c-region is dominated by small, non-polar residues [16], as illustrated in Fig. 1. Moreover, as illustrated in Fig. 2, the degree of hydrophobicity is also very different at different positions. It is believed that different sets of alphabets can complement each other in finding significant conserved regions along the amino acid residues. In case several alphabet sets indicate the same conserved region, that region is also likely to be of functionally important to the protein.

[1]This is called alphabet indexing [15] in the literature.

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 3. Grouping of amino acids according to their hydrophobicity and charge/polarity [17].

Table 2 shows an example amino acid sequence together with its hydrophobicity sequence and charge/polarity sequence. Note that either amino acid, hydrophobicity, charge/polarity, or their combinations can be used as observations to train a CRF.

4. FUSION OF CRF AND SIGNALP

We noticed from the outputs of SignalP and CRF that for some sequences, when CRF made a wrong decision, SignalP made a correct one. Similarly, there are also sequences whose cleavage sites are incorrectly predicted by CRF but correctly predicted by SignalP. This suggests a potential performance improvement by fusing the decisions of CRF and SignalP. To fuse the two decisions, some kinds of reliability scores need to be determined. For CRF, we used the probability of the best viterbi path, and for SignalP, we used the C_{max} scores. Hereafter, we refer to these scores as CRF scores and SignalP scores, respectively.

Table 4 (upper part) shows the number of sequences with CRF scores smaller than some pre-defined thresholds, below which the predicted sites are deemed untrustworthy. The table shows that less than 40% of these untrustworthy decisions are correct, suggesting that CRF has difficulty in predicting the cleavage sites of these sequences. On the other hand, among these sequences, over 60% of them can be correctly predicted by SignalP. The situation is reversed in the lower part of Table 4. In particular, while SignalP can only predict the difficult sequences at a rate of 54%–69%, the CRF achieves 97% accuracy on these sequences.

Based on these observations, we implemented the fusion as follows.

Step 1 Given a query sequence $\mathbf{x}$, present it to the CRF and SignalP to obtain a CRF score (denoted $\texttt{crf}(\mathbf{x})$) and a SignalP score (denoted $\texttt{snp}(\mathbf{x})$), respectively.

Step 2 Perform z-norm independently on these two scores to obtain the z-norm scores, namely $\texttt{crfn}(\mathbf{x})$ and $\texttt{snpn}(\mathbf{x})$.

AA Sequence ($\mathbf{x}$)	T	–	Q	–	T	–	W	–	A	–	G	–	S	–	H	–	S
Hydrophobicity ($\mathbf{x}$)	H_2	–	H_1	–	H_2	–	H_3	–	H_3	–	H_2	–	H_2	–	H_2	–	H_2
Charge/Polarity ($\mathbf{x}$)	C_3	–	C_3	–	C_3	–	C_4	–	C_4	–	C_3	–	C_3	–	C_2	–	C_3
Label ($\mathbf{y}$)	S	–	S	–	S	–	S	–	C	–	M	–	M	–	M	–	M

Table 2. An example amino acid sequence with the corresponding hydrophobicity sequence and charge/polarity sequence. The 2nd and 3rd rows represent the hydrophobicity and charge/polarity groups shown in Table 3.

Step 3 Determine the cleavage site position according to

$$\mathtt{p}(\mathbf{x}) = \begin{cases} \text{SignalP's decision} & \text{if } \mathtt{snpn}(\mathbf{x}) > \mathtt{crfn}(\mathbf{x}) - \epsilon \\ & \text{or} \\ & \mathtt{crfn}(\mathbf{x}) < \eta \\ \text{CRF's decision} & \text{otherwise} \end{cases}$$

where ϵ and η are predefined constants that can be be determined from training data. In this work, $\epsilon = 0.8$ and $\eta = -2$. A positive ϵ means that the cleavage site position is based on CRF only when the normalized CRF score is significantly higher than the normalized SignalP scores.

5. EXPERIMENTS AND RESULTS

5.1. Materials 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. [18], which results in 1,937 sequences. Ten-fold cross validations were applied to these sequences to obtain the prediction accuracies.

The property set $\mathcal{P}$ for the state-feature function $f_{jk}(\cdot)$ contains n-grams of amino acids, where $n = 1, \ldots, 5$, and bi-gram of hydrophobicity groups and polarity/charge groups. CRF++ was used to implement the CRFs.[2] The parameters `-c` and `-f` were set to 1.0.

To investigate the effect of the varying the maximum allowable offset for indexing amino acids in a sequence on prediction accuracy, various values of $\max\{d_n\}$ in Eq. 4 were tried.

5.2. Results and Discussions

5.2.1. Effect of Indexing Offsets

Table 5 shows the performance of CRF at different value of $\max\{d_n\}$. Evidently, varying the maximum allowable offset affects the prediction performance. The superiority of large offset seems to suggest that signal sequences exhibit long-range dependency. However, this conjecture needs to be confirmed biologically.

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

Maximum Allowable Offset	Prediction Accuracy
5	80.54%
6	81.41%
7	82.40%
8	83.17%
9	**83.32**%
10	83.12%
11	82.71%
12	82.40%

Table 5. Accuracy of CRF predictors at different maximum AA position offsets, i.e., $\max\{d_n\}$ in Eq. 4.

Cleavage Site Predictor	**Accuracy**
SignalP [7]	81.88%
PrediSi [3]	77.06%
CRF5 [8]	79.71%
CRF5 + SignalP [8]	83.12%
CRF9	83.32%
CRF9 + SignalP	**85.03**%

Table 6. Accuracy of different cleavage site predictors and the fusion of CRF and SignalP. CRF5 and CRF9 stand for CRFs with window size of 5 and 9 amino acids, respectively.

5.2.2. 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 6 shows that CRF with window size of 9 performs the best, followed by SignalP and PrediSi.

5.2.3. Fusion of CRF and SignalP

Table 6 suggests that fusing the decisions of SignalP and CRF can increase the prediction accuracy. In particular, the fusion strategy adopted in this study achieves an even higher performance than the one we used in [8].

6. WEB INTERFACE AND SERVICES

To facilitate researchers to use CRF for cleavage site prediction, a web server called CSitePred was developed.[3] CSitePred allows users to submit amino acid sequences by either coping-and-pasting FASTA format sequences into a window or uploading a FASTA file containing a large number of sequences. The web server returns the most likely cleavage site locations

[3]http://158.132.148.85:8080/CSitePred/faces/Page1.jsp

Predictor	Score Threshold	No. of seqs below threshold (deemed untrustworthy)	No. of seqs correctly predicted by CRF	No. of seqs correctly predicted by SignalP
CRF	0.60	94	32 (34.0%)	64 (68.1%)
	0.65	125	44 (35.2%)	81 (64.8%)
	0.70	156	61 (39.1%)	96 (61.5%)
SignalP	0.60	444	426 (96.0%)	243 (54.7%)
	0.70	668	647 (97.0%)	412 (61.7%)
	0.80	917	893 (97.4%)	628 (68.5%)

Table 4. The complement between CRF and SignalP. The 3rd column is the number of sequences whose cleavage-site scores are less than the CRF score threshold (upper part) and SignalP score threshold (lower part).

and their corresponding prediction scores of the submitted sequences to the user. Therefore, prediction on individual sequences or whole datasets are supported. Users can also invoke a collection of web services via the WSDL interface of the software.

7. CONCLUSIONS

This paper has demonstrated that there is a high degree of complementarity between CRF-based predictors and SignalP, and that this complementary information can be easily exploited to fuse the two types of predictors in a protein cleavage site prediction task. The paper also shows that the CRF can be further enhanced by constructing state features from more spatially dispersed amino acids along the peptide chain. A Web interface and a collection of web services of the CRF-based predictor are available online.

8. 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] M. W. Mak and S. Y. Kung, "Conditional random fields for the prediction of signal peptide cleavage sites," in *Proc. ICASSP*, Taipei, April 2009.

[9] 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.

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

[11] 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.

[12] 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.

[13] K. Sato and Y. Sakakibara, "RNA secondary structural alignment with conditional random fields," *Bioinformatics*, vol. 21, pp. 237–242, 2005.

[14] L. A. Ramshaw and M. P. Marcus, "Text chunking using transformation-based learning," in *Proc. of the Third Workshop on Very Large Corpora*, Cambridge, MA, 1995.

[15] S. Shimozono, "Alphabet indexing for approximating features of symbols," *Theor. Comput. Sci.*, vol. 210, no. 2, pp. 245–260, 1999.

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

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

[18] 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.