Conditional Random Fields 法を用いたシグナルペプチドの開裂部分の予測

岩西雄大[1]　三森智裕[1]　蓬莱尚幸[1]　土井晃一[1]　土居洋文[12]
1 奈良先端科学技術大学院大学　情報科学研究科
2 セレスター･レキシコ･サイエンシズ株式会社
E-mail: {k-iwani , mitsumor , hisayu-h , doy}@is.naist.jp , doi@cl-sciences.co.jp

近年、自然言語処理の分野において、既存の手法の問題を解決するとして期待されているConditional Random Fields (CRFs)と蛋白質機能予測の分野において、注目されているSupport Vector Machineを適用して、シグナルペプチド開列部分の予測を行った。本稿では、2つの機械学習法に対して先行研究と同様の実験方法を用いて、既存手法のNeural Network (NN)との予測性能の比較およびCRFsがSVMの予測性能が高いことについて論じる。

Prediction of Signal Peptides Cleavage Sites Using Conditional Random Fields.

Katsuhiro Iwanishi[1]　Tomohiro Mitsumori[1]　Hisayuki Horai[1]
Kouichi Doi[1]　Hirohumi Doi[12]
1 Graduate School of Information Science, Nara Institute of Science and Technology
2 Celestar Lexico-Sciences, Inc.
E-mail: {k-iwani , mitsumor , hisayu-h , doy}@is.naist.jp , doi@cl-sciences.co.jp

In recent year, Conditional Random Fields is expected to resolve existent methods show high performance in field of Natural Language Processing. Support Vector Machine used in field of prediction of protein function. We predicted cleavage sites of signal peptides with their two methods. This paper presents that we compared to comparison of Neural Network prediction performance in the experiment method of precedence research and show that the performance of CRFs is higher than SVM.

1．はじめに

真核生物および原核生物の分泌タンパク質や膜タンパク質の多くは，N 末端に 15～30 残基からなる疎水性アミノ酸残基を多く含むシグナルペプチドとよばれる配列をもっている。シグナルペプチドの役割は、分泌型のタンパク質が膜を通過することを指令することである。シグナルペプチドを見つけることは細胞生物学やタンパク質の研究においてとても重要なことである。例えばシグナルペプチドを知ることは、遺伝病を分子レベルで解明することや新薬開発、医薬品の服用による副作用発現抑制を期待することができる。

Bendtsen [1]らは Hidden Markov Model (HMM)と Neural Network (NN)を用いてシグ

ナルペプチドの開裂部分の予測を行った。Bendtsen らは、真核生物、グラム陽性細菌、およびグラム陰性細菌のシグナルペプチドの予測を行っている。彼らは注目しているアミノ酸の前後のアミノ酸 n 個(n は自然数)を機械学習の素性とし、また最適な n が生物種で異なることを示した。

本研究ではバイオインフォマティクスや自然言語処理の分野で注目されている Support Vector Machine (SVM) と Conditional Random Fields(CRFs)を用いてシグナルペプチドの開裂部分を予測する。またその結果を先行研究で用いられてきた HMM および NN の結果と比較、検討を行う。

・シグナルペプチド

54 41BB_MOUSE 24 T CELL ANTIGEN 4-1BB PRECURSOR.

MGNNCYNVVVIVLLLVGCEKVGAVQNSCDNCQPGTFCRKYNPVCKSCPPSTFSS

SSSSSSSSSSSSSSSSSSSSSSSCMMMMMMMMMMMMMMMMMMMMMMMMMMM

・細胞質内タンパク質, 核タンパク質

70 ACTM_HALRO ACTIN, MUSCLE.

MSDGEEDTTAIVCDNGSGLVKSGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYV.....

MMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMM.....

図 1 Bendtsen らのシグナルペプチド、細胞質タンパク質および核タンパク質のデータセットの例

2. 方法

2.1 シグナルペプチドのデータセット

シグナルペプチドの予測のために、Neilsen[3]ら[1]によって提供されたデータセットを使用した。このデータセットは SWISS-PROT version 29 [2] から作成された。このデータは真核生物 (eukaryote) と原核生物 (prokaryote) に分けられており、また原核生物はさらにグラム陽性細菌 (gram positive) とグラム陰性細菌 (gram negative) に分けられている。

データセットでは、シグナルペプチドと、シグナルペプチドの開裂部分から下流 30 個のアミノ酸配列が含まれている。また細胞質内のタンパク質（真核生物のみ）と核タンパク質の最初の 70 残基のアミノ酸配列も含まれている。図１にデータデットの例を示す。ここではシグナルペプチドと、細胞質タンパク質・核タンパク質を分けて示す。シグナルペプチドでは、１行目に配列の長さ(54)、SWISS-PROT の ID(41BB_MOUSE)、シグナルペプチドの長さ(24)、および SWISS-PROTDE (description) の情報(T CELL .....)が記されている。２行目にはアミノ酸配列が記されている。３行目では２行目のアミノ酸配列に対応させて、シグナルペプチド”S”、開裂部分”C”および開裂部分の下流”M”が記されている。細胞質内タンパク質および核タンパク質では、これらのタンパク質はシグナルペプチドを含まないので３列目

[1] http://www.cbs.dtu.dk/ftp/signalp/

は"M"のみである。

さらにこのデータセットでは Nielsen[3]らの方法を使って相同性のある配列は取り除かれている。相同性のある配列が含まれているとシステムの予測性能を過大評価するためである。この手続きによって 13%から 56%の配列が取り除かれた。最終的に取り出された配列の数を表 1 に示す。

表 1　Bendtsen らのデータセット

	Signal peptides	Non-secretory proteins
Eukaryote	1011	820
Gram -	266	186
Gram+	141	64

2.2　機械学習法

2.2.1 SVM

SVM[4]は、2000年にChristianiniとShawe-Taylorが提唱したカーネルトリックと呼ばれる方法を用いて、非線形の識別関数を構成できるように拡張した機械学習法である。現在知られている多くの手法の中でも最も認識性能の優れた学習モデルの一つであると考えられている。式(1)の正負によって、２つのどちらのクラスにあるかを判別する２クラス分類のためのパターン認識手法である。

$$f(\vec{x}_i) = sign(\vec{w} \cdot \vec{x}_i + b) \qquad (1)$$

この式は、重みベクトル $\boldsymbol{w}$ および係数 $\boldsymbol{b}$ によって分離超平面が規定されることを意味している。この超平面は、それぞれのクラス境界付近の特徴ベクトル $\boldsymbol{x_i}$ 同士の距離（マージン）を最大化するように選択され、結果的に、クラス間境界上の特徴ベクトル(この特徴ベクトルをサポートベクトル(SV)と呼ぶ)のみによって定義される。したがって、従来の機械学習法でみられた、データの次元数が増えることにより計算量や記憶容量が急増するようなことがなく、非常に高次元なベクトルを扱うことも容易である。本研究ではSVMを用いた汎用テキストチャンカーであるYamCha[2]を用いる。

また高次元の特徴空間に写像し、図 2 に示すような非線形クラス分類が可能となる。

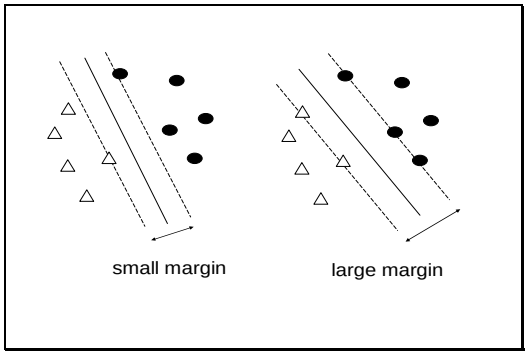


図 2　SVM におけるソフトマージンの概念

2.2.2 CRF

CRFは,系列ラベリング問題のために設計された識別モデル (discriminative model) であり、正しい系列ラベリングを他の全ラベリング候補と弁別するような学習を行う。通常の識別モデルとの違いは、出力が出力集合Y の部分集合ではなく、系列となる点にある。 CRFは、品詞付与、 テキストチャンキング、 固有表現抽出、HTML からの情報抽出、書誌データからの情報抽出といった系列ラベリング問題に適用され、いずれにおいても高い性能を示している。[5]

$\boldsymbol{x} = (x_1, x_2, \ldots, x_T), xt_2 \in \Sigma x$ を観測変数の集合とし、$\boldsymbol{y} = (y_1, y_2, \ldots, y_T), yt_2 \in \Sigma y$ を目的変

[2]http://chasen.org/~taku/software/yamcha/

数の集合とする。ただし、Σx と Σy はそれぞれ観測変数と目的変数が取りうる値の集合とする。ラベルを付与する構造データは、変数間の依存関係を示すグラフ構造で表現されているものと仮定する。グラフ構造の例を前節のシグナルペプチド、細胞質タンパク質および核タンパク質のデータセットの例を使って示す。図3は前節のアミノ酸配列に対するラベル付与の例である。

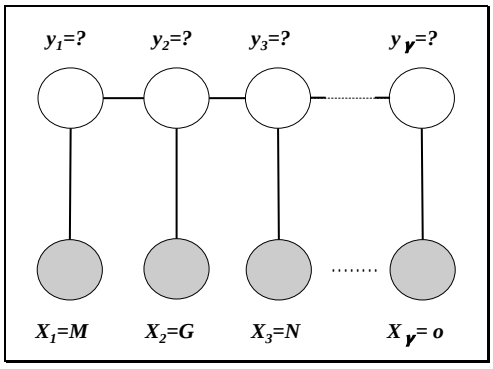


図3　アミノ酸残基列

CRF は条件付確率を直接表現する多クラスのロジスティック回帰の形をとる。

$$f(y \mid x) = \frac{\exp(< \Theta, \Phi(x, y) >)}{\sum_{\tilde{y}} \exp(< \Theta, \Phi(x, \tilde{y}) >)},$$

ただし、Θ はモデルのパラメータ、$\Phi(\boldsymbol{x}, \boldsymbol{y})$ は $(\boldsymbol{x}, \boldsymbol{y})$ に対する素性ベクトルで、$\Phi(\boldsymbol{x}, \boldsymbol{y})$ の要素 ϕi の値は i 番目の素性が $(\boldsymbol{x}, \boldsymbol{y})$ に現れた回数とする。図3 に示すように、素性は通常、連続する変数の組として定義される。図4***(a)*** の形の素性は観測素性、図4***(b)*** の形の素性は遷移素性と呼ばれる。ラベル付与は $\boldsymbol{x}$ に対する $\arg\max_{y} f(\boldsymbol{y}|\boldsymbol{x})$ による予測で行う。[6] この研究ではツールとしてCRF++[3]を用いる。

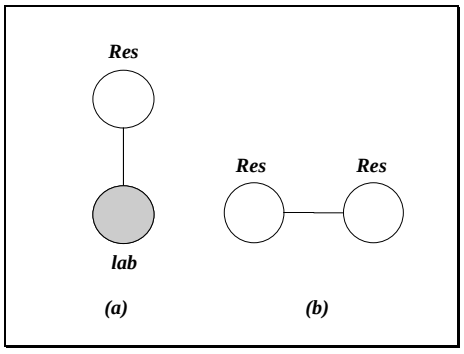


図 4　CRF で使用されている素性：***Res*** はアミノ酸残基、***lab*** はラベルを表す。

2.3　評価方法

評価方法として各生物種に対して five-fold cross validation(5CV)[4] を採用した。採用したデータセットをほぼ同じデータ量になるように 5 分割し、CRF と SVM 共にこれらのデータのうち４つをトレーニングデータ、残った１つをテストデータとして予測した。この過程は分割した 5 つのデータが 1 回ずつテストデータなるように繰り返し行った。

更にシステムの予測性能比較には Accuracy を用いた。以下にその定義を示す。

$$\text{Accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{fp} + \text{fn} + \text{tn}} \qquad (3)$$

ここで tp は true positive、fp は false positive、fn は false negative、tn は true negative を表す。

[3] http://www.chasen.org/~taku/software/CRF++/

[4] 5CV は先行研究と実験条件を合わせるために採用した。

3． 実験結果

表 2 に実験結果を示す。Euk は真核生物、gram+はグラム陽性細菌、gram-はグラム陰性細菌を表す。学習の素性として予測するアミノ酸と、その前後のアミノ酸配列を採用した。採用する前後の長さを window size と呼ぶ。本研究では window size として予測するアミノ酸残基を中心に±1 から 25 とした。

表 2　先行研究との Accuracy の比較

Method	Cleavage site (Accuracy %)		
	Euk	Gram－	Gram+
SignalP 1 NN	70.2	79.3	67.9
SignalP 2 NN	72.4	83.4	67.4
SignalP 2 HMM	69.5	81.4	64.5
SignalP 3 NN	79.0	**92.5**	**85.0**
SignalP 3 HMM	75.7	90.2	81.6
SVM	71.9	68.6	54.6
CRF	**79.1**	79.2	67.8

各生物種に対する CRF と SVM の window size の最適化を図 5 から図 7 に示す。先行研究での最適な window size は真核生物で－20 から+4、グラム陰性細菌で－11 から+3、グラム陽性細菌で－11 から+2 である。

図 7　Window size の最適化 (gram negative)

これに対し、SVM での最適な window size は、eukaryote では±18、gram positive では±21、gram negative では±24 であった。また、CRF での最適な window size は、eukaryote では±6、gram positive、gram negative 共に±3 であった。最適な window size での性能を表 2 に示す。これらの結果から我々が適用した機械学習法は先行研究の結果

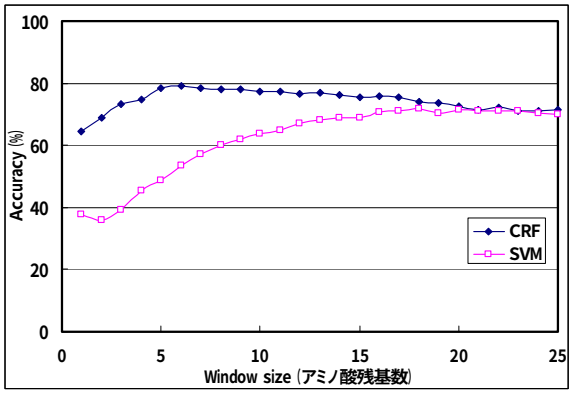


図 5　Window size の最適化 (eukaryote)

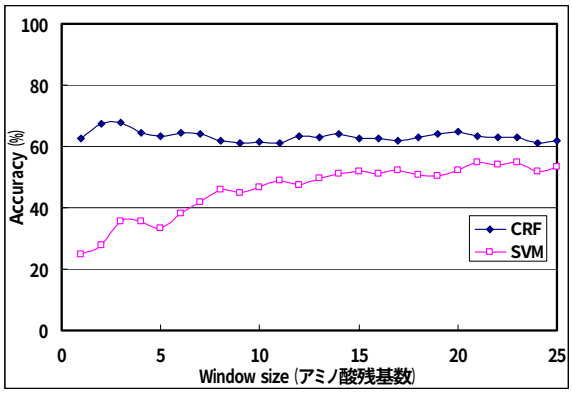
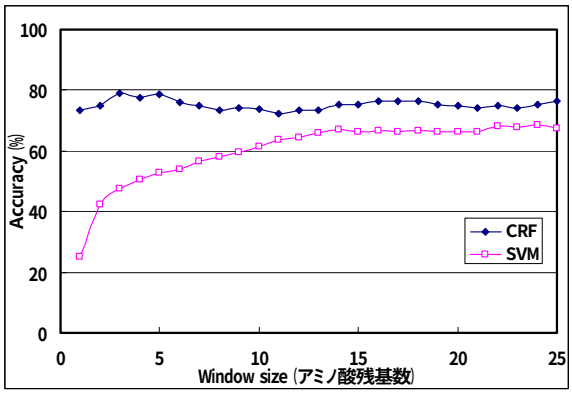


図 6　Window size の最適化 (gram positive)



と比べ、グラム陽性菌とグラム陰性菌は予測性能が及ばなかったが、真核生物の場合に限って性能がほぼ同等であることが分かった。

また、CRF は SVM よりシグナルペプチド開裂部分の予測に対して window size が小さい場合、性能が大きく上回っており、全体を見

ても CRF の方が性能を上回っていることが分かった。

4．考察とまとめ

実験結果より CRF を適用したシグナルペプチドの開裂部分予測において、真核生物が一番良い値となったのに対し、先行研究ではグラム陽性細菌の予測が一番良い値となっている。

先行研究と傾向が異なることから、生物種によって適切な機械学習法が異なることを示唆されていると思われる。また CRF は SVM に比べ、対称の window size の幅を小さくしても比較的高いシグナルペプチドの開裂部分予測ができると考えられる。

また今回の研究は Neilsen[3]らが提供した version 1 のデータセットを使用した。表 2 において、グラム陰性細菌とグラム陽性細菌において最高値を示した Bendtsen[1]らは version 3 のデータセットを用いている。Bendtsen ら[1]が文献中で述べているように、version 3 以前のデータセットにはアノテーションの間違いなど、多くの誤ったデータが含まれている version 3 はまだ公開されていないので今回は version 1 を用いたが、version 3 のデータセットを使うと今回の CRF, SVM の性能はさらに向上することが期待される。

5．おわりに

本稿では、CRFs と SVM ともに素性にアミノ酸配列のみを採用した。更に適当な素性を検索し、採用することで予測性能が向上するか確認する必要がある。また今回は window size として対称の window size を用いたが、先行研究では非対称の window size が最適とされている。非対称の window size を用いた CRF および SVM の予測も今後の検討課題である。

参 考 文 献

[1] Jannick Dyrlo Bendtsen, Henrik Nielsen, Gunnar von Heijne and Soren Brunak: Improved Prediction of Signal Peptides: SignalP 3.0: *J. Mol. Biol.* Vol.340, pp.783-795 (2004).

[2] Bairoch A. and Boeckmann B.: The SWISS-PROT protein sequence data bank: current status, *Nucleic Acids Res.* Vol.22, No.17 pp.3578-3580 (1994).

[3] Nielsen H., Engelbrecht J., von Heijne G. and Brunak S.: Defining a Similarity Threshold for a Functional Protein Sequence Pattern: The Signal Peptide Cleavage Site, *PROTIENS: Structure, Function, and Genetics* Vol.24, pp.165-177 (1996).

[4] Cristianini N. and Shawe-Taylor j.: An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods: Cambridge University Press. Cambridge, MA. (2000)

[5] Kudo, T, Yamamoto and Y, Matsumoto Y.: Applying Conditional Random Fields to Japanese Morphologiaical Analysis: 情報処理学会自然言語処理研究会 SIGNL-161 (2004)

[6] Tsuboi Y and Kashima H.: Design of Discriminative Models for Labeling Structured Data: *2005 Worlshop on Information-Based Induction Science (IBIS2005)*.