［特集］ポストゲノム時代に高まるバイオ自然言語処理への期待：バイオ自然言語処理最新事情

6 バイオ自然言語処理のための機械学習技術

平　博順
taira@cslab.kecl.ntt.co.jp /NTTコミュニケーション科学基礎研究所

前田 英作
maeda@cslab.kecl.ntt.co.jp / NTTコミュニケーション科学基礎研究所

情報更新が頻繁なデータを扱うための機械学習の技術

近年，医学生物学の分野では，ハイスループットの実験手法が開発されたこともあり，以前には考えられなかったような大量の実験結果が得られるようになってきている．これに伴い，その実験結果を報告する論文の出版も爆発的に増加し，論文のアブストラクトだけ読んで全体像を把握することさえもだんだん困難になりつつある．そこで，大量の文献の中から必要な情報だけを取り出す情報抽出技術は，医学生物学研究者の切実な要求となってきている．

テキストからの情報抽出技術では，抽出のためのテンプレートをあらかじめ人手で作成しておき，入力されたテキストに対しテンプレートマッチングを行い，情報を抽出する，という手法がよく用いられている．対象分野が狭い範囲に限定されている場合や，対象分野で用いられる用語，概念の時間的変動が少ない場合には，テンプレートマッチングによる手法で高精度の抽出が行えることも多い．しかし，現在の医学生物学分野のように，新しい事実，概念が次々と発見され，専門用語が急速に増加していっているような分野では，テンプレートマッチングによる抽出は，年々困難になっていくと考えられる．その理由としては，最新の専門用語に対応できる新たなテンプレートを記述できるほどの，分野に精通した人材の確保が難しいことが挙げられる．また，仮に新たなテンプレートが作成できたとしても，新たなテンプレートを既存のテンプレート群に追加する際，全体の抽出精度を保ったまま追加することは，副作用の問題があり，一般に人手では難しい，という問題もある．よって，将来的には情報抽出システム構築において，なるべく多くの部分が人手を介すことなく，自動で行われることが望ましい．その自動化のための１つのキーが機械学習技術である．

自然言語処理の研究では，すでに1990年代より，機械学習を用いた手法が，さまざまなタスクにおいて多く用いられている．機械学習を用いる際の利点には，対象となるデータの特徴が変化しても，すぐに対応できる保守性の良さや，専門知識を持った人が細かなチューニングを行わなくても，ある程度の精度が自動で得られる点などがある．最近，医学生物学分野のテキストに対する自然言語処理に関しても，学習を行うために必要なGENIAコーパス☆1等のコーパスが構築され，機械学習

☆1 http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/index.html

単語	A	nuclear	factor	from	both	peripheral	blood	monocyte	and
BIO タグ	O	B	I	O	O	B	I	I	O
固有表現クラス	その他	タンパク質名	タンパク質名	その他	その他	細胞種名	細胞種名	細胞種名	その他

表 -1　BIO タグの付与

を適用する環境も整いつつある．

本稿では，医学生物学文献に対する自然言語処理（バイオ自然言語処理）における代表的なタスクとして，固有表現抽出，タンパク質間関係抽出，遺伝子機能分類を取り上げ，これらのタスクで用いられている機械学習技術について紹介する．

固有表現抽出

バイオ自然言語処理における最も基礎的なタスクの1つに，固有表現抽出（Named Entity Extraction）がある．ここで，固有表現とは，遺伝子名，タンパク質名，細胞種名などの固有名詞や，測定値，実験時間などの数値表現などを指す．一般のテキストに対する固有表現抽出の研究は，Message Understanding Conference（MUC）や Information Retrieval and Extraction Exercise（IREX）などの会議で，評価用データセットに対するシステム構築を通して，発展を遂げてきた．これらの会議では，主にニュース記事中から，システムが人名，組織名，場所の名称などの文字列を特定することを目的としている．

近年，MUC等で培われた固有表現抽出技術を医学生物学分野にも適用しようと，さまざまな研究がなされている．しかし，これまでの固有表現抽出技術をそのまま使っても思うような高い精度が得られないことが分かってきている．その原因として，以下のような医学生物学分野独特の固有表現があることが指摘されている．

- 機能をそのまま説明的に記述したような固有表現
  （例）adenylate cyclase activating polypeptide 1
  （アデニル酸シクラーゼ活性化ポリペプチド1）
  （全体が1つのタンパク質の名前であると同時に，アデニル酸シクラーゼもタンパク質（酵素）の 名前になっている）
- 固有表現がand，orなどの接続詞で長くつながった表現の存在
  （例）alpha- and beta- globin
  （αグロビンとβグロビン）
- 短い略称名となっている固有表現
  （例）IL2
  （タンパク質Interleukin 2の略称名）
- 遺伝子が発見され名前が付与される速度が速く，遺伝子名辞書への登録が未登録の固有表現

そこで，医学生物学分野に特化した素性を学習に用いるなどの工夫がなされている．

■バイオ固有表現抽出問題の定式化

次に，固有表現抽出問題を機械学習を用いて解く際の定式化について述べる．固有表現抽出問題は，単語列が与えられたとき，単語列のどの部分が，どの固有表現クラスに属すかを識別する問題である．たとえば単語列として，

“A nuclear factor from both peripheral blood monocyte and T cell binds the peri-kappa B site.”
（末梢血単球とT細胞両方から由来する核因子がDNA上のペリκBサイトに結合する．）

が与えられたとき，“nuclear factor”がタンパク質ファミリ（タンパク質の類似構造による分類）名，“peripheral blood monocyte”および“T cell”が細胞種名，“peri-kappa B site”がDNAドメイン（DNA上の特徴的な配列のある場所）名であることを識別する．ここで，「タンパク質ファミリ名」「細胞種名」「DNAドメイン名」は，固有表現の種類であり，固有表現クラス，と呼ばれる．

一般に1つの固有表現は複数の単語から構成されるため，そのままでは，SVMなどの分類学習手法を適用できない．そこで，BIOタグと呼ばれる固有表現境界タグを用いて，問題を分類問題に変換することがよく行われる．BIOタグには，いくつかの方式があるが，ここでは，IOB1と呼ばれる方式を例に説明する．IOB1方式では，固有表現の先頭の単語をB（Beginの意），固有表現中の単語をI（Inの意），固有表現でない単語をO（Otherの意）で表す．上記の“A nuclear factor from both peripheral blood monocyte and ...”という単語列に対し，BIOタグを付与した例を，**表-1**に示す．

このようにBIOタグを用いることによって，固有表現抽出問題は，与えられた単語列中の各単語を，固有表現クラス（NC）とBIOタグ（BIO）の組，すなわち（NC, BIO）に分類する問題として捉えることができる．たとえば，表-1における，nuclearという単語は，クラス（タンパク質名，B）に，bloodは，クラス（細胞種名，I）に分類できればよい．

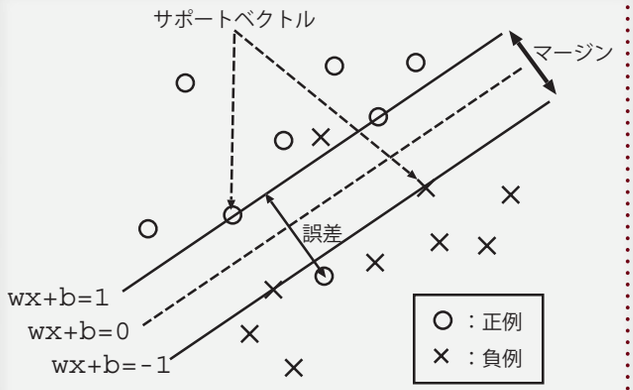

図-1　SVMの概念図

学習には，コーパスなどから，d文からなる訓練データ，

$$D = (\mathbf{x}_1^n, y_1^n)_{(1)}, \ldots, (\mathbf{x}_1^n, y_1^n)_{(d)}$$

を用意しておき，分類学習を行う．ここで，$\mathbf{x}_1^n$は，系列$\mathbf{x}_1,\ldots,\mathbf{x}_n$を省略して書いたもので，$\mathbf{x}_i$は，文中の$i$番目の単語$t_i$の持つ性質を数値化したベクトルである．単語の持つ1つ1つの性質は素性と呼ばれる．同様に，y_1^nは，$y_1,\ldots,y_n$を省略して書いたもので，y_iは，単語t_iに対する，分類すべき固有表現クラスとBIOタグとの組（NC, BIO）である．

こうして分類学習が行われた分類器に，未知のテストデータ$\mathbf{x}_1^n$を入力し，固有表現クラスとBIOタグとの組の列y_1^nを予測する．

■バイオ固有表現抽出の学習で用いられる主な素性

ここで，バイオ固有表現抽出の学習の際，よく用いられている素性を紹介する．

- 単語素性
  注目している単語，もしくはその単語の前後数単語が，ある単語である（1）か否か（0）の数十万次元程度の素性．
- 形態素素性
  注目している単語が，ある接頭辞（hydro-など）や接尾辞（-teinなど）を含む（1）か否か（0）を表す素性．
- 品詞素性
  注目している単語，もしくはその単語の前後数単語の品詞が，ある品詞である（1）か否か（0）を表す素性．
- 単語形の素性
  英字大文字をX，英字小文字をx，数字をdで表すなどして，単語が“T-cells”ならばX−xxxxx，“IL2”ならばXXdと表すなどして，注目している単語がある形である（1）か否か（0）を表す素性．
- 外部情報源中の登録状態の素性
  注目している単語が，タンパク質データベースSwiss-Prot☆2など外部情報源に登録されている固有表現である（1）か否か（0）を表す素性．

これらの素性を合わせて，全体で数十万から数百万次元のベクトルを用いて学習を行う．

■バイオ固有表現抽出に用いられている機械学習手法

ニュース記事からの固有表現抽出の場合と同様に，最近では，さまざまな機械学習手法が試みられている．ここでは，そのうち，サポートベクタマシン，隠れマルコフモデル，条件付確率場について説明する．

サポートベクタマシンの利用

サポートベクタマシン（support vector machine; SVM）は，高次元空間内で，訓練データを正例と負例とに分け，かつ，正負例間のマージンが最大になるような超平面を求める機械学習手法である[12]．**図-1**にSVMの概念図を示す．

最も負例よりの正例側の境界面と，最も正例よりの負例側の境界面の間の距離をマージン（margin）と呼ぶ．このマージンが最大となるような超平面を求め，$\mathbf{w}\cdot\mathbf{x}+b=0$を最終的な分類境界面とする．ただし，完全に線形分離できない場合には誤差も考慮に入れて（ソフトマージン），分離境界面を決定する．

バイオ固有表現抽出にSVMを適用する場合は，一般に1つの$\mathbf{x}_i$ごとにy_iを求める．山田ら[3]の研究では，各分類クラスごとに注目している分類クラスを正例，その他のクラスを負例としたone vs restと呼ばれるやり方で，クラス数個の分類器を学習し，分類境界面から正例側に最も遠く分類されたクラスを最終的な分類クラスとしている．y_iを系列の先頭から求めていく際，分類結果によっては，（タンパク質名，B）の次に（細胞種名，I）が来てしまうといった，禁止されているBIOタグの系列を得てしまうケースがある．このようなケースを回避した結果を得るために，y_iの各候補に対する分類確率を各々求めておき，禁止された系列は除いて，Viterbiアルゴリズムなどで，系列全体として最ももっともらしい系列を求める，という方法もよく用いられる．

ただし，SVMでは確率値は得られないため，SVMの

☆2 http://us.expasy.org/sprot/

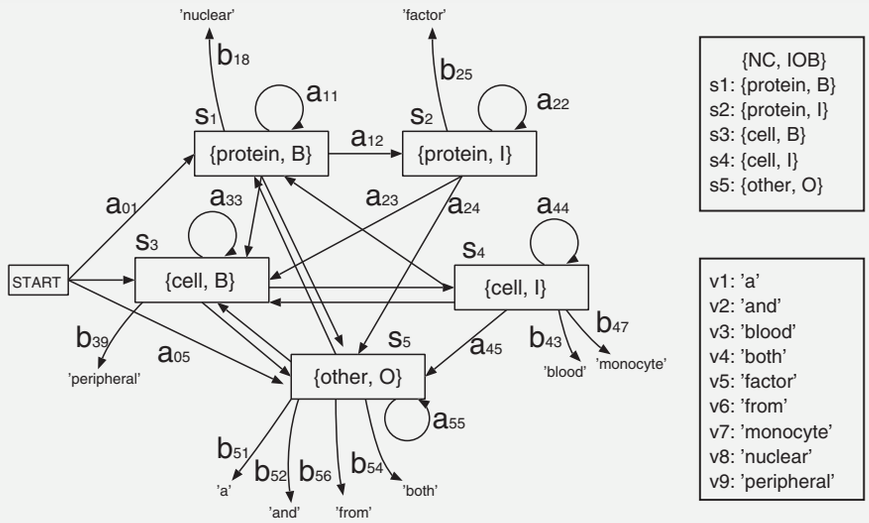

図 -2　HMM の例

i	1	2	3	4	5	6	7	8	9
可視シンボル系列 $\{x_i\}$	v_1	v_8	v_5	v_6	v_4	v_9	v_3	v_7	v_2
状態系列 $\{y_i\}$	s_5	s_1	s_2	s_5	s_5	s_3	s_4	s_4	s_5

表 -2　可視シンボル系列と状態系列の例

出力を近似的に確率値に変換する方法として，Plattによるシグモイド関数を用いる方法[8]がしばしば使われている．これは，SVMの出力値$f(\mathbf{x}) \equiv \mathbf{w}\cdot\mathbf{x} + b$に対し，

$$p(y|\mathbf{x}) = \frac{1}{1 + \exp(Af(\mathbf{x}) + B)} \quad (1)$$

で確率値を与える方法である．ここで，AおよびBは，訓練データから決定されるパラメータである．

隠れマルコフモデルの利用

隠れマルコフモデル（hidden Markov model; HMM）は，モデルが状態sの系列

$y_1^n = y_1, y_2, ..., y_i, ..., y_n$ $(y_i \in s = \{s_1, s_2, ..., s_j, ...\})$

をマルコフ性（各状態が，前の時刻での状態に依存して決まる）を持って生成するが，その状態の系列を外部から観測できず，各状態jが確率b_{jk}で出力した可視シンボルv_kの系列

$x_1^n = x_1, x_2, ..., x_i, ..., x_n$ $(x_i \in v = \{v_1, v_2, ..., v_k, ...\})$

しか外部から観測できない，というモデルである[9]．**図 -2**にHMMの例，**表 -2**に可視シンボル系列と状態系列の例を示す．

HMMを用いたバイオ固有表現抽出では，可視シンボルx_iを，単語tの性質を数値ベクトル化した素性ベクトルと考える．また，モデルの隠れ状態sは，固有表現クラス（NC）とIOB1などの固有表現境界タグ（BIO）との組，すなわちs =（NC, BIO）とする．

最も単純には，訓練データとして，

$D = (x_1^n, y_1^n)_{(1)}, ..., (x_1^n, y_1^n)_{(d)}$

を用意しておき，前向き・後向きアルゴリズムなどを用いて，状態iから状態jへの遷移確率a_{ij}と，状態jにおけるシンボルv_kの出力確率b_{jk}を推定する．そして，状態遷移が未知のテストデータに対して，可視シンボルv_kの系列 $x_1^n = x_1, x_2, ..., x_i, ..., x_n$ $(x_i \in v = \{v_1, v_2, ..., v_k, ...\})$を最も出力しやすい状態遷移，すなわち$P(y_1^n | x_1^n)$が最大となるような状態遷移$y_1^n$を求めることで，最適な固有表現クラス（NC）と固有表現境界タグ（BIO）とを求める．

ただし，HMMを用いたバイオ固有表現抽出では，訓練データの量が十分でないケースが多いので，実際の応用では，モデルの訓練を精緻化するために，以前の状態で現れた単語なども状態に含める，といったさまざまな工夫がなされている．

条件付確率場の利用

最近，系列の学習で高精度の結果を得る方法として条件付確率場（Conditional Random Field; CRF）が注目されており[6]，バイオ固有表現抽出においても，高い精度の抽出を行えることが示されている[7]．

CRFでは，可視シンボル系列$x_1^n = x_1, x_2, ..., x_n$に対応

する状態遷移を$y_1^n = y_1, y_2, ..., y_n$としたとき，各$y_i (i = 1, ..., n)$が，$x_1^n$と$y_i$近隣の$y_j (i \neq j)$から決まる値であると仮定し，観測列が与えられた時の状態系列の確率を以下のように定義する．

$$p(y_1^n | x_1^n) = \frac{1}{Z_0} \exp\left(\sum_{i=1}^{n} \sum_{j=1}^{m} \lambda_j f_j(y_{i-1}, y_i, x_1^n, i) \right) \quad (2)$$

ここで，Z_0は，正規化のための数で，全状態系列に対する$\exp\left(\sum_{i=1}^{n} \sum_{j=1}^{m} \lambda_j f_j(y_{i-1}, y_i, x_1^n, i) \right)$の和である．

$f_j(y_{i-1}, y_i, x_1^n, i)$は，前状態が$y_{i-1}$で表される特徴を持つ単語で，現状態が$y_i$で表される特徴を持つ単語が入力され，可視シンボル系列がx_1^nであったとき１，それ以外の場合０の値を持つ関数である．λ_jは各素性関数f_jに対応する重みで，この重みについて学習が行われる．素性jが状態に対し正の相関を持つとき，λ_jは正の値，負の相関を持つとき，負の値を取り，無相関の場合にはλ_jは０に近い値を取る．mは全素性数である．

重みの学習は，訓練事例$(x_1^n, y_1^n)_{(k)} (k = 1, ..., d)$における，可視シンボル列の条件付対数尤度

$$LL = \sum_{k=1}^{d} \ln P(y_{1(k)}^n | x_{1(k)}^n) - \sum_{j=1}^{m} \frac{\lambda_j^2}{2\sigma^2}$$

を最大にするようなλ_jを見つけることで行われる．ここで，第２項は，Gaussian Priorと呼ばれる過学習を回避するための項で，σ^2はλ_jに関する分散である．未知データに対しては，Viterbiアルゴリズムなどを用いて，最も尤もらしい状態遷移y_1^nを求める．

■ タンパク質等IDの自動付与

次章で述べるタンパク質間関係抽出などの高レイヤの処理に固有表現抽出を利用するには，実は，タンパク質や遺伝子のデータベース上でのIDまで特定しておく必要があることが多い．たとえば，タンパク質間関係抽出においては，“Interleukin 2”や“IL-2”がタンパク質名であることが分かっただけでは不十分で，どちらも同じタンパク質（タンパク質データベースSwissProt中ではヒト細胞の場合 ID:P60568）であることを特定する必要がある．

タンパク質や遺伝子の略称名などは，同じ名前でも異なるタンパク質を指す曖昧性が見られることや，名前の表現が文献ごとに微妙に異なることがあるため，単純に既存のデータベースが持っているIDへの対応付け辞書を用意するだけではこの問題は解決できない．そこで，鶴岡らは，タンパク質辞書に登録されたタンパク質名に対して高速な近似文字列探索を行ったあと，グローバルな文脈を学習させた分類器でフィルタリングする方法により，高精度なタンパク質IDの自動付与を実現している[11]．

タンパク質間関係抽出

医学生物学文献を対象とした情報抽出における高いレイヤのタスクに，タンパク質間関係抽出がある．タンパク質は生体内で酵素として働き，通常起こりにくい化学反応を緩やかな条件のもとで進行させる．これらの酵素などが複雑に関連してマクロな生体の機能を実現しているため，タンパク質間の関係についての情報は，非常に重要である．膨大な医学生物学文献の中から自動的にタンパク質間関係についての情報を得ることができれば，医学生物学の研究スピードが上がることが期待できる．

たとえば，

“hEpoR transcription activity depends on coordination between Sp1 and GATA-1.”

（タンパク質hEpoRの転写活性はタンパク質Sp1とタンパク質GATA-1との間の協調に依存する．）

のような文が与えられたとき，ここから“hEpoR”と“Sp1”と“GATA-1”がタンパク質の名前であり，「タンパク質 hEpoRの転写活性」が，「タンパク質 Sp1」と「タンパク質 GATA-1」が協調して「タンパク質 hEpoR」を転写活性させる，という関係を抽出することが課題となる．

このような関係についての情報抽出は，人手でパターンを書き，そのパターンに従って情報抽出する方式が多いが，少数ながら学習手法を用いた研究もある．たとえば，Cravenらは，２つの手法を試みている[2]．

１つは，抽出対象の文に現れる単語群（bag-of-words）を素性とした学習が行われた分類器で，抽出したい関係を述べている文を特定した後，抽出したい関係を構成するオブジェクトが辞書に含まれる場合に，関係があるとして抽出する方法である．

もう１つは，抽出対象の文に対して構文解析を行って構文木を作成したあと，「タンパク質名を含む句と場所名を含む句が１つの句を挟んで出現する」などの背景知識を設定した上で「タンパク質Ａが細胞内の場所Ｂに存在する」などの関係の学習を行う方法である．

またAlphonseらは，背景知識を利用しながら帰納推論を行うことができる，帰納的論理プログラミング（Inductive Logic Programming; ILP）の１つであるPropalアルゴリズムを用いて，関係の学習を行う方法を提案している[1]．

これらの手法は，あらかじめ関係の定義を人手で記述する必要があり，まだまだ低コストの手法とは言いがたい．今後，関係の定義について，記述量が少なくても学習できるような枠組みが求められる．

遺伝子機能分類

現代の生物学は研究分野が細分化され，研究が進められている．ある分野の生物学者が他の分野の生物学的知識を得ようとする際，多様な語彙が同じものを指していて，情報抽出が困難であることが問題となっている．この問題の1つの解決策として，生物学における用語の標準化が進められている．有名なものでは，Rileyらによる大腸菌遺伝子の機能に関する用語分類，ミュンヘン・タンパク質配列情報センタによる機能用語分類（MIPS）☆3，Gene Ontology コンソーシアムによる複数の生物をカバーする機能用語の分類（GO）☆4などがある．これらの標準化作業も，専門家が，各遺伝子に関係する論文や実験データなどを調査した上で行っているため，非常にコストがかかっている．また，前述のように，年々，医学生物学の情報は増え，それに伴い新しい概念や分野が生まれるため，付与される情報も逐次，更新される必要がある．そのため，機械学習手法を用いて，これらの機能用語を遺伝子に対して自動で機能分類を付与する技術の進歩が期待されている．

これらの，機能用語の分類体系のうち，GOの分類を機械学習手法を用いて付与した例には，SVMや最大エントロピー法を用いた研究があるが，これらは，従来の枠組，つまり，各GO分類について，その分類クラスがその遺伝子に当てはまるか否かを学習し，分類するものである[10),4)]．しかし，一般に，遺伝子は複数の機能を持っており，また，分類クラスが相互に関連し合っているケースも多い．そこで，それらの複数クラスの関係も考慮することにより高精度の分類を実現した，最大マージン原理にもとづく多重トピック分類（MML）によるGOの分類も最近行われている[5)]．MMLでは，SVMと同様のマージンを最大化する考え方に基づいて学習を行うが，複数クラスの組を1つのクラスと見なして，最大マージンの学習を行う点がSVMと異なっており，MMLではSVMなどの他の分類学習手法を上回る高い分類精度が得られている．なお，これらの手法では，素性として，主に目的とする遺伝子が掲載されていたテキストのbag-of-wordsが使われている．

☆3 http://mips.gsf.de/projects/funcat/
☆4 http://www.geneontology.org/

より複雑な構造を扱う機械学習へ

バイオ固有表現抽出や，機能分類に関しては機械学習の最新手法が数多く用いられているが，深い解析や構造を扱うような学習については，まだ十分に研究されているとは言えない．また，関係抽出への学習についても，労力の低減という意味で十分なレベルには達していない．さらに本稿では触れなかったが，大規模な実データを対象としたときの処理の高速化についての研究もこれから活発に行われる必要がある．医学生物学文献が爆発的に増える中，ますます，機械学習への期待は高まるであろう．より一層の大規模な処理，高精度化が実現できる機械学習手法の登場が期待されている．

参考文献

1) Alphonse, E.: Event-based Information Extraction for the Biomedical Domain, Proc. of Coling-2004 International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications (NLPBA/BioNLP 2004), pp.43-46 (2004).
2) Craven, M. and Kumlien, J.: Constructing Biological Knowledge Bases by Extracting Information from Text Sources, Proc. of ISMB-1999, pp.77-86 (1999).
3) 山田寛康，工藤　拓，松本裕治：単語の部分文字列を考慮した専門用語抽出と分類，情報処理学会研究報告2000-NL-140, pp.77-84 (2000).
4) Izumitani, T, Hideto, H., Kazawa, T. and Maeda, E.: Assigning Gene Ontology (GO) Codes to Yeast Genes using Text-based Super-vised Learning Methods, Proc. of IEEE Bioinformatics Conference (CSB-2004) (2004).
5) Kazawa, H., Izumitani, T, Taira, H. and Maeda, E.: Gene Category Prediction by Support Vector Multi-learning Machinesd, Proc. of NIPS-2003 Workshop on New Problems and Methods in Bioinformatics (2003).
6) Lafferty, J., McCallum, A. and Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, Proc. of the 18th International Conference on Machine Learning (ICML '2001).
7) McDonald, R. and Pereira, F.: Identifying Gene and Protein Mentions in Text using Conditional Random Fields., Proc. of BioCreAtIvE: Cretical Assessment for Information Extraction in Biology (2004).
8) Platt, J.: Probabilistic Outputs for Support Vectior Machines and Comparisons to Regularized Likelihood Methods, Advances in Large Margin Classifiers (1999).
9) Rabiner, L.R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Reconginition, Proc. of the IEEE, Vol.77, No.2, pp.257-286 (1989).
10) Raychaudhuri, S., Chang, J.T., Sutphin, P.D. and Altman, R.B.: Associating Genes with Gene Ontology Codes using a Maximum Entropy Analysis of Biomedical Literature, Genome Research, Vol.12, No.1, pp.203-214 (2002).
11) Tsuruoka, Y. and Tsujii, J.: Boosting Precision and Recall of Dictionary-based Protein Name Recognition, Proc. of ACL '2003 Workshop on Natural Language Processing in Biomedicine, pp.41-48 (2003).
12) Vapnik, V.N.: The Nature of Statistical Learning Theory, Springer-Verlag (1995).

（平成17年1月10日受付）