

## Multiple-Instance 学習によるタンパク質相互作用の属性予測

山川 宏<sup>†</sup> 仲尾 由雄<sup>†</sup> 丸橋 弘治<sup>†</sup>

<sup>†</sup>(株)富士通研究所 IT コア研究所 〒211-8588 神奈川県川崎市中原区上小田中 4-1-1

E-mail: †{ymkw,ynakao,maruhashi.koji}@jp.fujitsu.com

あらまし 我々は、未知のタンパク質ペアに対して特定の相互作用属性の有無を予測する機械学習技術を開発し、この技術を利用して、タンパク質毎の属性 (Gene Ontology) からリン酸化属性の有無を判別した。複数のサブユニットを含むタンパク質間の相互作用が、組合せとして含まれるサブユニットペアの何れかに還元できると仮定することで、判別課題を Multiple-Instance 学習 (MIL) として定式化した。そして、よく知られている Diverse Density 法を改良し、特徴空間内で負例バッグに比べて正例バッグの寄与が大きな部分領域を概念として獲得する多数決型の MIL 手法を提案した。KEGG データベースから相互作用属性を学習データとして計算機実験を行なうと、86.1%の正解率が得られた。さらに偽陽性の発生原因を解析し、これが MIL 特有の陽性判定へのバイアスに起因することを示した。キーワード タンパク質相互作用, Multiple-Instance 学習, Diverse Density, リン酸化, サブユニット

## Predicting type of protein-protein interaction as a multiple-instance learning problem

Hiroshi YAMAKAWA<sup>†</sup>, Yoshio NAKAO<sup>†</sup>, and Koji MARUHASHI<sup>†</sup>

<sup>†</sup>FUJITSU LABORATORIES LTD. 1-1, Kamikodanaka 4-chome, Nakahara-ku, Kawasaki, Kanagawa, 211-8588 Japan

E-mail: †{ymkw,ynakao,maruhashi.koji}@jp.fujitsu.com

**Abstract** We propose a method for predicting types of protein-protein interactions using a multiple-instance learning (MIL) model. By assuming cause of each interaction between proteins containing two or more subunit is results from local subunit pairs, we formulate this problem as MIL. In this problem, influences from instances in negative bag to target concepts are too strong by using well-known Maron's method. We propose a new MIL method based on decision by majority and apply to the KEGG interaction data. In an experiment using the KEGG pathways and the Gene Ontology, the method successfully predicted an interaction type (phosphorylation) at the accuracy rate of 86.1%. We find that cause of false positive is caused by positive bias peculiar to MIL method by analyzing prediction results.

**Key words** protein-protein interaction, multiple-instance learning, diverse density, phosphorylation, subunit

### 1. はじめに

バイオインフォマティクス分野では、ドライ技術とウェット技術の融合による生体内の分子生物学機構の解明を目指す。細胞内での分子機構の多くがタンパク質間相互作用を基盤とし、例えば、ある疾患の原因となるタンパク質と、その治療に役立つようなタンパク質を関連づける等といった現実的な課題において、タンパク質間相互作用を知ることは重要である。現状では未知のタンパク質相互作用が多数存在し、不足した情報をドライ技術を用いて補完することが有用である。これを用いて、未知相互作用を推定しながらの仮説構築を支援などができる。

相互作用に関わる予測技術の先行研究として、たとえば、[1]ではオルソログ、タンパク質発現、Gene Ontology(以下 GO)、ドメインを使って相互作用を推定し、[2]では機能カテゴリ、細胞局在、ネットワークから相互作用を検証した。しかし現状では、相互作用の方向やタイプ(活性化、リン酸化、抑制など)といった相互作用属性の予測技術は見あたらない。

タンパク質相互作用を蓄積状況を見回すと、相互作用属性に関する情報は一般的にこの種のデータベース上に蓄積されていない[5]。実際、代表的なデータベースである Human Protein Reference Database[3]では2005年9月時点で33,710個のデータを蓄積しており、その数はさらに増え続けているが、相

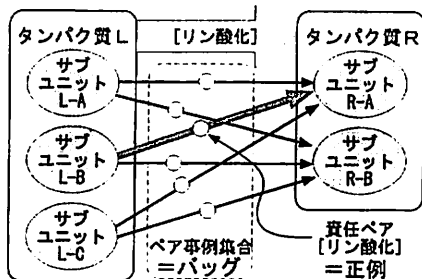


図1 タンパク質ペアの相互作用とバッグモデル  
サブユニットの組合せであるペア事例の少なくとも一つがリン酸化(正例)であれば、全体としてリン酸化(正例)であるとする。

相互作用性は付与されていない。逆に,[4]のデータには相互作用属性が記述されているが、登録されている相互作用の件数は数千件程度で小規模なものである。

以上のような背景から、与えられたタンパク質ペアに対して相互作用属性を予測する技術は有用である。そこで、本研究では、学習データとして既知の相互作用属性([4]の一部)を用い、タンパク質毎の属性(GO等)を入力とし相互作用属性の予測を出力する技術を提案する。

### 1.1 タンパク質ペアにおける相互作用の構造

タンパク質は図1に示すようにしばしば複数のサブユニットの複合体として形成され、詳細な属性は個々のサブユニット毎に得ることが可能である。一方で、学習データにおいて得られる相互作用属性はタンパク質ペア毎である。

2章の議論と関連するが、学習技術の観点では、予測すべき変数であるリン酸化属性の有無と、入力変数との間にギャップがあり、通常の教師有り学習として定式化できない課題がある。

そこで、本研究では、「局所相互作用仮定」を導入することで、相互作用をモデル化する。この仮定は、タンパク質ペアにおける、リン酸化などの相互作用属性は、直接的には各タンパク質に含まれるサブユニット間(責任ペア)の相互作用に還元できるという仮定である。これより、二つのタンパク質に含まれるサブユニットの組み合わせの一部が持つ相互作用属性は全体としての相互作用属性になるとモデル化される。

局所相互作用仮定の妥当性について議論する。タンパク質複合体ペア間の相互作用は複数の部位が関連する場合があり、それらが必ず一つのサブユニット上に存在するとは言えない。しかし、現在一般的な分子生物学のアプローチを考慮すると、相互作用の要因を1つのサブユニットペアに還元するという近似には一定の有用性があると考え得る。つまり、分子生物学のアプローチでは、一般に、疾患原因と考えられる1つのタンパク質機能に着目し、ノックアウト実験等によりその機能を変化させた場合の病態の変化を観察し、そのタンパク質が疾患原因であるという仮説を検証する。これは、疾患原因を1つのタンパク質の機能に帰着できることを暗黙に仮定していると考えられ、タンパク質機能が「サブユニット間の関係」と対応する多くの場合には、上記「局所相互作用仮定」と同様の仮定が一般的で

あることを示唆している。上記議論を勘案すれば、本仮定の基礎のデータ解析には、一定の有用性があるといえる。

なお相互作用属性予測は、タンパク質間のレベルではシステム全体の振る舞い理解に、サブユニット間のレベルでは創薬支援等における基礎情報として、それぞれに有用である。

## 2. MILとその改良手法の提案

相互作用毎にタンパク質のペア事例集合が含まれる(図1参照)。「局所相互作用仮定」に基づいて相互作用をモデル化したので、ペア事例集合はバッグとみなせ、その中の少なくとも一つが正例であれば、バッグを正例とみなせる。

すると、相互作用属性予測課題は、本章で説明する Multiple-Instance 学習(MIL)として定式化できる。

### 2.1 Multiple-Instance 学習(MIL)とは

MILは、学習データに対するラベル付けが不完全な教師有り学習である。通常の教師有り学習では、学習データ内の全事例に正解のクラスラベルや実数値が割り当てられるが、MILではバッグ(任意の数の事例集合)にラベル等が与えられるのみで個々の事例にラベル等は与えられない。

MIL課題の目標は、バッグがラベル付けされた学習データを用いて、未知のバッグと事例をラベル付けすることである。二値分類の場合には、バッグ内の少なくとも1つが正例であればバッグは正例で、バッグ内のすべての事例が不例なら、バッグは負例とみなす。

- 正例バッグ: 少なくとも1つの正例を含む場合
- 負例バッグ: その中の事例がすべて負例の場合

MILの初期研究では,[6]が薬物の分子活性に対する予測課題に用いた。その後多くのMIL手法が提案され、画像分類[7],学習やテキスト分類,株式市場予測,動画の顔ラベリング[8],Webマイニングへの利用[9]など多くの分野で適用されている。

### 2.2 標準的なMIL手法(Diverse Density)

本研究は,[7]による、MILでよく知られた Diverse Density( $d_D$ )を用いる手法を基盤とする。

このアプローチでは、正例バッグと負例バッグの集合が与えられたとき、特徴空間内で、全ての負例バッグ内の全ての事例から遠く、正例バッグ内の少なくとも1つの事例に近い部分領域を概念として獲得する。さらに、獲得すべき概念は、正例バッグに含まれる事例の密度が高いだけでなく、多様な正例バッグからの寄与を含む部分領域である。

形式的には、特徴量空間中の座標  $x$  における Diverse Density( $d_D(x)$ )は、与えられた  $i$  番目の正例バッグ  $B_i^+$  と、負例バッグ  $B_j^-$  からの寄与の積として定義される。

$$d_D(x) \propto \prod_i \Pr(x|B_i^{\pm}) \quad (1)$$

ただし、所与のバッグ  $B_j^{\pm}$  には各々に任意の数の事例  $B_{ij}^{\pm}$  を含んでいるとする。(ここで、 $j$  はバッグ内の事例インデックス)

事例毎の評価の最大値  $d_D(B_X) = \max_m d_D(x^{(m)})$  を、バッグ  $B_X = \{x^{(m)}\} (m \in M)$  に対する評価とする(ここで、 $x^{(m)}$  はバッグ内の任意の数の事例)。そして、この値が大きければ

バッグが正例であると判定する。

一方、各バッグ  $i$  の寄与は、noisy-or 型の関数で評価する。

$$\Pr(x|B_i^+) = 1 - \prod_j (1 - \Pr(x|B_{ij}^+))$$

$$\Pr(x|B_i^-) = \prod_j (1 - \Pr(x|B_{ij}^-)) \quad (2)$$

そして、各事例  $j$  の寄与はガウス分布  $\Pr(x|B_{ij}^\pm) = \exp[-\sum_k s(B_{ijk} - x_k)^2]$  で評価する。ここで、 $k$  は次元のインデックス、 $s$  はスケールファクターである。

### 2.3 重み付き多数決としての改良 Diverse Density

Diverse Density( $d_D(x)$ )の値は、座標  $x$  の近傍の一つでも負例バッグの事例が存在すると強烈に抑制される。一方、本研究が扱う課題は、特徴量空間内の多くの範囲で負例バッグの事例が散在し、このままでは負例の影響が強すぎた。

本研究では、周辺に若干の負例バッグの事例が存在しても、正例バッグの事例が多数存在する部分領域ではポジティブな方向で評価したい。そこで、特徴量空間上のある座標  $x$  における評価を、周辺バッグからの重み付き多数決とした。形式的には、前記  $d_D(x)$  において積としていたバッグからの寄与を、和に変更した改良 Diverse Density( $\tilde{d}_D(x)$ )を新たに定義する。

$$\tilde{d}_D(x) \propto \sum_i \text{sign}_i \left[ 1 - \prod_j (1 - \Pr(x|B_{ij}^\pm)) \right] \quad (3)$$

(ここで  $\text{sign}_i$  はバッグ  $i$  が正例のとき +、負例のとき -。<sup>(注1)</sup>  $x^{(m)}$  を事例として含む任意のバッグ  $B_x = \{x^{(m)}\} (m \in M)$  に対する評価は、バッグ内事例の少なくとも一つが正 ( $\max_m \tilde{d}_D(x^{(m)}) > 0$ ) であればバッグが正例であるとする。

## 3. 相互作用属性と Gene Ontology

本研究では、相互作用属性を KEGG から取得し、個々のタンパク質の属性として GO を用いた。

### 3.1 KEGG における相互作用属性の出現状況

KEGG には細胞レベルでの生命システムの機能に関する知識が、分子間相互作用ネットワークの情報として蓄積されている。データ取得時において、1279 件のタンパク質間相互作用 (PPrel) と、2910 件の Enzyme-enzyme relation (ECrel) が含まれていた。

相互作用情報の入手は、XML 形式のファイルをダウンロードし<sup>(注2)</sup>、そこから相互作用属性を含むタンパク質間相互作用 (`<relation>` タグで記述) と、タンパク質 (`<entry>` タグで記述) を取り出した。なお、`<entry>` タグで記述された各タンパク質には、複数のサブユニットが含まれている。

次に、タンパク質間相互作用 (PPrel) について、相互作用属性の出現状況を調べた (表 1 参照)。ここで、相互作用属性として付与されるラベルは 11 種類である。一つの相互作用に複数の属性が付与されるが、その分布は著しく偏っている。例えば、activate & inhibit といった組み合わせは存在せず、17 種類の相互作用属性の組合せ (排他的カテゴリ) が存在する。

表 1 KEGG から得られた相互作用属性の分布 (PPrel) 予測実験 (4 章) で、正例に用いた排他的カテゴリはピンク、負例に用いたのは青で表す。

排他的カテゴリ	compound	hidden compound	activation	inhibition	binding association	phosphorylation	prohibition	dissociation	dephosphorylation	ubiquitination	starts	合計
1												4
2												7
3												8
4												13
5												145
6												8
7												8
8												150
9												1
10												41
11												151
12												97
13												5
14												525
15												11
16												52
17												27
合計	27	0	102	181	249	188	14	23	22	4		1279

### 3.2 タンパク質属性としての GO の利用

タンパク質毎の属性には、ゲノム分野のスタンダードなオントロジとも呼べる [10](GO) を利用する。GO のトップレベル直下には、生物学的プロセス、分子機能、局在位置の 3 カテゴリがある。さらに GO とタンパク質を結びつける情報として (gene2go) を利用する。なお、GO 階層性をもつので、GO ターム毎に見て推移律が成立する先祖属性を追加して利用する。以上で、個々の相互作用情報 (バッグ) 単位に、ペアとなるタンパク質毎の GO タームと、その間の相互作用属性が整えられた。

### 4. リン酸化属性の有無判別実験の設定

Multiple-Instance 学習 (MIL) 手法を用い、タンパク質相互作用についてのリン酸化属性の有無の判別実験を行う。このあと問題設定を明確化し、次に GO タームを特異値分解 (SVD) で圧縮した上位次元を用い、改良 Diverse Density による MIL 手法により判別予測実験を行う。

計算機実験には、計算ソフトウェアである MATLAB を利用した。これは、疎行列の性質を利用して大規模な行列に対応できる特異値分解 (SVD) ツールが利用可能なためである。

#### 4.1 リン酸化属性の有無についての分類課題

前章で述べたように、相互作用属性データには、従属変数として設定しうる属性が複数有り、排他的なカテゴリに着目した分類課題なども設定しうる。

ここでは、リン酸化属性の有無という二値分類課題を扱う。これは予備実験において、リン酸化属性の有無に着目すれば、GO タームを手がかりとしてある程度良い性能が得られる可能性が示されていたためである。さらに、独立変数として利用しうる、GO タームの 3 カテゴリの中から、分子機能のタームのみを利用した。これも予備実験からの考察による。

次に、現状の不完全なデータ蓄積状況を考慮しつつ、KEGG から取得した 1279 件のタンパク質間相互作用情報から、リン

(注 1) : 式 3 は式 1、式 2 の内容を含んでいる。

(注 2) : ftp.genome.jp/pub/kegg/xml/KGML.v0.5/haa

酸化学習データにおける正例と負例を決定する。

**正例** あるタンパク質ペアに対しリン酸化属性が付与されていれば、少なくとも特定の状況において、それは存在するだろう。そこで表1の排他的カテゴリである{8,9,10,16}に含まれる249個の相互作用を正例とした。

**負例** 逆にリン酸化属性が付与されていなくても、その属性が存在する可能性はかなり残されている。そこで、表1の分布からみて正例となりえない排他的カテゴリである{1,2,3,11,17}に含まれる227個の相互作用を負例とした。

#### 4.2 特異値分解 (SVD) による特徴量空間の圧縮

GO タームをそのまま MIL に利用することには困難が伴う。まず、GO の階層性のために、上位カテゴリの GO タームは多数の事例に付与されすぎ、一方で下位カテゴリの GO タームの付与は疎であり事例間の共通性を見つけるのが難しい。また、GO タームの付与には依存関係があるために情報が冗長である。また分子機能のみに絞り込んでも GO タームは 622 種類と次元数が大きいため、MIL での計算量が大きくなる。

そこで、分布を考慮しながら属性空間を圧縮できる特異値分解 (SVD)<sup>(18)</sup> を用い、GO タームのペアの圧縮した固有空間に写像する。

##### 4.2.1 GO 階層構造に対応する情報量荷重

特異値分解は、多くのタンパク質に現れる属性を強調する。一方、拡張属性は GO の階層に基づくため上位の属性ほど頻出する。このため、GO タームをそのまま特異値分解しても、単に上位概念が優先的に選択され、良い結果は得られない。

適度な抽象度 (深さ) の情報を利用するために、GO 階層構造に対応する情報量  $W_j$  を用いて属性の重み付けを行う。

$$W_j = -\log(p_j) = -\log\left(\frac{m_j}{m}\right) \quad (4)$$

ここで、gene2go に現れるタンパク質数  $m$  に対する属性  $j$  を持つタンパク質数  $m_j$  の比率を  $p_j$  とする。すると、全タンパク質に付与される最上位概念の荷重  $W_j = 0$  となり無視され、特殊化された下位概念ほど強調される。

##### 4.2.2 特異値分解 (SVD)

特異値分解 (SVD) の準備として、KEGG に現れたタンパク質について、そこで一度以上付与されている、GO タームを取り出して行列  $G$  を生成する。事例  $i$  の属性  $j$  における行列要素  $G_{ij}$  の値に  $W_j$  を設定し、それ以外の行列要素は 0 とする。特異値分解は行列  $G$  を、以下の形に分解する。

$$G = USD^T \quad (5)$$

ここで  $U$ 、 $D$  はそれぞれ  $U^T U = I_n$ 、 $D^T D = I_m$  を満たすユニタリ行列で、 $U$  の列ベクトルを左特異ベクトル、 $D$  の列ベクトルを右特異ベクトルと呼ぶ。また  $S$  は  $n$  行  $m$  列の非負の (広義の) 対角行列で、その対角要素を特異値という。

(注3) : 左右のタンパク質毎の GO 属性を OR 論理で集約したあと、左右のタンパク質を結合した GO タームのベクトルを一事例とした。

## 5. 判別予測の実験結果

### 5.1 判別性能

SVD で得られた固有空間では、上位の次元が入力情報源の特徴を強く反映すると期待できる。そこで、2章で説明し MIL を適用するにあたり、上位から [2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 20] 個の次元を選択利用した。

一方、Diverse Density を用いる手法では、その性能はガウス分布のスケールファクター  $s$  に敏感であるから、この値を [500, 1000, 2500, 5000, 10000, 20000, 40000, 80000, 160000, 320000, 640000, 1280000] と変化させた。

MIL によるリン酸化属性の有無の予測を、Leave one out のクロスバリデーションで評価した結果を図2に示す。探索した範囲では、4次元、 $s=160,000$  にて正解率=86.1%という最高性能を得られた。10次元、 $s=10,000$  においても正解率=85.6%であり、この間をつなぐように性能の高い峰が現れている。おそらく、使用する次元数を増やすと、事例間の距離が増大するために、 $s$  が小さい方が良い性能が現れると考えられる。

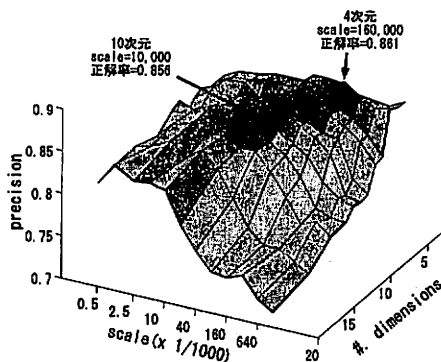


図2 パラメータチューニングによる性能変化  
利用次元数とスケールファクターの調整した。  
4次元、 $S=160,000$  にて最大の正解率=86.1%が得られた。

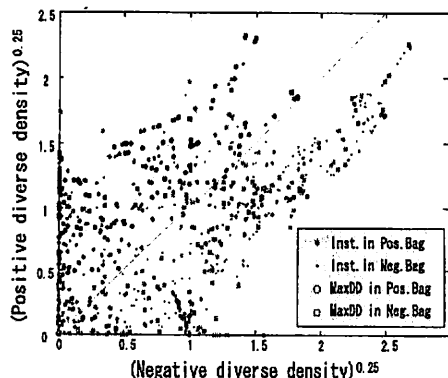


図3 Diverse Density の正例寄与分と負例寄与分  
横軸:負例寄与分, 縦軸:正例寄与分。(4次元,  $S = 160,000$ )

### 5.2 Diverse Density の分布

次に、図3に最高性能を得られた4次元、 $s=160,000$ において、式(3)で $sign_1$ が+となる正例バッグからの寄与分を縦軸に $sign_1$ が-となる負例バッグからの寄与分を横軸に、それぞれ0.25乗して表した。改良 Diverse Density( $d_D(x)$ )値は、両軸の引き算に相当するので、点線(緑)よりも左上が正例と判定される領域で、右下が負例と判定される領域である。

まず正例バッグ内の事例(\* (ピンク))と、負例バッグ内の事例(● (黒))に着目すると、正例バッグ内の事例が、点線より右下にやや多くあらわれる場合がある。参考のため、事例単位での正解率は計算すると79.5%であった。正例バッグ内の事例(\* (ピンク))に比べて、正例バッグを代表する事例(○ (赤))の多くは点線の左上に現れる。MILでは、バッグ内の事例で、 $d_D(x)$ の値が最も大きな値を予測結果に用いるため、バッグを代表する事例の分布は平均的に左上側に移動したためである。これに伴って、負例バッグ(□ (青))も左上から選ばれる場合が増えるため、これが偽陽性の原因となっている。

### 5.3 誤判別の原因解析

今後の性能向上をめざした準備として、誤判別が発生状況を解析した。

正解が正例である239バッグの判定状況を図4に示す。ここで、横軸はバッグ内で陽性判定されたインスタンスの比率で、縦軸はバッグ内でのインスタンス毎の改良 Diverse Density( $d_D(x)$ )の最大値である。そして数字は、バッグに含まれるインスタンスの数である。バッグ内の一つのインスタンスが陽性判定されれば、そのバッグは陽性とみなされるので、陽性比率が正の値をもつ225バッグは真陽性となり、全てのインスタンスが陰性であった14バッグが偽陰性となる。同様に、正解が負例である192バッグの判定状況を図5に示す。ここで、陽性比率が0の146個のバッグが真陰性と判定されている一方で、バッグ内にも一つでも陽性を含む46バッグは偽陽性と判定されている。

以上より判別性能の低下は、偽陽性バッグからの影響が大きいたことがわかった。偽陽性が発生する原因には、インスタンス

表2 着目する偽陽性バッグ(390番)

Diverse density(dd)が正の値である2番目のインスタンスのために偽陽性となっている

#	diverse density	pred ictio	Left protein		Right protein	
			geneID	symbol	geneID	symbol
1	-0.09	0	10379	ISGF3G	6772	STAT1
2	0.11	0	10379	ISGF3G	6773	STAT2
3	-0.02	0	10379	ISGF3G	6774	STAT3
4	-0.44	0	10379	ISGF3G	6775	STAT4
6	-0.44	0	10379	ISGF3G	6776	STAT5
6	-0.44	0	10379	ISGF3G	6777	STAT5B
7	-0.44	0	10379	ISGF3G	6778	STAT6

を正負に判別する能力の他に、MIL固有の原因が存在する。つまり、実際には正例バッグ中の全てのインスタンスが、正例であるとは限らないのであるが、学習時にその点が考慮されていないという原因である。

その影響を調査するために、7個中で2番目のインスタンスのみがが陽性判定された、第390バッグに着目する(表2参照)。ここで、2番目のインスタンスはISGF3G-STAT2間の相互作用において、リン酸化が存在すると推定している。

第390バッグ第2インスタンスの陽性判定をサポートするのは、特徴量空間中でその近傍にある正例バッグ内のインスタンスである。そこで、学習データに含まれるインスタンスを近い順に上位10個を列挙したのが表3である。これより、2番目(第129バッグの第5インスタンス)と4番目(第215バッグの第6インスタンス)が、陽性判定に貢献しているといえる。

次に、表4を見ると、第120バッグにおいて、第5インスタンスは陰性と判定されており、このインスタンスを負例とみなしても、このバッグは正しく陽性と判定できる。同様に、表5から、第215バッグにおいて、第6インスタンスを負例とみなしても、このバッグも、正しく陽性と判定できる。

現状の手法では、正例バッグ中の全インスタンスが、陽性判定への影響力をもつが、本解析で示したような状況が一部に存在することを考えると、その影響力は過剰である。適切に正例バッグ内のインスタンスに対し、正例としての適格性を判定できれば、性能向上ができる可能性を示唆した。

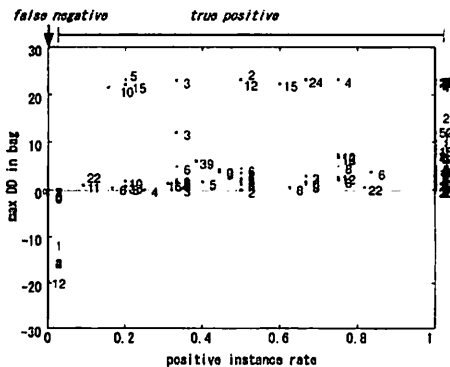


図4 正例バッグ(239個)のバッグ毎の判別状況  
横軸:陽性判定されたインスタンス比率、縦軸: $d_D(x)$ のバッグ内最大値、数字:バッグが含むインスタンス数。

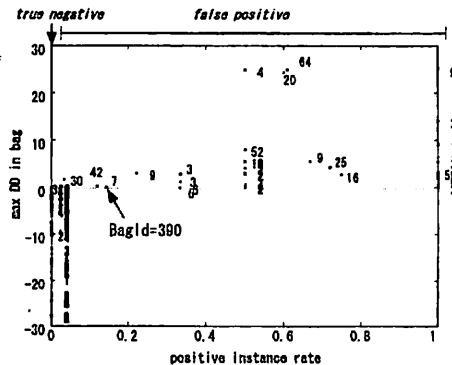


図5 負例バッグ(192個)のバッグ毎の判別状況  
横軸:陽性判定されたインスタンス比率、縦軸: $d_D(x)$ のバッグ内最大値、数字:バッグが含むインスタンス数。

表 3 第 390 番バッグの第 2 インスタンスから近い 10 インスタンス  
正例バッグである第 129 番バッグの第 5 インスタンスと、

第 215 バッグの第 6 インスタンスの影響が大きい

rank	bagID	instanceID	distance	pos/neg
1	390	2	0.0000	-
2	229	2	0.0017	-
3	301	1	0.0017	-
4	229	6	0.0023	-
5	260	1	0.0038	-
6	214	2	0.0040	+
7	246	1	0.0040	-
8	403	36	0.0045	-
9	403	40	0.0045	-
10	403	48	0.0045	-

表 4 偽陽性原因インスタンスをもつ第 129 番バッグ  
偽陽性の原因である第 5 インスタンスは負例と判定されている

#	diverse density	pred ictio	Left protein geneID symbol	Right protein geneID symbol
1	0.00	0	898 CCNE1	1027 CDKN1B
2	-0.20	0	898 CCNE1	1028 CDKN1C
3	0.07	0	1017 CDK2	1027 CDKN1B
4	4.85	0	1017 CDK2	1028 CDKN1C
5	-1.96	0	9134 CCNE2	1027 CDKN1B
6	-16.27	0	9134 CCNE2	1028 CDKN1C

表 5 偽陽性原因インスタンスをもつ第 215 番バッグ  
偽陽性の原因である第 5 インスタンスは負例と判定されている

#	diverse density	pred ictio	Left protein geneID symbol	Right protein geneID symbol
1	4.15	0	6885 MAP3K7	1147 CHUK
2	4.01	0	6885 MAP3K7	3551 IKBKB
3	3.25	0	6885 MAP3K7	8517 IKBKG
4	1.31	0	10454 MAP3K7IP1	1147 CHUK
5	-9.08	0	10454 MAP3K7IP1	3551 IKBKB
6	-1.80	0	10454 MAP3K7IP1	8517 IKBKG
7	-1.94	0	23118 MAP3K7IP2	1147 CHUK
8	-9.14	0	23118 MAP3K7IP2	3551 IKBKB
9	-4.59	0	23118 MAP3K7IP2	8517 IKBKG

## 6. おわりに

局所相互作用仮定をおくことで、複数のサブユニットを含むタンパク質間での相互作用を、サブユニット間ペアの組合せの何れかに還元するアプローチをとった。すると、サブユニット間のペア事例の少なくとも一つが正例なら、相互作用 (バッグ) が正例となり、相互作用属性の予測課題は Multiple-Instance 学習 (MIL) として定式化できた。タンパク質毎の属性として Gene Ontology を利用し、KEGG から相互作用データを取り出して相互作用毎にリン酸化属性の有無を判別予測する実験を行った。前処理の特異値分解 (SVD) で得た上位 4 次元を利用し、MIL による予測結果を、Leave one out のクロスバリデーションで評価したところ 86.1% の正解率を得た。なお、問題の性質を考慮して、MIL の手法として多数決型で判別を行う改良 Diverse Density 法を提案した。また、偽陽性の発生状況を解析することで現状手法に陽性判定への影響力が強すぎることを指摘し、正例バッグ内のインスタンスに対する正例適格性を判定できれば、性能向上ができる可能性を示唆した。

今後は、タンパク質毎の属性としてドメイン情報の利用や、リン酸化以外の相互作用属性の予測を行いたい。また、KEGG 上の個々のタンパク質を含むサブユニットには、同質な性質も

つ (バリエーション) サブユニットと、異質な性質をもつサブユニットが混在しており、これらを区別したの解析や予測を行いたい。また、技術的には、インスタンス毎の正例適格性の考慮や、従風変数に適応した特徴選択を含めた手法に展開したい。

附辞 本研究の推進にあたり、本技術の応用分野などについてご議論頂いた富士通 (株) の土井健太郎氏に感謝する。

## 文 献

- [1] Rhodes DR, Tomlins SA, et. al., "Probabilistic model of the human protein-protein interaction network." 2005.Nat Biotechnol. 2005 Aug;23(8):951-9. Related Articles, Links
- [2] Min Su Lee, Seung Soo Park, Min Kyung Kim, "A Protein Interaction Verification System Based on a Neural Network Algorithm", CSB2005, 2005.
- [3] Peri S, Navarro JD, et. al., "Development of human protein reference database as an initial platform for approaching systems biology in humans.", Genome Res. , 13 (10):2363-71, 2003.
- [4] Kyoto Encyclopedia of Genes and Genomes, <http://www.genome.jp/kegg/>
- [5] Ekins S, Nikolsky Y, Nikolskaya T., "Techniques: application of systems biology to absorption, distribution, metabolism, excretion and toxicity.", Trends Pharmacol Sci. ;26(4):202-9. 2005.
- [6] G. Dietterich, Richard H. Lathrop, and Tomas Lozano-Perez, "Solving the multiple-instance problem with axis-parallel rectangles, by Thomas", Artificial Intelligence, 89(1-2), pp. 31-71, 1997.
- [7] Maron, O., and Lozano-Perez, T., "A framework for multiple-instance learning.", Advances in Neural Information Processing Systems, 10. MIT Press, 1998).
- [8] Jun Yang, Rong Yan, Alexander G. Hauptmann, "Multiple instance learning for labeling faces in broadcasting news video.", ACM Multimedia 2005: 31-40, 2005.
- [9] Z.-H. Zhou, K. Jiang, and M. Li, "Multi-instance learning based web mining," Applied Intelligence, in press.
- [10] Gene Ontology Consortium, <http://www.geneontology.org/>