3A1-5

タンパク質相互作用属性の出現解析とその予測

Analyzing and predicting attribute of protein-protein interaction

山川 宏[*1] 丸橋 弘治[*1] 仲尾 由雄[*1]
Hiroshi Yamakawa Koji Maruhashi Yoshio Nakao

[*1](株) 富士通研究所
FUJITSU LABORATORIES LTD.

We propose a method for predicting types of protein-protein interactions using a multiple-instance learning (MIL) model. Given an interaction type to be predicted, the MIL model was trained using interaction data collected from biological pathways, where positive bags were constructed from interactions between protein complexes of that type, and negative bags from those of other types. 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%.

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，その治療に役立ちそうなタンパク質を関連づける等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，[Rodes 05] 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，[Lee 05] 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 [Ekins 05]．実際，代表的なデータベースである Human Protein Reference Database [HPRD] 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，相互作用属性は付与されていない．逆に，[KEGG] 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 ([KEGG] 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，学習データで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．

連絡先: 山川宏，(株) 富士通研究所 IT コア研究所，
〒 211-8588 川崎市中原区上小田中 4-1-1，tel:044-754-2658, fax:044-754-2693, e-mail:*ymkw@jp.fujitsu.com*

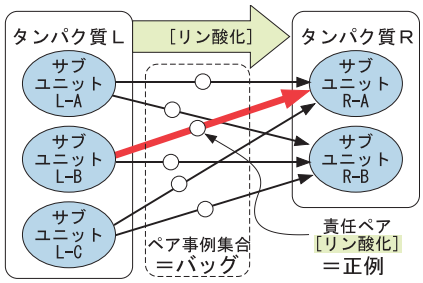

図 1: タンパク質ペアの相互作用とバッグモデル
サブユニットの組合せであるペア事例の少なくとも一つがリン酸化 (正例) であれば，全体としてリン酸化 (正例) であるとする．

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，本仮定の基でのデータ解析には，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，[Dietterich 97] が薬物の分子活性に対する予測課題に用いた．その後に多くの MIL 手法が提案され，画像分類 [Maron 98], 学習やテキスト分類，株式市場予測，動画像の顔ラベリング [Yang 05],Web マイニングへの利用 [Zhou 06] 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，[Maron 98] 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得たとき，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_i^- からの寄与の積として定義される．

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

ただし，所与のバッグ $B_i^{\pm}$ 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．

$$\begin{aligned}\Pr\left(x|B_i^+\right) &= 1 - \prod_j \left(1 - \Pr\left(x|B_{ij}^+\right)\right) \\ \Pr\left(x|B_i^-\right) &= \prod_j \left(1 - \Pr\left(x|B_{ij}^-\right)\right)\end{aligned} \tag{2}$$

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

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	indirect	binding_association	phosphorylation	inhibition	dissociation	dephosphorylation	ubiquination	state	合計
1	.	.	.	.	.	.	.	.	.	.	1	4
2	.	.	.	.	.	.	.	.	.	1	.	7
3	.	.	.	.	.	.	.	.	1	.	.	8
4	.	.	.	.	.	.	.	1	.	.	.	13
5	.	.	.	.	.	.	1	.	.	.	.	145
6	.	.	.	.	.	.	1	.	.	1	.	6
7	.	.	.	.	.	.	1	.	1	.	.	6
8	.	.	.	.	.	1	.	.	.	.	.	150
9	.	.	.	.	.	1	.	1	.	.	.	1
10	.	.	.	.	.	1	1	.	.	.	.	41
11	.	.	.	.	1	.	.	.	.	.	.	181
12	.	.	.	1	.	.	.	.	.	.	.	97
13	.	.	.	1	.	1	.	.	.	.	.	5
14	.	.	1	.	.	.	.	.	.	.	.	525
15	.	.	1	.	.	.	.	.	1	.	.	11
16	.	.	1	.	.	1	.	.	.	.	.	52
17	1	.	.	.	.	.	.	.	.	.	.	27
合計	27	0	588	102	181	249	198	14	25	13	4	1279

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 の近傍に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，このままでは負例かの影響が強すぎた．

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．

$$\tilde{d_D}(x) \propto \sum_i \mathrm{sign}_i \left[1 - \prod_j \left(1 - \Pr\left(x|B_{ij}^{\pm}\right)\right)\right] \tag{3}$$

(ここで sign_i はバッグ i が正例のとき +，負例のとき −．[*1])

$x^{(m)}$ 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[*2]，そこから相互作用属性を含むタンパク質間相互作用 (

*1 式 3 は式 1，式 2 の内容を含んでいる．
*2 ftp.genome.jp/pub/kegg/xml/KGML_v0.5/hsa

⟨*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うるが，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．

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言える [Gene Onotolgy](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. リン酸化属性有無の判別予測実験

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 Divers 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，計算ソフトウエアである 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みつけるのが難しい．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なため，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)[*3]を用い，GO タームのペアのを圧縮した固有空間に写像する．

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．

$$W_j = -\log(p_j) = -\log(\frac{m_j}{m}) \tag{4}$$

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 $W_i = 0$ となり無視され，特殊化された下位概念ほど強調される．

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．

$$G = USD^T \tag{5}$$

ここで U，D はそれぞれ $U^TU = I_n$，$D^TD = I_m$ 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．

4.3 MIL による判別予測の結果

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．

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_i が+となる正例バッグからの寄与分を縦軸に sign_i 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($\tilde{d_D}(x)$) 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

*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．

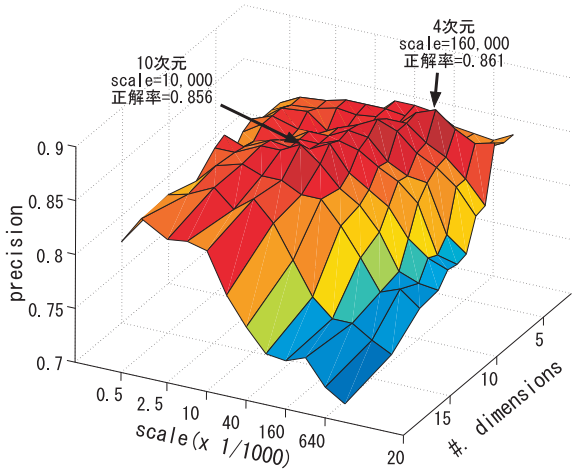

図 2: パラメータチューニングによる性能変化
利用次元数とスケールファクターの調整した.
4 次元,S=160,000 にて最大の正解率=86.1%が得られた.

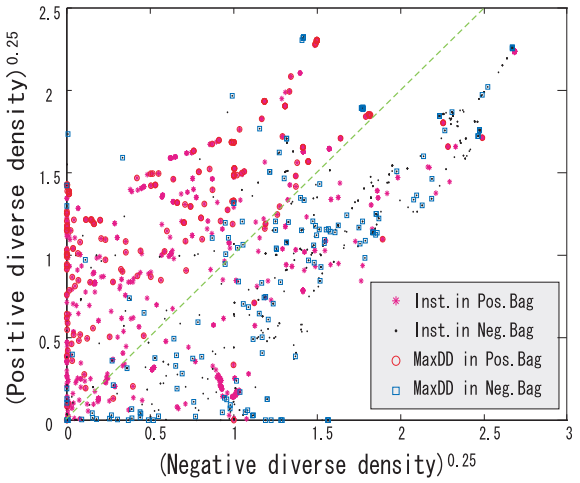

図 3: Diverse Density の正例寄与分と負例寄与分
横軸が不例寄与分，縦軸が正例寄与分 (4 次元,S=160,000).

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，$\tilde{d_D}(x)$ 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，これが擬陽性の原因となっている．

5. 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．

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が含むサブユニットには，同質な性質もつ (バリアント) 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，従属変数に適応した特徴選択を含めた手法に展開したい．

謝辞 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．

参考文献

[Rodes 05] 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

[Lee 05] Min Su Lee, Seung Soo Park, Min Kyung Kim, “ A Protein Interaction Verification System Based on a Neural Network Algorithm ”, CSB2005, 2005.

[HPRD] 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.

[KEGG] Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg/

[Ekins 05] 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.

[Dietterich 97] 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.

[Maron 98] Maron. O., and Lozano-Perez, T., “A framework for multiple-instance learning.”, Advances in Neural Information Processing Systems, 10. MIT Press, 1998).

[Yang 05] Jun Yang, Rong Yan, Alexander G. Hauptmann, “ Multiple instance learning for labeling faces in broadcasting news video.”, ACM Multimedia 2005: 31-40, 2005.

[Zhou 06] Z.-H. Zhou, K. Jiang, and M. Li, “Multi-instance learning based web mining,” Applied Intelligence, in press.

[Gene Onotolgy] Gene Ontology Consortium, http://www.geneontology.org/