社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－MPS－46 (20)
2003／9／19

タンパク質間の相互作用強度予測

林田守広† 上田展久† 阿久津達也†

タンパク質間相互作用の推定のために、様々な方法が提案されている。Association 法や EM 法といった既存の手法では、相互作用するかしないかだけを扱っているが、我々はどのくらいの強度で相互作用するかについても扱った。タンパク質間相互作用の確率モデルとしては、タンパク質の部品であるドメインを使う。同じ種類のドメインは複数のタンパク質に共通に含まれることを利用する。我々はこのモデルに対し線形計画法による定式化を行い、既存の手法と比較して同程度の性能が得られた。さらに、強度予測では極めて良好な結果が得られた。一方、学習データに対する分類精度を最大化する問題は MAX SNP 困難であることを証明した。

Inferring strengths of protein-protein interactions

Morihiro Hayashida† ,Nobuhisa Ueda† and Tatsuya Akutsu†

Several computational methods have been proposed for inference of protein-protein interactions. The existing methods such as the association method and the EM method assume only whether or not each protein pair interacts. However, we also consider strengths of interactions. A probabilistic model is constructed based on domains, where each protein consists of a few of several domains. We propose a new method by formalizing the problem as a linear program. The results show that our method outperforms the existing methods for numerical data. We proved that the problem to maximize the classification accuracy for the training data is MAX SNP-hard.

1. はじめに

タンパク質間相互作用のコンピュータによる推定のために、様々な方法が提案されている。最近、既知のタンパク質データから、タンパク質の「部品」であるドメインの相互作用を推定しようという研究が進展しつつある。ドメイン間相互作用の推定は、タンパク質間相互作用のより詳細な理解のために有用なだけでなく、新たなタンパク質間相互作用の推定にも有用である。本稿では、タンパク質間の相互作用の確率をドメイン間の相互作用の確率で定義し、それらを推定する既存の手法として、Association 法[1] と EM 法[2] を説明した後、提案法である線形計画問題への定式化による推定法[3] を説明する。さらに、SVM を使った方法を説明し、提案法と既存の手法との比較を行う。また、学習データの分類精度を最大化する問題を定義し、MAX SNP 困難であることを証明する。

2. タンパク質間相互作用の確率モデル

Deng らにより提案された相互作用の確率モデルを説明する[2]。$P_1, \ldots, P_N$ をそれぞれタンパク質とし、$D_1, \ldots, D_M$ をドメインとする。また、P_i でタンパク質 P_i に含まれるドメイン集合も表すものとする。P_{ij} を P_i と P_j 間の相互作用を表す確率変数とし、D_{mn} を D_m と D_n 間の相互作用を表す確率変数とする。ここで P_i, P_j 間のドメインペアのうち、一組でも相互作用すれば、P_i, P_j は相互作用するものとし、また、各ドメインペアの相互作用は独立であると仮定する。すると、P_i と P_j が相互作用する確率は、

† 京都大学化学研究所バイオインフォマティクスセンター
Bioinformatics Center, Institute for Chemical Research, Kyoto University

$$\Pr(P_{ij}=1) = 1 - \prod_{(D_m,D_n)\in P_i\times P_j}(1-\Pr(D_{mn}=1))$$

と記述される。

3. ドメイン間相互作用の推定

上で述べた確率モデルは、既知のタンパク質間の相互作用データからドメイン間の相互作用の確率を推定（学習）するために利用できる。

3.1 Association 法 (Sprinzak et al., 2001)

Sprinzak らは、アミノ酸のスコア行列推定などに広く用いられている「頻度分布の比からスコアを計算する」という方法を用いて、ドメイン間相互作用のスコアを推定した[1]。

$$\Pr(D_{mn}=1) = \frac{I_{mn}}{N_{mn}}$$

N_{mn} はドメインペア (D_m, D_n) を含むタンパク質ペア (P_i, P_j) の総数を表し、I_{mn} はそのうち、相互作用するタンパク質ペアの数を表す。

3.2 EM 法 (Deng et al., 2002)

Deng らは、最尤法に基づき、EM（Expectation Maximization）アルゴリズムを用いて、ドメイン間相互作用の確率を推定する手法を開発した[2]。

3.3 LPBN 法

線形計画問題へ変換するために、「相互作用する」ことを閾値 Θ $(0 \le \Theta \le 1)$ を用いて、$\Pr(P_{ij}=1) \ge \Theta$ と定義すると、この不等式は、$\lambda_{mn} = \Pr(D_{mn}=1)$ とおいて、

$$1 - \prod_{(D_m,D_n)\in P_i\times P_j}(1-\lambda_{mn}) \ge \Theta$$

$$\sum_{(D_m,D_n)\in P_i\times P_j} \log(1-\lambda_{mn}) \le \log(1-\Theta)$$

log をそれぞれ変数 $\gamma_{mn} = \log(1-\lambda_{mn})$, $\beta = \log(1-\Theta)$ で置き換えれば、線形不等式

$$\sum_{(D_m,D_n)\in P_i\times P_j} \gamma_{mn} \le \beta$$

を得る。β より定数マージンをとって、以下の線形計画問題を得る。

最小化 $\sum_{(P_i,P_j)} \xi_{ij}$

制約条件

$O_{ij}=1$ となる (P_i, P_j) について、

$$\sum_{(D_m,D_n)\in P_i\times P_j} \gamma_{mn} \le \beta - const + \xi_{ij}$$

$O_{ij}=0$ となる (P_i, P_j) について、

$$\sum_{(D_m,D_n)\in P_i\times P_j} \gamma_{mn} > \beta + const - \xi_{ij}$$

$(\forall\gamma_{mn})\ \gamma_{mn} \le 0$

$(\forall\xi_{ij})\ \xi_{ij} \ge 0$

$\beta < 0$

ただし、$O_{ij}=1$ はタンパク質 (P_i, P_j) 間に相互作用が観測されたことを表し、逆に $O_{ij}=0$ は観測されなかったことを表す。

3.4 LPNM 法

これらの手法では、既知のタンパク質間相互作用データを相互作用するかしないかの二値データとして扱っていたが、実際には、何度も実験を繰り返した後の頻度データ Θ_{ij} が得られる。LPNM 法ではこの頻度 Θ_{ij} と、学習データから予測される相互作用の確率 $\Pr(O_{ij}=1)$ とを近づけるように、線形計画問題を立てる。$\beta_{ij} = \log(1-\Theta_{ij}) \le 0$ とおくと、上と同様にして、

最小化 $\sum_{(P_i,P_j)} \alpha_{ij}$

制約条件

$$\left|\sum_{(D_m,D_n)\in P_i\times P_j} \gamma_{mn} - \beta_{ij}\right| \le \alpha_{ij}$$

$(\forall\gamma_{mn})\ \gamma_{mn} \le 0$

$(\forall\xi_{ij})\ \alpha_{ij} \ge 0$

3.5 SVM 法

特徴ベクトル f_{ij} がタンパク質ペア (P_i, P_j) を表すように、(mn) 要素をドメインペア (D_m, D_n) で次のように定義する。

$$f_{ij}^{(mn)} = \begin{cases} 1 & ((D_m,D_n)\in P_i\times P_j) \\ 0 & ((D_m,D_n)\notin P_i\times P_j) \end{cases}$$

3.6 LPBN 法と EM 法の組み合わせ

3.6.1 LPEM 法

LPBN 法を実行後に、結果を初期値として EM 法を実行。

3.6.2 EMLP 法

EM 法を実行後、結果 $\gamma_{mn}^{(EM)}$ を用いて、$\gamma_{mn}^{(EM)}$ からあまり外れない範囲で探索するように、LPBN 法の制約条件に以下の不等式を加えて実行する。

$$\log(1-\delta) \le \gamma_{mn} - \gamma_{mn}^{(EM)} \le \log(1+\delta)$$

ここで、δ は正の定数とする。

4. 実験結果

4.1 二値データ

DIP データベース[4] のコアデータ (core20020404.lst) を使い、2/3 を学習に、残りをテストに使った。ドメインは Pfam[5], InterPro[6] を利用した。結果は表 1 のように、既存の EM 法に対して少しだけ EMLP 法の性能が上回っている。

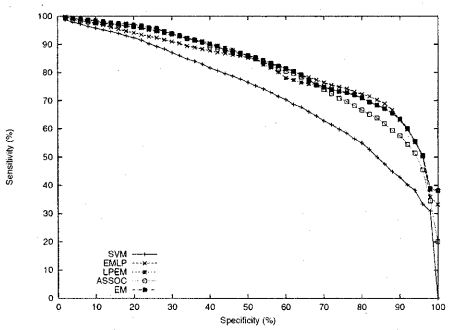

図 1　二値データに対する ROC グラフ

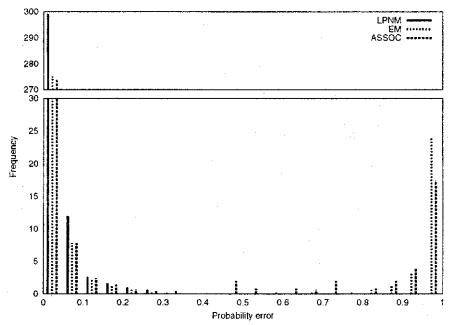

図 2　相互作用確率誤差のヒストグラム

4.2 頻度データ

伊藤ら[7][8]の Yeast Interacting Proteins データベースを使った。頻度データとして、IST (Interaction Sequence Tags) の値を使い、5 回のクロスバリデーションで評価した。表 2 は、頻度を実験回数で割った値と予測した相互作用の確率との差をヒストグラムで表した。LPNM 法は誤差 0 の付近に集中している。EM 法は、尤度極大化手法で、確率を 0 か 1 に近づけるため、誤差 1 の付近にも多く見られるが、LPNM 法では見られなかった。また、その平均誤差は、LPNM 法が 0.03 程度であるのに対し、EM 法では 0.29 程度であった。

5. 相互作用推定の計算量

上述の相互作用のモデルに基づいて次のように問題を定義し、それが MAX SNP 困難となることを証明する。

5.1 MAX PPI 問題

以下の不等式が与えられたとき、できるだけ多くそれらの不等式が成立するようにドメイン間の相互作用の確率を表すパラメータ λ_{mn} を見つける問題と定義する。

$$\begin{cases} P_i \text{ と } P_j \text{ が相互作用するとき、} \\ \prod_{(D_m,D_n)\in P_i\times P_j}(1-\lambda_{mn}) \le 1-\Theta \\ P_i \text{ と } P_j \text{ が相互作用しないとき、} \\ \prod_{(D_m,D_n)\in P_i\times P_j}(1-\lambda_{mn}) > 1-\Theta \end{cases}$$

ただし、全ての不等式が成立可能である場合は線形計画法により多項式時間でパラメータを見つけることができる。

5.2 MAX SNP 困難であることの証明

MAX SNP 完全問題である MAX 2SAT-B[9] からの L-リダクション[9] が存在することを示す。それぞれの定義は以下の通りである。

5.2.1 MAX 2SAT-B

n 個の変数 $x_1,...,x_n$ からなる m 個の節 $C_1,...,C_m$ をできるだけ多く充足するように真偽値割り当てを決める問題と定義する。ただし、同じ変数は高々 B 回までしか現れない。

5.2.2 L-リダクション

多項式時間アルゴリズム f,g と正の定数 α,β が、最適化問題 Π の各インスタンス I について以下の 2 つの条件を満たすとき、最適化問題 Π から Π' への L-リダクションが存在するという[9]。

(a) アルゴリズム f によって生成される Π' のインスタンス $I' = f(I)$ は、$OPT(I), OPT(I')$ をそれぞれ I, I' の最適値をとると、$OPT(I') \le \alpha OPT(I')$ を満たす。

(b) I' の解がコスト c' で与えられたとき、アルゴリズム g は、$|c - OPT(I)| \le \beta|c' - OPT(I')|$ を満たす、コスト c の解を生成する。

5.2.3 証　明

MAX 2SAT-B からの変換アルゴリズム f を次のように設計する。MAX 2SAT-B の各リテラルを MAX PPI のドメインとみなす。同じ変数の正リテラル x、負リテラル $\bar{x}$ も全く別のドメインとみなす。MAX 2SAT-B には現れない変数 α を導入し、ドメイン α のみを持つタンパク質 P_α と各節 C_k のリテラルからなるタンパク質 P_k との相互作用を考える。f はこれら全てが相互作用するとして、次のように不等式を立てる。

$$\begin{cases} k = 1,\ldots,m \\ P_k = \{x_i, x_j\} \text{ と } P_\alpha = \{\alpha\} \text{ について} \\ (1-\lambda_{i\alpha})(1-\lambda_{j\alpha}) \le 1-\Theta \end{cases} \quad (1)$$

ここで、変数 y_i を $1-\lambda_{i\alpha} \le \sqrt{1-\Theta}$ が満たされるなら真、そうでないなら偽とする。上の各不等式は、$y_i \vee y_j$ と変形できる。同じ変数 x_i に対して、正リテラル x_i と負リテラル $\bar{x}_i$ があると、それぞれについて $y_i, y_{\bar{i}}$ も存在する。$y_i, y_{\bar{i}}$ がともに同じ真偽値を持たないように MAX PPI のインスタンスに以下の不等式を追加する。

$$\begin{cases} i = 1,\ldots,n \\ P'_i = \{x_i, \bar{x}_i\} \text{ と } P_\alpha = \{\alpha\} \text{ について} \\ (1-\lambda_{i\alpha})(1-\lambda_{j\alpha}) > 1-\Theta \\ \Rightarrow \bar{y}_i \vee \bar{y}_{\bar{i}} \Leftrightarrow \overline{y_i \wedge y_{\bar{i}}} \end{cases} \quad (2)$$

ただし、各不等式は $2B$ 本づつ存在するとする。不等式 (2) は $2B$ 本づつあるから、(1) で

同時に $y_i, y_{\bar{i}}$ がともに真となるよりは、(2) で同時に真とならない方が有利である。その上で (1) に依存してどちらかが真となる。最適解をとるとき、x_i と y_i、$\bar{x_i}$ と $y_{\bar{i}}$ に対するそれぞれの真偽値割り当ては等しくなるものが存在する。したがって

$$OPT(f(I)) = OPT(I) + 2Bn \qquad (3)$$

f は明らかに多項式時間で $f(I)$ を生成する。全ての MAX SNP 問題はインスタンスの大きさのある定数比で近似できる [9] から $OPT(I) \geq m/\alpha_1$。変数の数は節の数の 2 倍以内 $n \leq 2m$ より、$n \leq 2\alpha_1 OPT(I)$。したがって、(3) より、

$$OPT(f(I)) \leq (1 + 4B\alpha_1)OPT(I) \qquad (4)$$

$\alpha = 1+4B\alpha_1$ とおけば、L-リダクションの (a) が満たされる。

$f(I)$ の解がコスト c' で得られたとし、各 i について $y_i, y_{\bar{i}}$ の c' に貢献するコストを $c'(i)$ とする。$c' \leq \sum_{i=1}^{n} c'(i)$ である。同様にこの解をアルゴリズム g によって I の解に変換したときのコスト c についても $c(i)$ を定義する。$c'(i)$ はさらに不等式 (1)、(2) それぞれについて $c'_1(i)$、$c'_2(i)$ を定義する。$c'(i) = c'_1(i) + c'_2(i)$ である。そこで g を $y_i, y_{\bar{i}}$ にしたがって次のように定義する。

(a) $y_i, y_{\bar{i}}$ がともに真のとき、g は適当に x_i に真か偽を割り当てる。このとき、$c'(i) = 0$。$x_i, \bar{x_i}$ はそれぞれ高々 B 回までしか I に現れないから、$y_i, y_{\bar{i}}$ も (1) 中に高々 B 回までしか現れず、$c'_1(i) \leq c(i) + B$。したがって、$c'(i) \leq c(i) + B$

(b) y_i が真で $y_{\bar{i}}$ が偽のとき、x_i を真に割り当てる。$c'_1(i) = c(i)$、$c'_2(i) = 2B$。したがって、$c'(i) = c(i) + 2B$

(c) y_i が偽で $y_{\bar{i}}$ が真のとき、x_i を偽に割り当てる。(b) と同様にして、$c'(i) = c(i) + 2B$

(d) $y_i, y_{\bar{i}}$ がともに偽のとき、適当に x_i に真か偽を割り当てる。$c'_2(i) = 2B$ で、$c'_1(i) \leq c(i)$ より、$c'(i) \leq c(i) + 2B$。

以上より、全ての i について、$c'(i) \leq c(i) + 2B$。それ故、

$$c' \leq c + 2Bn$$

両辺から $OPT(f(I))$ を引き、

不等式 (3) より

$$\Leftrightarrow \quad c' - OPT(f(I)) \leq c - OPT(I)$$

$c - OPT(I) \leq 0$ より

$$\Leftrightarrow |c - OPT(I)| \leq |c' - OPT(f(I))|$$

g は多項式時間で解を変換し、L-リダクションの条件 (b) を $\beta = 1$ で満たす。よって、MAX PPI 問題は MAX SNP 困難である。

参 考 文 献

1) Sprinzak,E. and Margalit,H., Correlated sequence-signatures as markets of protein-protein interaction, *J. Mol. Biol.*, 311:681-692, 2001.

2) Deng,M., Mehta,S., Sun,F. and Chen,T., Inferring domain-domain interactions from protein-protein interactions, *Genome Research*, 12:1540-1548, 2002.

3) Hayashida,M., Ueda,N. and Akutsu,T., Inferring strengths of protein-protein interactions from experimental data using linear programming, *Bioinformatics* (Suppl. for ECCB, in press), 2003.

4) Xenarios,I., Salwínski,L., Duan,X.J., Higney,P., Kim,S. and Eisenberg,D., DIP: The database of interacting proteins. A research tool for studying cellular networks of protein interactions. *Nucl.Acids.Res.*, 30:303-305, 2002.

5) Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe, K.L., Marshall,M. and Sonnhammer,E.L.L., The Pfam protein families database. *Nucl.Acids.Res.*, 28:45-48, 2002.

6) Zdobnov,E.M. and Apweiler,R., InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17: 847-848, 2001.

7) Ito,T., Tashiro,K., Muta,S., Ozawa,R., Chiba,T., Nishizawa,M., Yamamoto,K., Kuhara,S. and Sakaki,Y., Towards a protein-protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins *Proc.Natl.Acad.Sci* 97:1143-1147, 2000.

8) Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y., A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc.Natl.Acad.Sci* 98:4569-4574, 2001.

9) Papadimitriou,C.H. and Yannakakis M., Optimization, approximation, and complexity classes *J.Comp.Sys.Sci.* 43: 425-440, 1991.