Vol.2009-BIO-18 No.3
2009/9/17

# 網羅的タンパク質間相互作用予測システムにおける判別精度の改良

大上雅史[†1]　松崎裕介[†1]
松崎由理[†1]　秋山泰[†1]

タンパク質間相互作用 (Protein-Protein Interaction：PPI) ネットワークの解明は細胞システムの理解や構造ベース創薬に重要な課題であり，網羅的 PPI 解析法の確立が求められている．我々がこれまで行ってきた PPI 予測では，タンパク質ドッキング (Protein-Protein Docking：PPD) のトップスコアや，クラスタリングのメンバ数を用いていたが，タンパク質の数が大きい系では予測精度が悪化するため，手法の改善が望まれていた．本稿では，PPD の結果から機械学習によって相互作用判別モデルを生成し，PPI 予測に用いることで精度が向上したことを示す．

# Improvement of the classification performance in all-to-all protein-protein interaction prediction system

Masahito Ohue,[†1] Yusuke Matsuzaki,[†1]
Yuri Matsuzaki[†1] and Yutaka Akiyama[†1]

The elucidation of the protein-protein interaction (PPI) network is an important problem in the understanding of the cellular system and structure-based drug design. Our previous PPI prediction system was based on maximum scores of the protein-protein docking (PPD) pairs and the number of cluster members with the clustering. However, improvement of the technique was required to improve prediction precision for the system with large number of proteins. We developed a classification model by machine learning technique from a result of the PPD and improved precision of a PPI prediction.

†1 東京工業大学 大学院情報理工学研究科 計算工学専攻
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

## 1. はじめに

タンパク質間相互作用 (Protein-Protein Interaction: PPI) は生命現象において中心的な役割を果たしていることが近年解明されつつあり，この相互作用ネットワークの解明は，細胞内のシグナル伝達経路の特定や，それをターゲットとした創薬に対する重要課題となっている．相互作用予測は，タンパク質のアミノ酸配列モチーフを用いた手法[1)–3)] や，既知のドメイン間相互作用情報などのデータベースに基づく手法[4),5)] などがよく用いられていたが，近年タンパク質の 3 次元構造が次々と決定されており[⋆1]，またタンパク質立体構造が複合体形成に深く関わっていることが分かってきたため，今後は豊富に得られつつある 3 次元情報を利用した手法が主流になると考えられている．3 次元情報を利用した相互作用予測手法は既にいくつか存在するが，自由エネルギー変化の推定など厳密に物理化学的な量を見積もろうとするアプローチでは，タンパク質 1 対 1 の相互作用予測にも数日から数週間かかるものがほとんどであった．また，相互作用ネットワーク解明のためには複数のタンパク質群の相互作用予測シミュレーションを網羅的に行う必要があり，予測計算回数は数万回から数百万回にのぼるため，現実の時間内でネットワーク予測を行うことは不可能であると見られてきた．しかし比較的計算時間の小さい形状相補性に基づくタンパク質ドッキング (Protein-Protein Docking: PPD) を利用することで，網羅的 PPI 予測を現実的な時間で行うことが可能となった．ここで問題となるのは，PPD システムが本来複合体を形成しないタンパク質ペアも，また相互作用をするタンパク質ペアも同等にドッキング計算を行い，システムが尤もらしいとする複合体構造を出力する点である．すなわち，出力されたドッキング構造から相互作用をするかしないかの判断を行う必要があるということであり，その判定を行う PPI 予測システムの確立が求められる（**図 1**）．

## 2. タンパク質ドッキング

本稿が指すタンパク質ドッキング (Protein-Protein Docking: PPD) とは，タンパク質を剛体とみなし，複合体形成の際にタンパク質構造が変化しないという仮定の元，タンパク質を 3 次元ボクセル空間上で表わして表面形状の相補性を主として計算するシミュレーション手法である．これまでに MolFit[6)] や FTDock[7)]，ZDOCK[8)–10)] などの PPD ソフトウェ

⋆1 構造決定されたタンパク質は Protein Data Bank (PDB) に登録される．2009/8/18 時点での PDB エントリ数は 59,618 である．

Vol.2009-BIO-18 No.3
2009/9/17

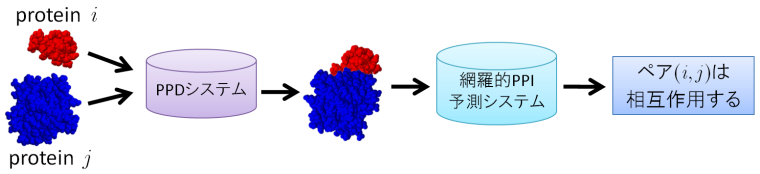


**図 1** タンパク質ドッキングによる網羅的 PPI 予測
Fig. 1 All-to-all protein-protein interaction prediction by protein-protein docking

アが開発され，現在まで用いられてきたが，いずれも網羅的 PPI 予測を目的としたものではないこともあり，我々は独自に “MEGADOCK”[11] という PPD システムを開発してきた．以下ではその現行バージョンである MEGADOCK Ver.2.1[12] について示す．

### 2.1 MEGADOCK Ver.2.1

MEGADOCK Ver.2.1 のドッキングは形状相補性と静電的相互作用の計算からなる．MEGADOCK で用いている PPD の良さを表すスコア **S** は以下の式で与えられる．

$$\mathbf{R}_{l,m,n} = G^R_{l,m,n} + iE^R_{l,m,n}$$
$$\mathbf{L}_{l,m,n} = G^L_{l,m,n} + iwE^L_{l,m,n}$$
$$\mathbf{S}_{\alpha,\beta,\gamma} = \Re\left[\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N}\mathbf{R}_{l,m,n}\mathbf{L}_{l+\alpha,m+\beta,n+\gamma}\right]$$
$$= \sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N}\left(G^R_{l,m,n}G^L_{l+\alpha,m+\beta,n+\gamma} - wE^R_{l,m,n}E^L_{l+\alpha,m+\beta,n+\gamma}\right)$$

**R** はレセプタータンパク質を，**L** はリガンドタンパク質を $N \times N \times N$ の 3 次元ボクセルに分割したときの離散関数で，$(l,m,n)$ はそのボクセルの座標である．それぞれの離散関数は形状相補性による項 $(G)$ と静電的相互作用による項 $(E)$ で表わされ，スコア **S** は $(\alpha,\beta,\gamma)$ をリガンドの平行移動ベクトルとした相関関数（の実部）として計算される．$w$ は静電的相互作用の重みを表すパラメータである．

形状相補性による項は，以下で表わされる real Pairwise Shape Complementarity(rPSC) スコア[12] を用いている．

$$G^R_{l,m,n} = \begin{cases} \#\text{ of R atoms within } (D + \text{R atom radius}) & \text{open space} \\ 3\rho & \text{solvent excluding surface of the R} \\ 9\rho & \text{core of the R} \end{cases}$$

$$G^L_{l,m,n} = \begin{cases} 0 & \text{solvent accessible surface layer of the L} \\ 1 & \text{solvent excluding surface layer of the L} \\ \delta & \text{core of the L} \\ 0 & \text{open space} \end{cases}$$

ただし，$\rho, \delta$ はマッチや衝突のペナルティの大きさを決めるパラメータであり，$D$ は距離のカットオフパラメータである．Ver.2.1 では $\rho = -3, \delta = 2, D = 3.6$Å を用いている．

また，静電的相互作用による項は以下のように計算される．ボクセル $i(l,m,n)$ に対する電界 $\phi_i$ を，

$$\phi_i = \sum_j \frac{q_j}{\varepsilon(r_{ij})r_{ij}}, \qquad \varepsilon(r) = \begin{cases} 4 & (r \le 6\text{Å}) \\ 38r - 224 & (6\text{Å} < r < 8\text{Å}) \\ 80 & (r \ge 8\text{Å}) \end{cases}$$

と定義する．ただし $q_j$ はボクセル $j$ の電荷，$r_{ij}$ は $i$ と $j$ の Euclid 距離[⋆1]，$\varepsilon(r)$ は誘電率をモデル化した関数である．アミノ酸残基ごとに CHARMM19[13] に基づいて原子に電荷を与え，ボクセルに分割してボクセル電荷 $q_{l,m,n}$ を決定し，$E^R_{l,m,n}, E^L_{l,m,n}$ を，

$$E^R_{l,m,n} = \begin{cases} \phi_{i(l,m,n)} & \text{entire voxel excluding core} \\ 0 & \text{core of the R} \end{cases}$$

$$E^b_{l,m,n} = q_{l,m,n}$$

とする．

MEGADOCK ではこのようにして計算されるスコア **S** を，レセプターを固定しリガンドを回転・移動させながら計算していく．1 つのリガンド回転角に対して一番スコアの高い値をとる $\alpha, \beta, \gamma$ を探索することを，15° 刻みで 3600 通り行い，そのうちの上位 2000 個の予測を出力する．

### 2.2 all-to-all PPD による網羅的 PPI 予測

all-to-all PPD による網羅的 PPI 予測は，対象とする $M$ 個のタンパク質群に網羅的にドッキング計算を行い，得られたドッキングスコアやリガンドの結合部位からそのタンパ

⋆1 2Å 以下の場合は 2Å と定める．

ク質ペアが相互作用しているかを判断するという問題である．従来手法としては，PPD のトップスコアの偏差を利用した手法 ($\eta$-method) や，Grouping を用いた手法 (AEP) がある．以下でそれぞれを紹介する．

#### 2.2.1 $\eta$-method

$\eta$-method[12] とは，$M \times M$ 個の組み合わせのタンパク質ペア $(i, j)$ に関するドッキング予測の 1 位のスコア $S_{ij}^{top}$ から，全 $M^2$ 個の値を母集団とする $z$-score を求め，ある閾値 $z_\tau$ (1.0 など) を超える値を持つペア $(i, j)$ は相互作用すると判定する方法である．ドッキングスコアはタンパク質の大きさにある程度依存するため，ここではタンパク質の表面積によってスコアの補正を行っている．補正されたスコア $\eta_{ij}$ は，

$$\eta_{ij} = \frac{(S_{ij}^{top})^{2/3}}{\min\{A_i, A_j\}} \tag{1}$$

と表され，$z$-score は，

$$z_{ij} = \frac{\eta_{ij} - \mu}{\sigma}$$

となる．ただし，$A_i$ はタンパク質 $i$ の表面積であり，$\mu$ と $\sigma$ はそれぞれ全 $M^2$ 個の $\eta$ の値を母集団とする平均，標準偏差である．

#### 2.2.2 Affinity Evaluation and Prediction

Affinity Evaluation and Prediction(AEP)[14),15)] とは，Tsukamoto らが提案した PPD に基づく相互作用判定法である．AEP の処理は，以下の 4 つのステップから構成されている．

(1) 形状相補性による PPD
(2) Grouping とよばれるクラスタリングに基づく独自の処理
(3) $z$-score の計算
(4) 親和性スコアとよばれるクラスタの重み付け和を用いた指標の計算

## 3. 機械学習による相互作用判別関数の生成

### 3.1 学 習 手 法

機械学習の手法は，C4.5[16] やサポートベクターマシン[17] など現在までに様々な手法が考案されているが，本研究では計算量が小さく，過学習が起こりにくいという理由からブースティング[18] を採用する．ブースティングは機械学習メタアルゴリズムであり，弱学習器の重み付き多数決によって判別関数を生成する手法である．ブースティングで最もポピュラーな手法に AdaBoost[19] があるが，サンプルのノイズや外れ値に比較的弱いことが知られている[20),21)] ため，本研究では以下に示す LogitBoost[22] を用いることにする．

### 3.2 LogitBoost

LogitBoost[22] は 1998 年に Friedman らによって提案されたアルゴリズムであり，1997 年に Freud らが提案した AdaBoost[19] を改良したものとなっている．弱学習器 $h(\in H) : D \to \{+1, -1\}$ の学習方法は与えられているものとすると，サンプル $S = ((\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N))$, $\mathbf{x} \in D$, $y \in \{+1, -1\}$ が与えられたときのブースティングアルゴリズムは以下のようになる．

**ブースティングアルゴリズム**

初期化 $(t := 1)$
　初期分布 $P_1$, 弱学習クラス $H = \{h_1, h_2, ..., h_{|H|}\}$, 閾値 $\varepsilon(> 0)$
　各弱学習器の重要度を表すベクトル $\mathbf{a}_1 = (\alpha_1, \alpha_2, ..., \alpha_{|H|}) := \mathbf{0}$
**repeat**{
　**Step1**: 弱学習クラス $H$ から良い弱学習器 $h_{opt}$ を選択する．

$$h_{opt} := \underset{h \in H}{\operatorname{argmin}} \sum_{(\mathbf{x}, y)} P_t(\mathbf{x}, y)[h(\mathbf{x}) \neq y]$$

　**Step2**: $h_{opt}$ の重要度 $\alpha_{opt}$ を算出する．

$$\alpha_{opt} := \underset{\alpha \geq 0}{\operatorname{argmin}} R_U(f + \alpha h_{opt}, P_t)$$

　求めた重要度 $\alpha_{opt}$ で $\mathbf{a}$ を更新する．

$$\mathbf{a}_{t+1} := \mathbf{a}_t + (0, 0, ..., 0, \alpha_{opt}, 0, ..., 0)$$

　統合決定関数 $f_{\mathbf{a}_t}$ をつぎのように定義する．

$$f_{\mathbf{a}_t}(\mathbf{x}) := \sum_{h \in H} \alpha_h h(\mathbf{x})$$

　**Step3**: 分布を更新する．

$$P_{t+1} := \frac{P_t(\mathbf{x}, y) U'(-f(\mathbf{x}) y)}{\sum_{(\mathbf{x}, y)} P_t(\mathbf{x}, y) U'(-f(\mathbf{x}) y)}$$

　**if** $R_U(f_{\mathbf{a}_t}, P_0) < \varepsilon$　**then break**
　$t := t + 1$
}
出力: 統合決定関数 $f_{\mathbf{a}_t}$

Vol.2009-BIO-18 No.3
2009/9/17

**表 1** 23 個のタンパク質複合体サブセット
Table 1 List of 23 protein complex subset

| |
|---|
| 1ACB 1AK4 1AVX 1AY7 1B6C 1CGI 1D6R 1E96 1EAW 1EWY 1GCQ 1GHQ 1GRN 1HE1 1KAC 1KTZ 1PPE 1SBB 1UDI 2PCC 2SIC 2SNI 7CEI |

ここで $R_U$ は $f$ の期待損失であり，$R_U(f,P)=\sum_{(\mathbf{x},y)}U(-yf(\mathbf{x}))P(\mathbf{x},y)$ と表される．$U$ は損失関数と呼ばれ，判別解の信頼性評価に用いられる．この $U$ の取り方によってアルゴリズムは多様化する．例えば AdaBoost では $U_{\mathrm{Ada}}(z)=e^z$，本研究で用いる LogitBoost では $U_{\mathrm{Logit}}(z)=\ln(1+e^{2z})$ としている．

### 3.3 弱学習クラス

LogitBoost の弱学習クラスとして，本研究ではブースティングで広く用いられている Decision stump を採用する．Decision stump は高さ 2 の決定木であり，エントロピーに基づく分割を行うものである．

### 3.4 対象データ

本研究で用いるデータは，all-to-all PPD による網羅的 PPI 予測の研究で広く用いられている PPD Benchmark 2.0[23] である．このデータセットには 84 のタンパク質複合体におけるレセプタータンパク質とリガンドタンパク質の立体構造が収録されているが，従来手法[12),15)] との比較のため，23 の複合体によるサブセットと，84 全データを用いたセットの 2 種類を利用する．さらにテスト事例として，84 のセットを 2 つに分けたサブセットによる交差検定を行う．23 の複合体のサブセットの内容を**表 1** に，84 全データの内容と交差検定用データセットを**表 2** にそれぞれ示す．

通常の学習問題の場合，サンプルが $M^2$ 個あれば $M^2/m$ 個ずつに分けた $m$ 個のサブセットによって交差検定を行うことができるが，相互作用予測の場合，2 つのサブセットにわたる組み合わせ，例えば表 2 における subset A の 1A2K のレセプターと，subset B の 1I4D のリガンドによる予測構造サンプルなどは用いることができない．従って $m$ 個のサブセットを作る際は，サンプルは $M^2/m^2$ 個ずつとなる．

PPD システムは前述の通り，ドッキングスコアの高いものから 2000 個の複合体予測を出力する．そのため，1 位から 2000 位までのドッキングスコア $S_1$〜$S_{2000}$ による 2000 次元の特徴量が得られることになる．しかし例えば表 1 の all-to-all PPD から得られるサンプル数は 529 であり，サンプル数より大きい 2000 次元という高次元をそのまま用いると，学習結果に悪影響を及ぼす可能性が高い．そこで特徴選択を行うことになるが，ラッパー法[24)] では可能な組み合わせが膨大になり計算時間がかかりすぎてしまう．一方フィルタ法ならば

**表 2** PPD Benchmark 2.0 の全複合体データセットと交差検定用サブセット
Table 2 List of PPD Benchmark 2.0 protein complex dataset and subset for validation

| | |
|---|---|
| subset A | 1A2K 1ACB 1AHW 1AK4 1AKJ 1ATN 1AVX 1AY7 1B6C 1BGX 1BJ1 1BUH 1BVK 1BVN 1CGI 1D6R 1DE4 1DFJ 1DQJ 1E6E 1E6J 1E96 1EAW 1EER 1EWY 1EZU 1F34 1F51 1FAK 1FC2 1FQ1 1FQJ 1FSK 1GCQ 1GHQ 1GP2 1GRN 1H1V 1HE1 1HE8 1HIA 1I2M |
| subset B | 1I4D 1I9R 1IB1 1IBR 1IJK 1IQD 1JPS 1K4C 1K5D 1KAC 1KKL 1KLU 1KTZ 1KXP 1KXQ 1M10 1MAH 1ML0 1MLC 1N2C 1NCA 1NSN 1PPE 1QA9 1QFW 1RLB 1SBB 1TMQ 1UDI 1VFB 1WEJ 1WQ1 2BTF 2HMI 2JEL 2MTA 2PCC 2QFW 2SIC 2SNI 2VIS 7CEI |

**表 3** $S_1$〜$S_{100}$ の相関係数 $r_{ij}$（一部抜粋）
Table 3 Correlation coefficient $r_{ij}$ of $S_1$–$S_{100}$ (extracted)

| | $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ | $S_7$ | $S_8$ | $S_9$ | $S_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 1.000 | | | | | | | | | |
| $S_2$ | 0.759 | 1.000 | | | | | | | | |
| $S_3$ | 0.775 | 0.972 | 1.000 | | | | | | | |
| $S_4$ | 0.772 | 0.963 | 0.989 | 1.000 | | | | | | |
| $S_5$ | 0.778 | 0.955 | 0.983 | 0.994 | 1.000 | | | | | |
| $S_6$ | 0.782 | 0.954 | 0.981 | 0.991 | 0.997 | 1.000 | | | | |
| $S_7$ | 0.782 | 0.951 | 0.978 | 0.988 | 0.994 | 0.998 | 1.000 | | | |
| $S_8$ | 0.785 | 0.948 | 0.975 | 0.986 | 0.992 | 0.995 | 0.998 | 1.000 | | |
| $S_9$ | 0.784 | 0.946 | 0.974 | 0.984 | 0.990 | 0.994 | 0.996 | 0.999 | 1.000 | |
| $S_{10}$ | 0.785 | 0.946 | 0.973 | 0.983 | 0.989 | 0.993 | 0.995 | 0.998 | 0.999 | 1.000 |

短時間で計算を終えることができるので，予備実験としてフィルタ法の代表的手法である Relief アルゴリズム[25)] を用いて，$23\times23$ サンプルに対する特徴のリランキングを行った．すると，$S_1$〜$S_{100}$ の順位にほとんど変動がなかったため，$S_1$〜$S_{100}$ の相関係数 $r_{ij}$ を計算して調べたところ，**表 3** のように $S_1$ 以外のスコア同士の相関に比べ，$S_1$ との相関係数の値が小さくなっていることがわかった．これより，特徴として多くのスコアを用いることにあまり意味はないと考えられるため，本研究では $S_1$ のみを特徴として用いることにする．

また，MEGADOCK のドッキングスコアの分布は**図 2** のように相互作用するものとしないものの間に大きな差異はなく，ドッキングスコア自体はタンパク質の大きさにある程度依存する．すなわちドッキングスコアのみでの判定は非常に難しい問題となる．そこで，レセプタータンパク質の表面積 $A_{\mathrm{R}}$ と，リガンドタンパク質の表面積 $A_{\mathrm{L}}$ も特徴として用いることにする．しかし，事前実験としてブースティングを何度か実行してみたところ，弱

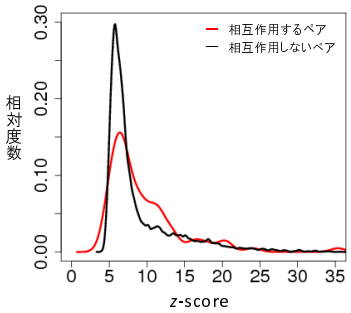


**図 2** トップスコア $S_1$ に対する $z$-score の相対度数（密度推定したもの）
Fig. 2 A relative frequency of $z$-score for maximum score $S_1$ (density estimated)

学習器のほとんどが $S_1$ を用いた規則を生成していた．これは，表面積がレセプターとリガンドでそれぞれ共通なのに対し，ドッキングスコアは組み合わせの数だけ存在しうるので，その多様性により現れた問題であると考えられる．そこで $\eta$-method のように，$S_1$ に対し式 (1) によって補正スコア $\eta$ を求め，これも特徴として利用することにする．

all-to-all PPD の正例と負例は，PDB に登録されているか否かで判断する．ベンチマークのデータセットでいえば，$84 \times 84$ の可能な組み合わせに対し，正例は 84 個，負例は 6972 個となる．一般に（ベンチマークによる）$M \times M$ の all-to-all PPD では正例と負例の数の比が $1 : M-1$ となり，そのままでは多くの学習器が「全て負例である」という予測を導きかねない．これは，数学的には正しいが実用上は無意味であるため，正例と負例の重みを同等に扱う必要がある．

### 3.5 オーバーサンプリング

正例と負例を同等に扱うための手法としてオーバーサンプリングとアンダーサンプリングがあるが，正例の数は全サンプルに対して極めて少ないので負例を削減して正例に合わせようとすると，サンプルの数が少なくなり予測精度が悪化する恐れがある．そこで本研究ではオーバーサンプリングを採用する．オーバーサンプリングの方法としては，ランダムオーバーサンプリング (ROS) や SMOTE[26] などいくつかあるが，ROS は正例同士で重みが異なってしまい，SMOTE では新たに生成された正例が何のタンパク質複合体なのか説明できないので，単純に正例の数を全て $M-1$ 倍することで負例と数を合わせる．

**表 4** 性能評価に用いる値一覧
Table 4 List of values to use for performance evaluation

| | |
|---|---|
| TP(True Positive) | 相互作用すると予測されたペアによる複合体が PDB に登録されている |
| FP(False Positive) | 相互作用すると予測されたペアによる複合体が PDB に登録されていない |
| FN(False Negative) | 相互作用しないと予測されたペアによる複合体が PDB に登録されている |
| TN(True Negative) | 相互作用しないと予測されたペアによる複合体が PDB に登録されていない |
| precision(適合率) | $\text{precision} = \frac{\text{TP}}{\text{TP+FP}}$ |
| recall(再現率) | $\text{recall} = \frac{\text{TP}}{\text{TP+FN}}$ |
| $F$-score | $F\text{-score} = \frac{2 \cdot \text{recall} \cdot \text{precision}}{\text{recall+precision}}$ |

### 3.6 $F$-score vs. $F_\beta$-score

学習の結果の評価に用いる指標として，**表 4** にあるような値が主に用いられ，従来の研究[12),14),15)] でも，網羅的 PPI 予測の性能は $F$-score（$F$-measure，$F$ 値ともよばれる）によって評価されてきた．しかし，いずれも precision は極端に小さく，recall は大きめの値を示しており，最終的な $F$-score は通常の学習問題に比べると非常に小さな値となっていた．学習問題であることに変わりはないので $F$-score で比較を行うこと自体に問題はないが，網羅的 PPI 予測の場合，正例と負例が $1 : M-1$ の比であること，確実に相互作用を検出しつつさらに新たな相互作用の可能性も示唆していく必要があることから，precision 対して recall にある程度の重み付けを行って判断する方が良いと考えられる．そこで本研究では $F$-score に重み付けを行った $F_\beta$-score[27)] を用いる．$F_\beta$-score は以下の式で表される．

$$F_\beta\text{-score} = \frac{(1+\beta^2) \cdot \text{recall} \cdot \text{precision}}{\text{recall} + \beta^2 \cdot \text{precision}}$$

$\beta$ の値としては，対象の系の大きさに合わせて，$\beta = \sqrt{M-1}$ とおく．このとき，

$$F_{\sqrt{M-1}}\text{-score} = \frac{M \cdot \text{recall} \cdot \text{precision}}{\text{recall} + (M-1) \cdot \text{precision}}$$

となる．

## 4. 結　　果

### 4.1 23×23 full training

$\eta$-method[12)] と，特徴次元 $(S_1, A_\text{R}, A_\text{L})$ の LogitBoost(Logit 3)，特徴次元 $(S_1, \eta, A_\text{R}, A_\text{L})$ の LogitBoost(Logit 4) を比較した．LogitBoost のパラメータは，弱学習器数を 30，shrink-

**表 5** 23×23 の相互作用予測結果
Table 5 Result of 23×23 interaction prediction

| | $\eta$-method | Logit 3 | Logit 4 |
|---|---|---|---|
| TP | 6 | 21 | 21 |
| FP | 51 | 184 | 123 |
| FN | 17 | 2 | 2 |
| TN | 455 | 322 | 383 |
| precision | 0.105 | 0.102 | 0.146 |
| recall | 0.261 | 0.913 | 0.913 |
| $F$-score | 0.150 | 0.184 | 0.251 |
| $F_\beta$-score | 0.245 | 0.679 | 0.743 |

age factor [⋆1] $\nu = 0.5$ とした．結果を**表 5** に示す．

表 5 より，LogitBoost は $\eta$-method よりも良い結果を示していることがわかる．特に，$F_\beta$-score において Logit 4 は 0.749 という高い値を示した．Logit 3 と Logit 4 では TP の数は同じだが，FP の数は Logit 4 の方が小さい．このことから，特徴 $\eta$ が FP の削減に効果があることがわかる．LogitBoost は $\eta$-method よりも FP の数は多いが，Logit 4 では FP の増加は 2.4 倍なのに対し，TP の増加は 3.5 倍であるため，Logit 4 の結果は良好であるといえる．

### 4.2 84×84 full training

$\eta$-method，AEP[15] と，Logit 3，Logit 4 を比較した．LogitBoost のパラメータは，弱学習器数を 100，shrinkage factor $\nu = 0.5$ とした．結果を**表 6** に示す．

表 6 より，$\eta$-method に比べて LogitBoost は良い結果を示している．AEP と比べると $F$-score では若干劣るものの，recall と $F_\beta$-score では倍以上の向上となった．この場合も Logit 3 と Logit 4 の TP は同じとなったが，FP の数から Logit 4 の方が性能が良いと判断できる．

### 4.3 two subset cross validation

Logit 3 と Logit 4 を表 2 におけるサブセットに適用し，交差検定を行った．LogitBoost のパラメータは，弱学習器数を 70，shrinkage factor $\nu = 0.5$ とした．結果を**表 7** に示す．

表 7 より，交差検定でも Logit 3 に比べて Logit 4 が良くなっていることが分かる．しかし，その精度は Logit 3，4 ともに十分であるとは言えず，Logit 4 の $F_\beta$-score の平均で

⋆1 LogitBoost におけるパラメータ．小さい値であるほど過学習しにくいと考えられている．
⋆2 TP,FP,FN,TN は文献[15] に値が記載されていないため省略．

**表 6** 84×84 の相互作用予測結果
Table 6 Result of 84×84 interaction prediction

| | $\eta$-method | AEP[⋆2] | Logit 3 | Logit 4 |
|---|---|---|---|---|
| TP | 9 | - | 66 | 66 |
| FP | 1008 | - | 3129 | 2456 |
| FN | 75 | - | 18 | 18 |
| TN | 5964 | - | 3843 | 4516 |
| precision | 0.009 | 0.035 | 0.021 | 0.026 |
| recall | 0.107 | 0.274 | 0.786 | 0.786 |
| $F$-score | 0.016 | 0.062 | 0.040 | 0.051 |
| $F_\beta$-score | 0.095 | 0.253 | 0.545 | 0.584 |

**表 7** 交差検定による相互作用予測結果
Table 7 Result of interaction prediction by cross validation

| | training=A, test=B | | training=B,test=A | |
|---|---|---|---|---|
| | Logit 3 | Logit 4 | Logit 3 | Logit 4 |
| TP | 11 | 17 | 11 | 14 |
| FP | 561 | 580 | 484 | 484 |
| FN | 31 | 25 | 31 | 28 |
| TN | 1161 | 1142 | 1238 | 1238 |
| precision | 0.019 | 0.028 | 0.022 | 0.028 |
| recall | 0.262 | 0.405 | 0.262 | 0.333 |
| $F$-score | 0.036 | 0.053 | 0.041 | 0.052 |
| $F_\beta$-score | 0.201 | 0.308 | 0.208 | 0.265 |

0.287 程度となった．精度が良くなかった原因としては，正例サンプル数が少ないことが主な原因であると考えられるが，図 2 に示したように正例と負例の両者の分布を分ける基準を見つけるのは困難であり，特徴が 4 つ程度（ユニークであるのは 3 つ）では判別を行うことが難しいことは先に述べたとおりである．LogitBoost で一定の精度を出せることは分かったので，本研究の結果を踏まえ，新たな特徴の模索を行うことを今後の課題とする．

## 5. おわりに

本研究では，all-to-all PPD による網羅的 PPI 予測の精度を向上させるために，相互作用判定をドッキングスコアを特徴とする学習問題とし，LogitBoost によるブースティング学習を行った．また，網羅的 PPI 予測特有の，サンプル数が正例と負例で大きく異なることに対して，その対策と結果の評価方法の提案を行った．実際に PPD Benchmark 2.0 に適用したところ，従来手法に比べて LogitBoost の精度は向上したが，実問題を想定した交

差検定での精度は良いとはいえないという結果となった．

今後の課題はいくつか考えられるが，特に重要なこととして予測構造に対する特徴量を増やすことが挙げられる．前節で述べたように交差検定で精度が上がらなかった原因として考えられるものの 1 つであるため，適切に特徴を選出し，現在のものに追加することで精度向上が期待できる．

また，サンプル数の問題に対しては，未知のサブセットのタンパク質に対して，既知サブセットとのドッキング予測の結果も利用することで，（負例のみではあるが）増やすことはできる．しかし，問題となっているのは正例の少なさであるため，根本的な解決には至らない．この問題に対しては，本研究では単純にサンプル数を倍化させたが，正例となるペアに対して分子動力学法などによって構造のアンサンブルをとり，アンサンブルドッキングを行うことでサンプル数を増やすという方法が考えられる．しかし，構造のアンサンブルをとるための計算時間を考慮した上で行う必要があり，また必ずしも精度が向上するかどうかは不明である．なお，2008 年にアップデートされた PPD Benchmark 3.0[28] を利用することも検討中である．

その他として，機械学習におけるパラメータ推定や，学習手法の再検討が挙げられる．

**謝辞** 本研究は，文部科学省 最先端・高性能汎用スーパーコンピュータの開発利用「次世代生命体統合シミュレーションソフトウェアの研究開発」，および科学研究費補助金（基盤研究（B）19300102）の支援を受けて行われたものである．

## 参 考 文 献

1) H.X. Zhou, Y. Shan: “Prediction of protein interaction sites from sequence profile and residue neighbor list”, *Proteins*, 44(3): 336-343, 2001.
2) Y. Ofran, B. Rost: “Predicted protein-protein interaction sites from local sequence information”, *Federation of European Biochemical Societies Letters*, 544(1-3): 236-239, 2003.
3) A. Koike, T. Takagi: “Prediction of protein-protein interaction sites using support vector machines”, *Protein Engineering, Design & Selection*, 17(2): 165-173, 2004.
4) M. Deng, S. Mehta, F. Sun: “ Inferring domain-domain interactions from protein-protein interactions”, *Genome Research*, 12: 1540-1548, 2002.
5) R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N.J. Krogan, S. Chung, A. Emili, M. Snyder, J.F. Greenblatt, M. Gerstein, “A bayesian networks approach for predicting protein-protein interactions from genomic data”, *Science*, 302(5644): 449-453, 2004.
6) E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A.A. Friesem, C. Aflalo, I.A. Vakser: “Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques”, *Proceedings of the National Academy of Sciences of the United States of America*, 89(6): 2195-2199, 1992.
7) H.A. Gabb, R.M. Jackson, M.J.E. Sternberg: “Modelling protein docking using shape complimentarity, electrostatics and biochemical information”, *Journal of Molecular Biology*, 272(1): 106-120, 1997.
8) R. Chen, Z. Weng: “Docking unbound proteins using shape complementarity, desolvation, and electrostatics”, *Proteins*, 47(3): 281-294, 2002.
9) R. Chen, L. Li, Z. Weng: “ZDOCK: An initial-stage protein-docking algorithm”, *Proteins*, 52(1): 80-87, 2003.
10) R. Chen, Z. Weng: “A Novel Shape Complementarirty Scoring Function for Protein-Protein Docking”, *Proteins*, 51(3): 397-408, 2003.
11) Y. Akiyama, T. Sato, Y. Matsuzaki, Y. Matsuzaki: “MEGADOCK - A rapid screening system for all-to-all protein docking analysis with pre-calculated Fourier library of protein structures”, *Proceedings of the 2008 Annual Conference of the Japanese Society for Bioinformatics*: P032，2008.
12) M. Ohue, Y. Matsuzaki, Y. Matsuzaki, T. Sato, Y. Akiyama: “Improvement of all-to-all protein-protein interaction prediction system by introducing physicochemical interaction”, *IPSJ-SIG Technical Report*, 2009-BIO-17(11): 1-8，2009.
13) B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M. Karplus: “CHARMM: A program for macromolecular energy, minimization, and dynamics calculations”, *Journal of Computational Chemistry*, 4(2): 187-217, 1983.
14) K. Tsukamoto, T. Yoshikawa, Y. Hourai, K. Fukui, Y. Akiyama: “Development of an affinity evaluation and prediction system by using the shape complementarity characteristic between proteins”, *Journal of Bioinformatics and Computational Biology*, 6(6): 1133-1156, 2008.
15) T. Yoshikawa, K. Tsukamoto, Y. Hourai, K. Fukui: “Improving the accuracy of an affinity prediction method by using statistics on shape complementarity between proteins”, *Journal of Chemical Information and Modeling*, 49(3): 693-703, 2009.
16) J.R. Quinlan, “C4.5: Programs for machine learning”, *Morgan Kaufmann*, 1993.
17) V.N. Vapnik: “The nature of statistical learning theory”, *Springer*, 1995.
18) R.E. Schapire: “The strength of weak learnability”, *Machine Learning*, 5(2): 197-227, 1990.
19) Y. Freud, R.E. Schapire: “A decision-theoretic generalization of on-line learning and an application to boosting”, *Journal of Computer and System Science*, 55(1): 119-139, 1997.
20) A. Grove, and D. Schuurmans, “Boosting in the limit: Maximizing the margin

Vol.2009-BIO-18 No.3
2009/9/17

of learned ensembles", *Proceedings of the 15th National Conference on Artificial Intelligence*: 692-699, 1998.
21) G. Rätsch, T. Onoda, and K.R. Müller: "Soft margins for AdaBoost", *Machine Learning*, 42(3): 287-320, 2001.
22) J.H. Friedman, T. Hastie, R. Tibshirani: "Additive logistic regression: A statistical view of boosting", *Technical Report Stanford University*, 1998.
23) J. Mintseris, K. Wiehe, B. Pierce, R. Anderson, R. Chen, J. Janin, Z. Weng: "Protein-protein docking benchmark 2.0: an update", *Proteins*, 60(2): 214-216, 2005.
24) R. Kohavi, G.H. John: "Wrappers for feature subset selection", *Artificial Intelligence*, 97: 273-324, 1997.
25) K. Kira, L. Rendell: "A practical approach to feature selection", *Proc International Conference on Machine Learning*: 249-256, 1992.
26) N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer: "SMOTE: Synthetic minority over-sampling technique", *Journal of Artificial Intelligence Research*, 16: 321-357, 2002.
27) C.J. Van Rijsbergen: "Information retrieval (2nd ed)", *Butterworth*, 1979.
28) H. Hwang, B. Pierce, J. Mintseris, J. Janin, Z. Weng: "Protein-protein docking benchmark version 3.0", *Proteins*, 73(3): 705-709, 2008.