RG-001

配列情報に基づくタンパク質間相互作用予測の構造情報付加による高精度化

Improvement of Sequence-based Protein-Protein Interaction Prediction by Introducing Tertiary Structural Information

中嶋 悠介 †‡　大上 雅史 §　越野 亮 †
Yusuke Nakashima†‡　Masahito Ohue§　Makoto Koshino†

1　序論

本研究は，生命科学研究において近年注目されているタンパク質間相互作用（Protein-Protein Interaction, PPI）と呼ばれる生命現象を計算機で予測するため，タンパク質の配列情報と立体構造情報を併用して予測精度を向上させることを目的としたものである．

タンパク質には単体で機能するものも存在するが，多くの場合，複数のタンパク質や核酸，糖鎖などの生体高分子と相互作用することでその機能を発揮する．特にタンパク質同士の相互作用は PPI と呼ばれ，細胞内におけるシグナル伝達や代謝経路，転写制御機構など，多くの生体内現象に関与している．この PPI の網羅的な解明は，タンパク質機能の理解や疾病メカニズムの解明，創薬ターゲットの決定などにおいて重要であり [1]，生命科学における重要な課題の 1 つとなっている．

PPI は酵母ツーハイブリッド法 [2] や蛍光共鳴エネルギー移動法 [3] などの生化学的実験によって決定される．現在判明している PPI の約 1/3 は酵母ツーハイブリッド法によるものであり，例えばヒトの PPI を収集したデータベースである Human Protein Reference Database (HPRD) [4] Release 9 では，全 39,420 エントリ中約 37%にあたる 14,444 エントリが酵母ツーハイブリッド法によって決定された PPI となっている．しかし，新たなタンパク質配列 [5] や構造 [6] が次々と決定され続けている近年の現状を鑑みると，決定対象となるタンパク質間相互作用の数は組み合わせ的に増大し，生化学的実験による PPI の網羅的な決定は困難となりつつある．このため，近年では計算機を用いて PPI を予測する [7] という試みが数多くなされてきた [8, 9, 10, 11, 12]．

PPI 予測とは，対象の 2 つのタンパク質に関する情報をもとに，それらの間の相互作用の有無を予測する 2 値分類問題である．この問題を解く手法としては，配列情報を用いた探索や学習を基にして予測する手法 [8]，共進化情報に基づく手法 [9]，立体構造情報に基づく手法 [10, 11, 12] が挙げられる．立体構造情報に基づく手法は，既知のタンパク質複合体の構造情報をテンプレートとして構造マッチングを基に予測を行う手法 [10] と，既知の複合体構造情報を用いることなく予測する *de novo* ドッキングの手法 [11, 12] の 2 つのアプローチに大別される．配列相同なタンパク質に囚われない予測が可能であることから，これらの立体構造情報に基づく手法は近年特に注目されているが，両手法ともに多くの偽陽性予測を出力してしまうという欠点を持つ．また，テンプレートベースの手法はテンプレートとなる類似複合体構造が未知である場合には予測が行えない，または予測の精度が低くなるという問題が存在する．この問題に対し，テンプレートベースの手法と *de novo* ドッキングの手法の両予測結果に対してコンセンサスをとり，新たな予測とする方法 [13] が提案されているが，テンプレート構造が必要であることには変わりはなく，多くの種類のタンパク質に対して広く予測を行うという目的には，適用範囲が限定されるテンプレートベースの手法は適さないものと考える．

そこで本研究では，配列情報からの予測手法と *de novo* ドッキングの手法を組み合わせた PPI 予測手法を提案する．配列情報を用いることで相同なタンパク質間相互作用の特徴を，*de novo* ドッキングによって相互作用の構造的な特徴を加味することができ，またテンプレート構造を必要としないことから，予測の適用範囲が限定されることなく精度を向上させることができると考えられる．また，近年では特定の構造を取らない天然変性領域が PPI に重要な役割を持つことが議論されており [14]，構造情報に基づく手法のみでは天然変性領域の情報を取り込むことができないが，配列情報を利用することで天然変性領域に関する情報も加味することができると考えられる．

2　関連研究

2.1　配列情報に基づく相互作用予測

以下では配列情報に基づく PPI 予測手法として，先に挙げた Shen らの手法 [8] について述べる．Shen らは HPRD の PPI 情報を用い，負例を HPRD 内のタンパク質のランダムな組み合わせにより作成して，訓練データセットを構築した．また，このデータセットから，Conjoint Triad Feature (CTF) [8] と呼ばれる特徴ベクトルを生成して，S カーネルを用いたサポートベクターマシンによって判別器を生成し，PPI の有無を判別した．

2.1.1　Conjoint Triad Feature (CTF)

CTF は以下の手順で求められるタンパク質配列についての特徴ベクトルである．

1. アミノ酸残基を，アミノ酸の持つ極性と体積を基準とした 7 グループに分類する（表 1）．
2. 連続する長さ 3 の部分文字列の出現頻度 $f_{ijk}\ (i,j,k=1,2,\ldots,7)$ を求める．
3. 出現頻度 f_{ijk} を次式によって正規化し，$7^3=343$ 次元の特徴ベクトル $\boldsymbol{D}\in[0,1]^{343}$ とする．

$$D_{ijk}=\frac{f_{ijk}-\min_{ijk} f_{ijk}}{\max_{ijk} f_{ijk}} \tag{1}$$

これによって得られるタンパク質 A についての特徴ベクトルを $\boldsymbol{D}_A$ と表す．また，$\boldsymbol{D}_A$ と $\boldsymbol{D}_B$ をつなげた 686 次元の特徴ベクトル $\boldsymbol{D}_{AB}$ をタンパク質 A と B のペアの特徴ベクトルとする．

†石川工業高等専門学校 電子情報工学科
‡東京工業大学 工学部 情報工学科
§東京工業大学 大学院情報理工学研究科 計算工学専攻

表 1: アミノ酸残基の 7 グループ分類

グループ	アミノ酸
1	Ala, Gly, Val
2	Ile, Leu, Phe, Pro
3	Tyr, Met, Thr, Ser
4	His, Asn, Gln, Tpr
5	Arg, Lys
6	Asp, Glu
7	Cys

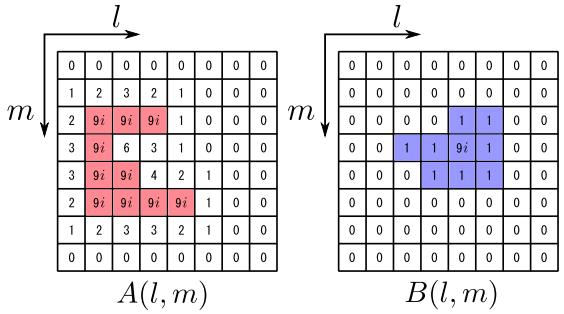

図 1: グリッドに対するスコアの配置例．ここでは簡単のため 2 次元で示している．図中の色の付いたセルはその部分にタンパク質が存在していることを表している．

2.1.2 サポートベクターマシン

サポートベクターマシン [15] はカーネル関数を用いた分類超平面の最適化によって学習を行う手法であり，現在最も利用されている機械学習手法の 1 つである．カーネル関数としては通常は多項式カーネルやガウシアンカーネルなどが用いられるが，タンパク質ペアの特徴ベクトルを扱う場合，ペアの順序に関する対称性が保たれない¶という問題がある．この問題に対し，Shen らはガウシアンカーネルに似た形をとる S カーネル [8] を用いることで解決を図った．S カーネル $k_S(\cdot,\cdot)$ を式 (2) に示す．

$$k_S(\boldsymbol{D}_{AB}, \boldsymbol{D}_{EF}) = \exp(-\gamma s) \tag{2}$$

$$s = \min \left\{ \begin{array}{l} \|\boldsymbol{D}_{AB} - \boldsymbol{D}_{EF}\|^2 \\ \|\boldsymbol{D}_{AB} - \boldsymbol{D}_{FE}\|^2 \end{array} \right\} \tag{3}$$

式 (2), (3) から S カーネルはペアの順序についての対称性を満たすことがわかる．

サポートベクターマシンの学習に関するパラメータとして，式 (2) に現れたカーネル関数のパラメータ γ と，超平面の最適化に関するソフトマージンパラメータ C がある．本研究ではグリッド探索によるパラメータの探索を行った．

2.2 立体構造情報に基づく相互作用予測

以下では立体構造情報に基づく PPI 予測手法として，Matsuzaki らによる *de novo* ドッキングを利用した手法 [11] について述べる．

2.2.1 *de novo* ドッキング計算

de novo ドッキング計算とは，既知の複合体構造とのマッチングを行うことなく，タンパク質単体の構造情報をもとに複合体構造を予測する計算のことを指し，ZDOCK [16]，PatchDock [17]，MEGADOCK [18]，PIPER [19] など，多数のソフトウェアが開発されている．このうち，ZDOCK の 2007 年に開発されたバージョン（ZDOCK 3.0）は，相互情報量を元に決定した独自の原子間統計ポテンシャルを用いることによって高精度な複合体構造予測を実現しており，Matsuzaki らの手法 [11] をはじめとして様々な研究に利用されている [20, 21]．

ZDOCK を含む多くのドッキングソフトウェアは，タンパク質構造をグリッド空間上で表現して計算を行っている．具体的には，2 つのグリッド空間にそれぞれのタンパク質についてのスコア $A(l,m,n)$, $B(l,m,n)$ を付与する．図 1 に模式図を示す．$B(l,m,n)$ を (α,β,γ) だけ平行移動させ，重なったボクセルについて積をとり総和をとった値をその平行移動位置 (α,β,γ) の評価値とする．即ち，評価値 $S(\alpha,\beta,\gamma)$ は，

$$S(\alpha,\beta,\gamma) = \sum_{l,m,n} A(l,m,n)B(l+\alpha, m+\beta, n+\gamma) \tag{4}$$

という畳み込み和の形で書くことができる．この評価値はドッキングスコアと呼ばれる．なお，式 (4) は，

$$S(\alpha,\beta,\gamma) = \mathcal{F}^{-1}\left[\mathcal{F}[A(l,m,n)]^* \mathcal{F}[B(l,m,n)]\right] \tag{5}$$

とすることで離散フーリエ空間上で計算することができ，高速フーリエ変換を用いることで計算量の削減が可能となっている．ここで，$\mathcal{F}[f(l,m,n)]$ は離散関数 $f(l,m,n)$ の離散フーリエ変換を表し，$\mathcal{F}^{-1}[F(o,p,q)]$ は離散フーリエ空間上の関数 $F(o,p,q)$ の逆離散フーリエ変換を表す．また，x^* は x の複素共役を表す．

2.2.2 相互作用判定

Matsuzaki らは ZDOCK 3.0 を用い，15° 刻みでタンパク質を回転させながら平行移動探索を行ったときのドッキングスコア S の上位 2,000 構造を用いて，次式で表される相互作用評価値 E を求め，タンパク質 A, B の間の相互作用の有無の判定を行った．

$$E = \frac{S_{top} - \mu}{\sigma} \tag{6}$$

ここで，S_{top} は出力されたドッキングスコアの中で最大のもの，μ, σ はそれぞれ 2,000 個のドッキングスコアの平均値と標準偏差である．相互作用の有無の判定には閾値 E^* を定め，

$$\mathrm{PPI}(A,B) = \begin{cases} \mathrm{True} & E \geq E^* \\ \mathrm{False} & \mathrm{otherwise} \end{cases} \tag{7}$$

¶カーネル関数 $k(\cdot,\cdot)$ に対し，$k(\boldsymbol{D}_{AB}, \boldsymbol{D}_{EF}) = k(\boldsymbol{D}_{AB}, \boldsymbol{D}_{FE})$ であるとき，対称性が保たれている，という．

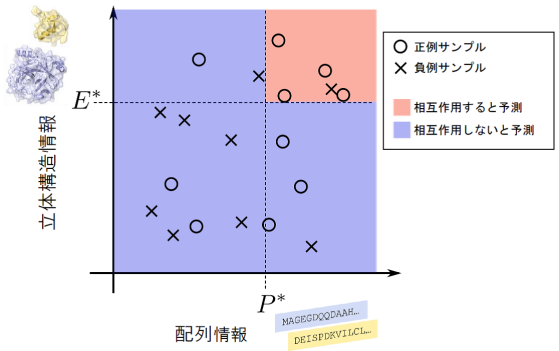

図 2: 提案手法の模式図．○と×はタンパク質ペアの真のラベル（未知）を表し，E^* と P^* で区切られた領域に対して相互作用すると予測する．

として評価値 E によって決定する．E^* によって感度を調節することができるが，予測の際は PPI が既知なデータセットを用い，F 値が最大となるように決定された E^* を用いることが多い [11]．

3 提案手法

従来の配列情報を用いた予測手法 [8] と，*de novo* ドッキング計算によって立体構造情報から予測を行う手法 [11] を組み合わせた提案手法について述べる．提案手法の概要図を図 2 に示す．本手法はタンパク質ペアに対して，配列情報からサポートベクターマシンによって求められる相互作用確率 P と，立体構造情報から求められる相互作用評価値 E を計算し，閾値 E^* と P^* に対して「$(P \geq P^*) \wedge (E \geq E^*)$」が満たされるとき，相互作用すると判定するものである．

配列情報についての評価値（相互作用確率）P は，Shen らの手法に基いて計算される．タンパク質ペアの配列から CTF 特徴ベクトルを生成し，サポートベクターマシンによって学習器を生成する．サポートベクターマシンは通常分類超平面によって 2 値の分類を行うものであるが，本研究ではサポートベクターマシンの分類超平面とサンプルの特徴ベクトルとの距離から求められる近似的な確率値 [22] を用いて，相互作用する確率 P を定義した．この確率値は，サンプル $\boldsymbol{D}_{AB}$ に対して以下のように計算される．

$$P = \Pr(\text{True}|\boldsymbol{D}_{AB}) = \frac{1}{1+\exp(\boldsymbol{w}^\top \boldsymbol{D}_{AB} - h)} \quad (8)$$

ただし，$\boldsymbol{w}$ と h は分類超平面を決定する変数である．また，立体構造情報についての評価値 E は式 (6) に示した通りである．各評価値に対する閾値 P^* と E^* を定め，評価値がいずれも閾値以上になる場合に相互作用すると予測した．ここで，$P^* = 0$ のときは配列情報を用いずに立体構造情報のみをもとに予測することと同値となり，$E^* = -\infty$ のときは配列情報のみをもとに予測することと同値となる．

閾値 P^* と E^* については，関連研究と同様に，既知のデータセットに対し F 値が最大となるように決定した．

表 2: 評価に用いたタンパク質複合体の PDB ID

1ACB,	1AK4,	1ATN,	1AVX,	1AY7,	1B6C,	1BUH,	1BVN,	1CGI,	1D6R,
1DFJ,	1E6E,	1E96,	1EAW,	1EWY,	1F34,	1FC2,	1FQ1,	1FQJ,	1GCQ,
1GHQ,	1GRN,	1H1V,	1HE1,	1HE8,	1I2M,	1IBR,	1KAC,	1KTZ,	1KXP,
1KXQ,	1M10,	1MAH,	1PPE,	1QA9,	1SBB,	1TMQ,	1UDI,	1WQ1,	2BTF,
2PCC,	2SIC,	2SNI,	7CEI						

4 評価実験

4.1 データセット

相互作用予測の学習器を生成するための訓練データセットとして，生化学的実験によって決定された PPI が登録されているデータベースである HPRD [4] から相互作用情報を収集した．使用したバージョンは，HPRD Release 9 である．このバージョンには相互作用情報が 39,240 件登録されている．Shen らが構築したデータセットと同様にして，これらを正例（相互作用する）として用い，負例（相互作用しない）ペアについてはデータベースに登録されていないタンパク質ペアからランダムに正例と同数作成‖した．ここから訓練データセットと検証用データセットに分割した．各データセットの内訳は，訓練データセットが正例数 = 負例数 =32,185，検証用データセットが正例数 = 負例数 =563 **である．

また，提案手法を評価するための構造データセットとして，タンパク質複合体の Protein Data Bank (PDB) 構造が収集された ZLAB Benchmark 2.0 [23] の複合体のうち，単量体同士のヘテロ複合体である 44 個から構成される $44 \times 44 = 1,936$ ペアを用いた．これらは Matsuzaki らが用いたものと同一のものである．44 個の複合体の PDB ID を表 2 に示す．

4.2 評価方法

本研究では，予測結果に対して真陽性 (TP)，偽陽性 (FP)，偽陰性 (FN)，真陰性 (TN) を決定し，適合率，再現率，F 値を用いて評価を行った．各評価尺度についての説明を表 3 に示す．

4.3 従来手法による予測結果

提案手法の評価に先立ち，配列情報に基づく予測と立体構造情報に基づく予測の精度を確認した．以下ではそれぞれの予測結果を示す．

4.3.1 配列情報に基づく予測

新たに S カーネル関数を追加した LIBSVM 3.12 [24] を用い，訓練データセットによるサポートベクターマシンの学習を行った．パラメータ C, γ を変化させたときの検証用データセットにおける F 値の変化を図 3 にそれぞれ示す．

‖タンパク質 A,B,C,D に対して A-B と C-D の情報がデータベースに登録されている場合，これらを正例とし，A-C, A-D, B-C, B-D を負例として 2 ペア選出した [8]．

**両データセットを足して HPRD の全件数と一致しないのは，例えばタンパク質ペア D-E が訓練データセットに含まれて，タンパク質ペア D-F が検証用データセットに含まれるといったような例を除外したためである．

表 3: 混合行列と適合率，再現率，F 値の定義

	DB に存在する	DB に存在しない
PPI 有と予測	TP	FP
PPI 無と予測	FN	TN

$$\text{適合率} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{再現率} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F 値} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}$$

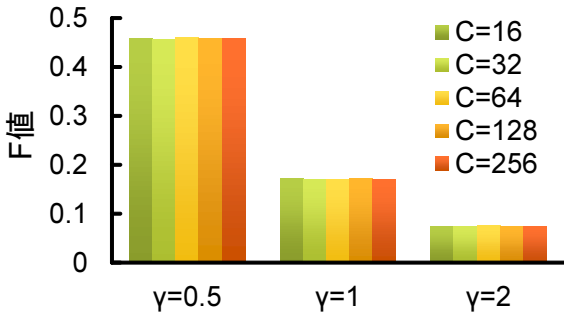

図 3: サポートベクターマシンによる予測を検証用データセットに適用したときの F 値．C と γ はサポートベクターマシンのパラメータである．

C に対してはほとんど F 値は変化せず，γ が 0.5 のときに最も良い F 値を示すことを確認した．なお，図 3 における最良値は，C が 64, 128, 256，γ が 0.5 のときの F 値 0.460 である．以降の解析において，サポートベクターマシンのパラメータは $C = 128$，$\gamma = 0.5$ を用いた．

4.3.2 立体構造情報に基づく予測

構造データセットのタンパク質ペア 1,936 個に対して ZDOCK による相互作用予測を行い，F 値を計算した．結果を図 4 に示す．F 値は E^* が 8.9 のときに最良値 0.361 となった．

4.4 提案手法の評価

提案手法におけるパラメータ P^*, E^* を変化させたときの F 値の変化を図 5 に示す．P^* が 0.15，E^* が 8.9 のとき F 値が最良の 0.366 となった．従来法と提案手法の構造データセットに対する予測結果をまとめたものを表 4 に示す．なお，表 4 における各パラメータは，配列情報に基づく予測が $C = 128, \gamma = 0.5$，構造情報に基づく予測が $E^* = 8.9$ であり，提案手法はこれらに加えて $P^* = 0.15$ を用いている．表 4 より，提案手法は従来法に比べ再現率を維持しつつ適合率を向上させることが確認された．

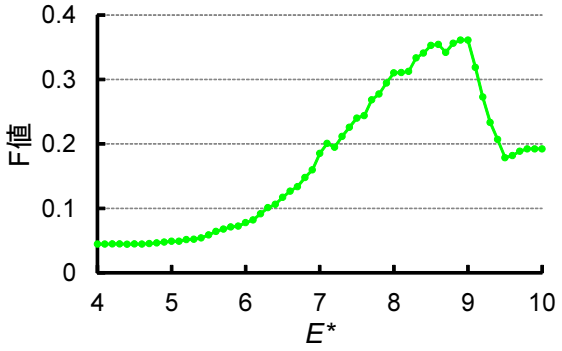

図 4: ZDOCK による予測を構造データセットに適用したときの F 値．

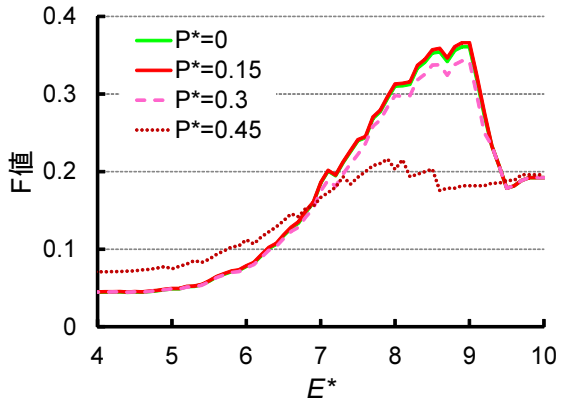

図 5: 提案手法による予測を構造データセットに適用したときの F 値．$P^* = 0$ は図 4 と同一の曲線である．

5 考察

5.1 ROC 解析

表 4 から分かるように，提案手法は従来法に比べて総合的な性能である F 値がほぼ変わらないものの，適合率が向上するという結果を得た．パラメータに対する提案手法の総合的な性能を検証するために，構造データセットに対する Receiver Operating Characteristic (ROC) 曲線 [25] をプロットして予測性能を確認した．ROC 曲線は予測手法の総合的な性能を表す指標の 1 つであり，真陽性率と偽陽性率から描かれる曲線である．対角線がランダムな予測の性能を表し，曲線が対角線より左上の領域にあると性能が良いことを表す．最も

表 4: 各手法の構造データセットに対する予測結果．

	適合率	再現率	F 値
従来法 (配列)	0.050	0.295	0.086
従来法 (構造)	0.464	0.295	0.361
提案手法	0.481	0.295	0.366

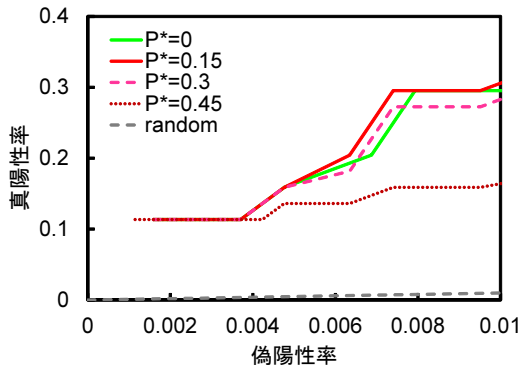

図 6: $ROC_{0.01}$ 曲線．灰色の破線はランダムな予測における性能を表す．

表 5: $ROC_{0.01}$ 曲線の AUC ($AUC_{0.01}$) 値．

	$AUC_{0.01}$
$P^* = 0$ (構造情報のみ)	1.74×10^{-3}
$P^* = 0.15$	1.80×10^{-3}
$P^* = 0.3$	1.70×10^{-3}
$P^* = 0.45$	1.27×10^{-3}
random	0.05×10^{-3}

性能の良い理想的な予測は，$(0,0),(0,1),(1,1)$ の 3 点を直線で結んだ形になる．真陽性率は $TP/(TP+FN)$，偽陽性率は $FP/(FP+TN)$ で計算される．

なお，本研究が対象とする PPI ネットワークは，一般に疎なネットワークとなる性質を持ち [26]，相互作用しないペアに比べて相互作用するペアの数が圧倒的に少ないアンバランスなデータとなることが多く，偽陽性率が少しでも高くなると大量の偽陽性を出力することになる．例として，偽陽性率が 20%で真陽性率が 60%であるとすると，表 2 の 1,936 ペアの場合，正例が 44 ペアに対して負例が 1,892 ペアなので，26 個の真陽性に対して 387 個の偽陽性が生成されることになる．このため，偽陽性率が高い領域での予測結果は実用的でなく，実際には偽陽性率が低い領域で予測を行うことになり，この領域での性能が重要となる．そのため，ROC 曲線の偽陽性率の低い領域でのプロットである $ROC_{0.01}$ 曲線を用いて性能評価を行った．$ROC_{0.01}$ 曲線を図 6 に示す．

また，これらの曲線下面積である AUC (area under the ROC curve) の値を表 5 に示す．表 5 より，いずれの P^* でもランダムな予測より良い結果を示しており，提案手法の $P^* = 0.15$ が，従来法よりも良い値を示すことが分かる．偽陽性率 0.01 までの AUC が優れている提案手法は，疎な性質を持つ PPI ネットワークの予測という問題に適した手法であると言える．

5.2 配列情報による予測について

表 4 から分かるように，配列情報による予測結果は適合率が低い結果となっている．この原因の 1 つとして前述した正例負例のバランスの問題が挙げられる．今回は機械学習として Shen らの報告と同一の手法をとったが，精度を向上させるためには，適用する PPI ネットワークの性質（疎性）を考慮した学習データセットの構築が肝要であると考える．

また，PPI の負例の扱いについては近年議論の的となっており，これまで相互作用するという正例の情報しか蓄積されてこなかった問題が指摘され，相互作用しないという負例の情報の蓄積について検討が進んでいる [27]．2013 年 4 月現在で文献ベースによる負例の情報は 1,291 件に留まっているが，今後更に蓄積されていくものと考えられる．このような負例に関する情報を取り入れることで予測性能を向上させることができると考えられ，本研究の今後の課題としたい．

6 結論

本研究では PPI 予測手法として，配列情報からサポートベクターマシンによって計算された相互作用確率値と，立体構造情報から *de novo* ドッキング計算によって求められた相互作用評価値をもとに予測を行う手法を提案した．提案手法は，配列情報と構造情報のそれぞれの評価値に対する二分決定木を構築したことに相当する単純な方法でありながら，立体構造情報のみによる PPI 予測と比較して精度が向上することを確認し，特に疎な性質を示す PPI ネットワークの予測のような正例に比べて負例が多くなりやすいという問題に適していることが，ROC 解析によって示された．

今後の課題として，前述した学習データセットの構築方法に関する問題の解決が挙げられる．また，今回は 2 つの予測手法の評価値に対する二分決定木を構築して組み合わせるということを行ったが，例えば機械学習によって立体構造情報に基づく特徴量を用いて識別モデルを生成したり，配列相同性の情報から相互作用することが知られている構造ドメインに限定してドッキング計算を行って予測するといった組み合わせ方に関する検討を進めていく予定である．

また，本研究の予測手法は，時空間的な相互作用実現性が考慮されておらず，同一時間・空間上で共存しないタンパク質ペアの偽陽性を生成してしまう可能性がある．遺伝子共発現情報や細胞内局在情報などを利用してそのような偽陽性の排除を行う手法を開発することも，今後の課題としたい．

謝辞

本研究は，科研費（特別研究員奨励費 23·8750）の支援を受けて行われたものである．

参考文献

[1] Wass MN, David A, Sternberg MJE. Challenges for the prediction of macromolecular interactions. *Curr. Opin. Struct. Biol.*, 21(3), 382–390, 2011.

[2] Fields S, Song O. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230), 245–246, 1989.

[3] Förster T. Zwischenmolekulare Energiewanderung und Fluoreszenz. *Ann. Physik*, 437(1–2), 55–75, 1948.

[4] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, *et al.* Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, 37(suppl 1), D767–772, 2009.

[5] UniProt Consortium. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.*, 41(D1), D43–47, 2013.

[6] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res.*, 28(1), 235–242, 2000.

[7] Nussinov R, Schreiber R. *Computational Protein-Protein Interactions*. CRC Press, 2009.

[8] Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H. Predicting protein-protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. USA.*, 104(11), 4337–4341, 2007.

[9] Valencia A, Pazos F. Prediction of protein-protein interactions from evolutionary information. *Structural Bioinformatics, Second Edition*, 617–634, Wiley and Sons: New York, 2009.

[10] Tuncbag N, Gursoy A, Nussinov R, Keskin O. Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nat. Protoc.*, 6(9), 1341–1354, 2011.

[11] Matsuzaki Y, Matsuzaki Y, Sato T, Akiyama Y. *In silico* screening of protein-protein interactions with all-to-all rigid docking and clustering: an application to pathway analysis. *J. Bioinform. Comput. Biol.*, 7(6), 991–1012, 2009.

[12] Ohue M, Matsuzaki Y, Ishida T, Akiyama Y. Improvement of the protein-protein docking prediction by introducing a simple hydrophobic interaction model: an application to interaction pathway analysis. *Lecture Notes in Computer Science*, 7632, 178–187, 2012.

[13] Ohue M, Matsuzaki Y, Shimoda T, Ishida T, Akiyama Y. Highly precise protein-protein interaction prediction based on consensus between template-based and *de novo* docking methods. In *Proc. Great Lakes Bioinformatics Conference 2013*, 100–109, 2013.

[14] Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, 6(3), 197–208, 2005.

[15] Vapnik VN. *The Nature of Statistical Learning Theory*. Springer: New York, 1995.

[16] Mintseris J, Pierce B, Wiehe K, Anderson R, Chen R, Weng Z. Integrating statistical pair potentials into protein complex prediction. *Proteins*, 69(3), 511–520, 2007.

[17] Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res.*, 33(suppl 2), W363–367, 2005.

[18] 大上雅史, 松崎由理, 松崎裕介, 佐藤智之, 秋山泰. MEGADOCK：立体構造情報からの網羅的タンパク質間相互作用予測とそのシステム生物学への応用. 情報処理学会論文誌 数理モデル化と応用, 3(3), 91–106, 2010.

[19] Kozakov D, Brenke R, Comeau SR, Vajda S. PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins*, 65(2), 392–406, 2006.

[20] Guo M, Shapiro R, Morris GM, Yang XL, Schimmel P. Packaging HIV virion components through dynamic equilibria of a human tRNA synthetase. *J. Phys. Chem. B.* 114(49), 16273–16279, 2010.

[21] Chen R, Robinson A, Gordon D, Chung SH. Modeling the binding of three toxins to the voltage-gated potassium channel (Kv1.3). *Biophysical J.*, 101(11), 2652–2660, 2011.

[22] Platt JC. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers*, 61–74, MIT Press, 1999.

[23] Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, Janin J, Weng Z. Protein-protein docking benchmark 2.0: an update. *Proteins*, 60(2), 214–216, 2005.

[24] Chang CC, Lin C-J. LIBSVM: A library for support vector machines. *ACM Trans. on Intelligent Systems and Tech.*, 2(3), Article 27, 2011.

[25] Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39(4), 561–577, 1983.

[26] Wuchty S, Oltvai ZN, Barabási AL. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat. Genet.*, 35(2), 176–179, 2003.

[27] Smialowski P, Pagel P, Wong P, Brauner B, Dunger I, Fobo G, Frishman G, Montrone C, Rattei T, Frishman D, Ruepp A. The Negatome database: a reference set of non-interacting protein pairs. *Nucleic Acids Res.*, 38(suppl 1), D540–544, 2010.