Vol.2014-MPS-98 No.12
Vol.2014-BIO-38 No.12
2014/6/26

タンパク質ドメインの特徴を用いたカーネルマシンによるタンパク質間相互作用強度予測

鎌田　真由美[1,a)]　佐久間　裕介[2]　林田　守広[3]　阿久津　達也[3]

概要：タンパク質は生体内において他のタンパク質や分子と結合する事で様々な機能を実現している．ゆえに，タンパク質間相互作用（PPI）に関する実験研究や計算機による予測手法開発がこれまでに多く行われている．更に，PPI を相互作用するか・しないかの 2 値ではなく，多値として扱う為に導入された，PPI の強度という概念がある．この PPI 強度がタンパク質の特異性や機能性にも関連すると考えられている．そこで本研究では，アミノ酸配列から PPI 強度を予測するために，タンパク質ペアに対するタンパク質ドメイン情報を用いた特徴ベクトルを提案する．そして，サポートベクター回帰と関連ベクトルマシンを用い，生物学実験により得られたデータに対して PPI 強度予測を行った．計算機実験の結果，我々の提案手法は既存手法よりも高い予測精度を示した．

1. はじめに

生体内においてタンパク質の多くは他の分子やタンパク質と相互作用することでその機能を発現している．つまり，タンパク質間相互作用（Protein-protein interaction, PPI）は我々の生命活動における重要な生体反応であり，その理解は生命システムを理解する上で重要な鍵となる．ゆえに，これまで多くの研究解析が行われており，計算機を用いた予測手法も多数開発されている．PPI は通常，相互作用するか・しないかで議論されることが多いが，その割合，つまり相互作用の強度もタンパク質複合体形成やその機能性に大きく関与しており，同様に重要である．例えば，転写因子複合体では，構成タンパク質の結合能が弱い場合，細胞内の環境によってはターゲット遺伝子は転写されないことがある．また，酵母菌から分離された NuA3 は，遺伝子の転写制御に関与するヒストンアセチル化酵素の 1 つであり，5 つのタンパク質 Sas3，Nto1，Yng1，Eaf6，Taf30 が結合した複合体として機能を果たすことがわかっている．しかしながら，NuA3 複合体内で Yng1 と Nto1 は単独に見つかっており，複合体内の各タンパク質間の結合能の違いが機能発現に関与していると考えられる．NuA3 複合体については未だ多くが明らかになっていないが，近年，このような過渡的なタンパク質間の相互作用とその安定性を同定する為の生物学的実験手法 [1] も提案されている．

PPI を理解する為にこれまでに多くの生物学実験 [2], [3] が構築されてきているが，PPI 強度が得られるものは限られている．PPI 強度に関連した実験として，T. Ito が酵母菌の全遺伝子に対して行った，大規模 two-hybrid 実験がある．この実験では，各々のタンパク質ペアに対して yeast two-hybrid 実験を複数回行い，相互作用が観測された実験回数を Interaction Sequence Tags（IST）としてカウントした．そして，IST が 3 以上のタンパク質ペアを相互作用していると判断し，それらのペアを相互作用タンパク質として発表した．

各タンパク質ペアの総実験回数に対する IST 数の割合は，そのペアの相互作用の強さとして捉える事が出来る．これに基づき，PPI 強度の予測手法がいくつか開発されている．LPNM[4] は PPI の割合を PPI 強度として初めて導入したものであり，線形計画法をベースとした予測手法である．また，E. Sprinzak によって提案された PPI 予測モデルである Association Method[5] を改良した ASNM[6] がある．更に L. Chen は ASNM を改良し，Association probabilistic method（APM）[7] を提案した．現在我々の知る既存手法では，APM が最も優れた PPI 強度予測手法である．これらの手法はどれもタンパク質の構造・機能のユニットであるドメインの情報を利用している．タンパク質ドメインはアミノ酸配列に進化上よく保存された領域であることが知られており，異なるタンパク質内で共通したドメインが同定されることが多くある．これらのドメイン情報は，Pfam[8] や InterPro[9] 等のデータベースに集められ，公開されている．上述した予測手法では，相互作用す

[1] 慶應義塾大学　理工学部　生命情報学科
[2] 楽天株式会社
[3] 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
[a)] kamada@dna.bio.keio.ac.jp

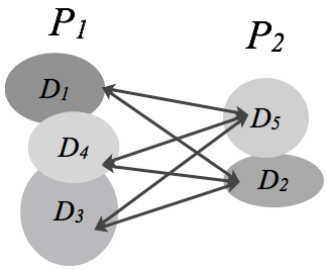

図 1 ドメイン間相互作用に基づくタンパク質間相互作用

ることが知られている既知のタンパク質ペアからドメイン間相互作用（Domain-domain interaction, DDI）を推定し，それを用いて新規のタンパク質ペアに対し PPI 強度を予測している．

本研究では，PPI 強度予測の為に，上記の手法と同様にドメイン情報を活用し，タンパク質ペアの特徴空間マッピングをいくつか提案する．更に，生物学実験から得られるデータセットに対し，提案する特徴ベクトルと教師あり学習によって PPI 強度予測を行う．本研究は国際会議で発表した導入研究 [10] を改良したものであり，主に異なる点は以下の通りである．

- 本研究では，2 つのカーネルマシン，サポートベクター回帰（SVR）と関連ベクトルマシン（RVM）を用いて PPI 強度の予測を行う．
- SVR，RVM の両手法に対し，ラプラシアンカーネルを用い，カーネルパラメータは 5-fold 交差検定により決定する．
- 公開データである WI-PHI データ [11] から高い信頼性が保たれるようデータセットの作成を行う．

計算機実験の結果，二乗平均平方根誤差（RMSE）の値は我々の提案手法が最も小さく，既存手法である APM の精度を上回ることを示した．

2. 提案手法

この節では PPI の確率モデルと関連する手法について簡単に紹介し，我々の提案するタンパク質ドメイン情報を用いた特徴空間マッピングについて述べる．

2.1 ドメイン間相互作用に基づくタンパク質間相互作用の確率モデル

PPI 強度の予測手法のいくつかは，M. Deng によって提案された PPI の確率モデル [12] に基づいている．このモデルは DDI を利用したもので，ある 2 つのタンパク質が相互作用する場合，それらに含まれるドメインの組み合わせの内，少なくとも 1 つが相互作用しているという仮定に基づいている．**図 1** に DDI に基づく PPI モデルの具体例を示す．この例では，ドメイン D_1，D_3，D_4 を持つタンパク質 P_1 と，ドメイン D_2，D_5 を持つタンパク質 P_2 間の関係を示している．Deng のモデルに基づくと，タンパク質 P_1 と P_2 が相互作用するならば，6 つ考えられるドメインペアの内，少なくとも 1 つが相互作用していることになる．逆に言えば，例えば，D_2 と D_4 が相互作用するならば，タンパク質 P_1 と P_2 は相互作用している，という事が出来る．このモデルの仮定から，2 つのタンパク質 P_i と P_j の相互作用に対し，以下のようなシンプルな確率モデルを考える事が出来る．

$$\Pr(P_{ij}=1) = 1 - \prod_{D_{mn}\in P_{ij}} (1 - \Pr(D_{mn}=1)) \qquad (1)$$

ここで，$P_{ij}=1$ はタンパク質 P_i と P_j が相互作用することを表しており（そうでない場合は $P_{ij}=0$），$D_{mn}=1$ も同様にドメイン D_m と D_n が相互作用することを示している（相互作用しない場合は $D_{mn}=0$）．また，P_i と P_j はそれぞれ，タンパク質 P_i と P_j を各々構成するドメインの集合も表わしている．Deng は，EM アルゴリズムを用いた尤度関数最大化によって 2 つのドメインの相互作用確率 $Pr(D_{mn}=1)$ を推定し，推定される DDI の確率を用いて PPI を予測する手法を提案した [12]．実際に，彼らは式 (1) を用いて $\Pr(P_{ij}=1)$ を計算し，閾値 θ を用いて $\Pr(P_{ij}=1) \geq \theta$ となるタンパク質 P_i と P_j を相互作用する，そうでない場合は相互作用しない，として予測を行った．

上記の Deng の手法を始めとするドメインに基づく PPI 予測手法は，大きく以下の 2 つのステップを行っている．まず，新規タンパク質ペアを構成するドメインの全てのペアについて，その相互作用を既知のタンパク質間相互作用データに基づき推定する．そして，推定したドメイン間相互作用とそれに適したモデルを用い，新規タンパク質ペアに対する相互作用予測を行う．しかし，相互作用部位が常に既知のドメイン領域に含まれているとは限らないことから，このフレームワークではしばしば予測精度の低下を引き起こすことがある．

2.2 既知の PPI データによる DDI 確率の推定

上述のように，新規タンパク質ペアに対する PPI の確率は，DDI の確率に基づき予測する事が出来る．ここでは，ドメインペアの相互作用確率を推定する手法について簡単に説明する．

- **Association Method**
  $\mathcal{P}$ を相互作用するか否かが観測されているタンパク質ペアの集合とする．E. Sprinzak が提案した Association Method[5] では，2 つのドメイン D_m と D_n に対するスコアを，各々を含むタンパク質を用いて下記のように計算する．

  $$ASSOC(D_m, D_n) = \frac{I_{mn}}{N_{mn}}. \qquad (2)$$

  ここで N_{mn} はドメインペア (D_m, D_n) を含んでいる

タンパク質ペア $(P_i, P_j) \in \mathcal{P}$ の数，I_{mn} はドメインペア (D_m, D_n) を含んでおり且つ相互作用するタンパク質ペア $(P_i, P_j) \in \mathcal{P}$ の数である．見て明らかなように，このスコアは D_m と D_n を含んでいる全てのタンパク質ペアにおいて，そのうちいくつのタンパク質ペアが相互作用しているのかを割合として表わしていることから，D_m と D_n の相互作用する確率として考える事が出来る．

- **Association method for NuMerical interaction（ASNM）**
  オリジナルの Association Method は，タンパク質相互作用を相互作用するか・しないかの，バイナリで予測する為に考えられた．そこで，相互作用の強度や割合などの数値的な相互作用を予測する為に，M. Hayashida は, オリジナルの Association Method を改良した ASsociation method for NuMerical interaction（ASNM）を提案した [6]．この手法では，PPI 強度を入力として用いる．ここで，ρ_{ij} をタンパク質 P_i と P_j 間の相互作用強度とし，ρ_{ij} は全てのタンパク質ペア $(P_i, P_j) \in \mathcal{P}$ に対して定義されているとする．ドメインペア (D_m,D_n) に対する ASNM スコアは，D_m，D_n を含むタンパク質ペアにおける平均強度として以下のように定義される．

$$ASNM(D_m, D_n) = \frac{\sum_{P_{ij}:D_{mn}\in P_{ij}} \rho_{ij}}{N_{mn}} \qquad (3)$$

ここで，ρ_{ij} が常に 0 か 1 の値を取る場合，$ASNM(D_m, D_n)$ は $ASSOC(D_m, D_n)$ と等しくなる．

- **Association Probabilistic Method（APM）**
  ASNM は PPI 強度の平均をとるシンプルなものであったが，L. Chen は，PPI 強度を更に改良した値に置き換えた Association Probabilistic Method（APM）[7] を提案した．これは，PPI 強度に対する 1 つ 1 つのドメインペアの貢献度が，タンパク質ペアに含まれるドメインペア数によって異なる，という考えに基づいている．彼らは各ドメインペアの相互作用確率がタンパク質ペアにおいては等しいと仮定し，式 (1) を下記のように変更した．

$$\Pr(D_{mn} = 1) = 1 - (1 - \Pr(P_{ij} = 1))^{\frac{1}{|P_i||P_j|}}. \qquad (4)$$

ここで $|S|$ は集合 S に含まれる要素の数を表わしている．そして式 (4) に基づき ASNM の PPI 強度の部分を置換し，以下のように APM を定義した．

$$APM(D_m, D_n) = \frac{\sum_{P_{ij}:D_{mn}\in P_{ij}} [1 - (1 - \rho_{ij})^{\frac{1}{|P_i||P_j|}}]}{N_{mn}}. \qquad (5)$$

L. Chen は，いくつかの計算機実験を行い，APM が予測精度において ASNM や LPNM 等の他の予測手法を上回ることを示した．

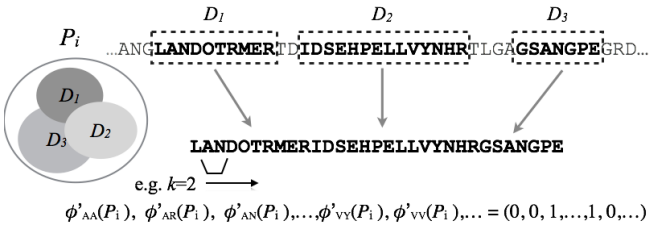

図 2 ドメイン領域に制限されたアミノ酸配列に対するスペクトルカーネルの適用

2.3 提案手法

これ以降，本研究で提案する特徴空間マッピングについて述べる．

2.3.1 ドメイン数に基づく特徴ベクトル（DN）

まず，タンパク質のドメイン数に基づいた特徴空間マッピングを提案する．ここでは DN と呼ぶ事とする．タンパク質が相互作用する確率は，それらのタンパク質に含まれるドメインの数が増えるに従い高くなると考えられる．そこで，タンパク質 P_i と P_j に対する DN の特徴ベクトルを以下の様に定義する．

$$f_{ij}^{(m)} = M(D_m, P_i) \quad \text{for } D_m \in P_i, \qquad (6)$$
$$f_{ij}^{(T+n)} = M(D_n, P_j) \quad \text{for } D_n \in P_j, \qquad (7)$$
$$f_{ij}^{(l)} = 0 \quad \text{for } D_l \notin P_i \cup P_j. \qquad (8)$$

ここで，T は全てのタンパク質におけるドメインの総数を表わしており，$M(D_m, P_i)$ は，タンパク質 P_i でドメイン D_m として同定されるドメイン数を表わしている．

2.3.2 ドメイン領域制限付きスペクトルカーネルによる特徴ベクトル（SPD）

本研究では更に，スペクトルカーネル [13] を応用し，ドメイン領域に制限をかけた特徴空間マッピングを提案する．ここでは，$\mathcal{A}$ を 20 種類のアミノ酸とその他，例えば複数の解釈ができるアミノ酸 X 等を表現する為の，合計 21 のアルファベットの集合とし，$\mathcal{A}^k$ $(k \geq 1)$ を集合 $\mathcal{A}$ から生成される長さ k の文字列の集合とする．ここで，文字列配列 x，y に対する k-スペクトルカーネルは下記のように定義される．

$$K_k(x, y) = \langle \Phi_k(x), \Phi_k(y) \rangle. \qquad (9)$$

$\Phi_k(x) = (\phi_s(x))_{s\in\mathcal{A}^k}$ と $\phi_s(x)$ は，文字列配列 x において生成される文字列 s の数を表わしている．

タンパク質ドメイン情報を活用する為，k スペクトルカーネルをドメイン領域に適用できるよう，タンパク質のアミノ酸配列に制限をかける．**図 2** に例を示す．この例えでは，タンパク質 P_i はドメイン D_1，D_2，D_3 を含んでおり，各ドメインの領域が破線四角によって囲まれている．

このドメインに対応する囲まれた領域を部分文字列として抜き出し，対象のタンパク質に含まれる全ての部分文字列をドメインの並びと同順に，1 つの文字列に連結する．そして，この連結された文字列に対して k スペクトルカーネルを適用する．ここで $\phi_s^{'}(x)$ を，タンパク質 x のドメイン領域に制限された文字列（アミノ酸配列）における文字列 s の出現頻度とすると，タンパク質 P_i と P_j に対する SPD の特徴ベクトルは下記のように定義される．

$$f_{ij}^{l} = \phi_{s_l}^{'}(P_i) \text{ for } s_l \in \mathcal{A}^k, \quad (10)$$

$$f_{ij}^{(21^k+l)} = \phi_{s_l}^{'}(P_j) \text{ for } s_l \in \mathcal{A}^k. \quad (11)$$

同じドメイン構成を持つタンパク質に対する $\phi_s^{'}$ は，それらタンパク質のアミノ酸配列によって異なる値を持つことを注記しておく．つまり，例えばタンパク質 P_i，P_j が，タンパク質 P_k，P_l と全く同じドメインによって構成されているとした時，DN による P_i，P_j に対する特徴ベクトルは P_k，P_l に対するそれと同じであるが，SPD による特徴ベクトルはしばしば異なるものになる．

2.4 サポートベクター回帰

PPI 強度を予測する為, 本研究では提案する特徴ベクトルとサポートベクター回帰（support vector regression, SVR）[14] を用いる．線形関数の場合，SVR は $f(x) = \langle w, x \rangle + b$ に対するパラメータ w と b を以下の最適化問題を解く事で決定する．

$$\begin{aligned} \text{minimize } & \frac{1}{2}||w||^2 + C\sum_i(\xi_i + \xi_i^{'}), \\ \text{subject to } & y_i - \langle w, x_i \rangle - b \leq \epsilon + \xi_i, \\ & y_i - \langle w, x_i \rangle - b \geq -\epsilon - \xi_i^{'}, \\ & \xi_i \geq 0,\ \xi_i^{'} \geq 0. \end{aligned}$$

ここで，C と ϵ は正の定数，(x_i, y_i) は訓練データであり，$f(x_i)$ と y_i の差が ϵ よりも大きい場合にのみ罰則が与えられる．本稿の PPI 強度の問題においては，x_i がタンパク質ペア，y_i が相互作用強度に対応している．

2.5 関連ベクトルマシン

本研究では SVR の他に, 関連ベクトルマシン（Relevance Vector Machine, RVM）[15] を用いて予測を行う．RVM は，SVM と同様のデータ依存のカーネルバイアスを用いている，ベイジアンモデルである．訓練データ $\{x_i, y_i\}_{i=0}^N$ が与えられた時，x による y の条件付き確率は下記の様にモデル化する事が出来る．

$$p(y|x, w, \beta) = \mathcal{N}(y|w^T\phi(x), \beta^{-1}).$$

ここで $\beta = \sigma^2$ はノイズパラメータ，$\phi(.)$ は入力された特徴の非線形写像である．RVM のフレームワークでは疎な解を得る為に，各重みに異なる分散パラメータが割り当てられるように重みの事前分布が下記の形で与えられる．

$$p(w|\alpha) = \prod_{i=0}\mathcal{N}(w_i|0, \alpha_i^{-1}).$$

ここで $M = N + 1$，$\alpha = (\alpha_1, ..., \alpha_M)$ は超パラメータである．RVM は超パラメータ α を，エビデンス近似による周辺尤度 $p(y|x, \alpha)$ の最大化で決定する．エビデンスの最大化において，いくつかの α_i が無限に近づき，これに対応する重み w_i がゼロに抑えられる．よってこれらのパラメータに対応する基底関数を取り除くことが出来，疎なモデルが導かれる．特に回帰問題において多くの場合に，RVM は SVM よりも良い予測精度をもたらすと報告されている．

3. 結果と考察

3.1 計算機実験

提案手法を評価する為に，計算機実験を行い既存手法 APM と精度の比較を行う．

3.1.1 データと実装

多くのタンパク質ペアに対して生物学実験によって直接的に PPI 強度を計測する事は難しい．そこで本研究では，50000 のタンパク質ペアによる WI-PHI データセット [11] を用いる．WI-PHI では，各 PPI に対して，PPI の信頼性を表現するよう考えられた “重み” を提供している．この重みは，複数の異なる PPI データセットから物理的なタンパク質の相互作用に対して統計的にランク付けを行い計算される．ここでは, WI-PHI により得られる重みをその最大値で割った値を, PPI 強度として用いる．また, タンパク質のアミノ酸配列や構成ドメイン, アミノ酸配列中のドメイン領域の情報をデータベース UniProt[16] から取得する．本研究で用いたデータファイルは “uniprot_sprot_fungi.dat.gz” である．

SPD を計算するには，タンパク質に含まれる全てのドメインに対して，その配列情報をきちんと取得することが必要となる．故に評価実験では，WI-PHI データに含まれるタンパク質ペアの内，UniProt データセットを用いて完全な配列を取得出来た 327 のドメインから成る，1387 のタンパク質ペアを用いる．また，WI-PHI データには相互作用強度が 0 のタンパク質ペアが含まれていないため，重みを持っていないタンパク質ペアをランダムに 100 個抽出し，PPI 強度 0 のタンパク質ペアとしてデータに追加した．

SVR と RVM は，R の “kernlab” パッケージ [17] により実装し，カーネルにはラプラシアンカーネル $K(x, y) = \exp(-\sigma||x - y||)$ を用いた．予測精度を評価する為，各予測に対して二乗平均平方根誤差（Root Mean Square Error, RMSE）を計算する．RMSE は予測値 $\hat{y}_i$ と実測値 y_i の差を計算するものであり，

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2},$$

として定義される．N はテストデータの数である．

3.1.2 計算機実験の結果

本研究では提案する 2 つの特徴ベクトルと SVR・RVM を用いて 3-fold の交差検定を行い，平均 RMSE の値によって既存手法の APM[7] と比較を行った．

APM 手法において PPI 強度は，対象タンパク質に含まれる各ドメインペアの APM スコアに基づき推定される．しかし，訓練データから常に全てのドメインペアの APM スコアが計算出来る訳では無い．故に本稿の実験では，訓練データによって APM スコアが計算されるドメインのペアで構成されるタンパク質ペアのみをテストデータとして用い，APM 手法と提案手法を評価した．ほとんど全ての場合で，元のテストデータの約 40%のタンパク質ペアが評価用テストデータとして用いられた．SVR と RVM で用いたラプラシアンカーネルは，カーネルパラメータとして σ を持つ．このパラメータ σ は，候補セット $\sigma \in \{0.01, 0.02, ..., 0.1\}$ の中から，5-fold 交差検定によって決定した．正則化項パラメータ C は，$C = 2$ を用いた．

訓練データと評価用テストデータに対する平均 RMSE を **表 1** に示す．表 1 には，DN による特徴ベクトルと k=1,2 とした SPD による特徴ベクトルの合計 3 つの特徴ベクトルを SVR と RVM に用いた結果と APM の結果を示している．訓練データでは，SPD を用いた RVM による平均 RMSE が他と比較して最も小さな値を示している．更に，評価用テストデータに対しては，我々の提案手法による全ての平均 RMSE が APM よりも小さな値となった．これらの結果から，SPD による特徴ベクトルと RVM を用いた場合に，最もよく PPI 強度を予測出来ていると見なす事ができる．

また，DN と k=1 の SPD は各々 654 と 42 次元の特徴ベクトルで各タンパク質ペアを表現しているが，訓練データに対する k=1 の SPD の平均 RMSE は DN の平均 RMSE よりも小さな値になっている．これは，データセットに即したモデルを構築するには，ドメインの構成情報よりもドメイン領域のアミノ酸配列情報の方がより有益であることを示していると言える．しかし一方で，訓練データに対して RVM+SPD が他に比べて圧倒的に当てはまりが良かったにも関わらず，テストデータでは他とあまり大きな差が見られない．特徴ベクトルの値に着目してみると，ドメイン数に基づく DN では対象としているタンパク質のほとんどが 1~3 個のドメインで構成される事から，ほぼ全ての次元で値 0 を持つ．一方，k=1 の SPD では各次元がアミノ酸頻度を表す事から，ほぼ全てで 1 以上の値を持っている．このことから，重みの多くが $w_i \neq 0$ となることから関連ベクトルの数が大きくなり，訓練データに過適合する傾向にあると考えられる．

表 1 訓練データと評価用テストデータに対する平均 RMSE 値

	Training	Test
SVR+DN	0.107	0.126
RVM+DN	0.092	0.129
SVR+SPD(k=1)	0.081	0.130
RVM+SPD(k=1)	**0.015**	0.127
SVR+SPD(k=2)	0.082	0.128
RVM+SPD(k=2)	0.023	**0.125**
APM	0.068	0.135

4. まとめ

本研究では，タンパク質間相互作用強度（PPI 強度）を予測する為，DN と SPD の 2 つの特徴空間マッピングを提案した．DN はタンパク質ドメインの数によって定義され，SPD はスペクトルカーネルとドメイン領域に制限したアミノ酸配列を用いて定義される．本研究では，PPI 強度の予測に，サポートベクター回帰（SVR）と関連ベクトルマシン（RVM）の 2 つカーネルマシンを用い，WI-PHI データに対して 3-fold 交差検定による計算機実験を行った．平均 RMSE を用いて評価した結果，訓練データとテストデータ共に，SPD を用いた RVM による予測が最も精度がよく，これは既存手法の APM を上回るものであった．計算機実験の結果から，ドメイン情報を用いた機械学習手法は，Association Method に基づく既存手法を上回る精度を示し，また，アミノ酸配列情報がドメインの構成情報のみよりも PPI 強度の予測には有益であることを示した．しかしながら，SPD の特徴ベクトルを用いた場合に，モデルが複雑になる傾向にあり訓練データに過適合していると考えられた．故に，今後さらに予測精度を向上させる為に，タンパク質ドメインとアミノ酸の物理的な特徴を組み合わせ，カーネル関数を改良することが有効であると考えている．

謝辞 本研究は，JSPS 科研費 #22240009,#24500361 の助成を受けたものです．

参考文献

[1] Byrum, S., Smart, S., Larson, S. and Tackett, A.: Analysis of stable and transient protein-protein interactions, *Methods in Molecular Biology*, Vol. 833, pp. 143–152 (2012).

[2] Uetz, P., Giot, L., Cagney, G., Mansfield, T., Judson, R., Knight, J., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. and Rothberg, J.: A comrehensive two-hybrid analysis to explore the yeast protein interactome, *Nature*, Vol. 403, pp. 623–627 (2000).

[3] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y.: A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proceedings of the National Academy of Sciences of USA*, Vol. 98, pp. 4569–4574 (2001).

[4] Hayashida, M., Ueda, N. and Akutsu, T.: Inferring strengths of protein-protein interactions from experimental data using linear programming, *Bioinformatics*, Vol. 19, pp. ii58–ii65 (2003).
[5] Sprinzak, E. and Margalit, H.: Correlated sequence-signatures as markers of protein-protein interaction, *Journal of Molecular Biology*, Vol. 311, pp. 681–692 (2001).
[6] Hayashida, M., Ueda, N. and Akutsu, T.: A simple method for inferring strengths of protein-protein interaction, *Genome Informatics*, No. 15, pp. 56–68 (2004).
[7] Chen, L., Wu, L., Wang, Y. and Zhang, X.: Inferring protein interactions from experimental data by association probabilistic method, *Proteins: Structure, Function, and Bioinformatics*, Vol. 62, pp. 833–837 (2006).
[8] Finn, R., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J., Gavin, O., Gunesekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E., Eddy, S. and Bateman, A.: The Pfam protein families database, *Nucleic Acids Research*, Vol. 38, pp. D211–D222 (2010).
[9] Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S., de Castro, E., Coggill, P., Corbett, M., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Fraser, M., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., McMenamin, C., Mi, H., Mutowo-Muellenet, P., Mulder, N., Natale, D., Orengo, C., Pesseat, S., Punta, M., Quinn, A. F., Rivoire, C., Sangrador-Vegas, A., Selengut, J. D., Sigrist, C. J. A., Scheremetjew, M., Tate, J., Thimmajanarthanan, M., Thomas, P. D., Wu, C. H., Yeats, C. and Yong, S.: InterPro in 2011: new developments in the family and domain prediction database, *Nucleic Acids Research*, Vol. 40, pp. D306–D312 (2012).
[10] Sakuma, Y., Kamada, M., Hayashida, M. and Akutsu, T.: Inferring strengths of protein-protein interactions using support vector regression, in Proceedings of International Conference on Parallel and Distributed Processing Techniques and Applications 2013, http://world-comp.org/p2013/PDP2162.pdf (2013).
[11] Kiemer, L., Costa, S., Ueffing, M. and Cesareni, G.: WI-PHI: A weighted yeast interactome enriched for direct physical interactions, *Proteomics*, Vol. 7, pp. 932–943 (2007).
[12] Deng, M., Mehta, S., Sun, F. and Chen, T.: Inferring domain-domain interactions from protein-protein interactions, *Genome Research*, Vol. 12, pp. 1540–1548 (2002).
[13] Leslie, C., Eskin, E. and Noble, W.: The spectrum kernel: a string kernel for SVM protein classification, *in Proceedings of Pacific Symposium on Biocomputing 2002*, pp. 564–575 (2002).
[14] Vapnik, V.: *The Nature of Statistical Learning Theory*, Springer (1995).
[15] Tipping, M. E.: Sparse Bayesian learning and the relevance vector machine, *Journal of Machine Learning Research*, Vol. 1, pp. 211–244 (2001).
[16] Consortium, T. U.: Reorganizing the protein space at the Universal Protein Resource (UniProt), *Nucleic Acids Research*, Vol. 40, pp. D71–D75 (2012).
[17] Karatzoglou, A., Smola, A., Hornik, K. and Zeileis, A.: kernlab – An S4 Package for Kernel Methods in R, *Journal of Statistical Software*, Vol. 11, No. 9, pp. 1–20 (2004).