Vol.2019-BIO-57 No.12
2019/3/9

代表タンパク質構造群との構造アラインメントスコアプロファイルに基づくタンパク質間相互作用予測の高速化

林 孝紀[1]　大上 雅史[1]　秋山 泰[1,a)]

概要：タンパク質間相互作用 (PPI) は生命現象を理解する上で重要な役割を果たしており，計算機による PPI 予測が注目されている．2010 年に Hue らはタンパク質の構造アラインメントから算出される立体構造の類似度を用いたカーネル法によって，アミノ酸配列に基づく手法に比べて予測精度の向上を実現した．しかし Hue らの手法には以下の 2 つの問題点が存在する：(1) タンパク質間グラム行列の作成に全組み合わせの構造アラインメントを行うため計算時間がかかる．(2) 新規タンパク質を予測するときに類似度行列の半正定値化変換と全データでの学習を都度行う必要がある．本研究では，あらかじめ用意した少数の代表的なタンパク質構造群（タンパク質ライブラリ）との構造アラインメントから得られるベクトル表現によって，擬似的な構造類似度を計算する手法を提案する．提案手法はタンパク質ライブラリのみとの構造アラインメントを行えばよく，問題点 (1) を解決可能である．さらに，半正定値性が保たれる一般的な実数ベクトル間の類似度を用いることができ，行列の変換が不要となるため問題点 (2) も解決される．先行研究に基づいて作成した 2 つの評価データセットを使用した計算機実験により，提案手法が Hue らの手法に比べて予測精度ではわずかに劣るものの，5–50 倍の高速化が可能であることを示した．

キーワード：タンパク質間相互作用，PPI 予測，機械学習，ペアワイズカーネル

A Fast Protein–Protein Interaction Prediction Method with a Small Number of Representatives

Takanori Hayashi[1]　Masahito Ohue[1]　Yutaka Akiyama[1,a)]

Abstract: Protein–protein interaction (PPI) plays an important role in understanding biological phenomena, and prediction of PPI by computers is required. Hue *et al.* achieved prediction accuracy better than the method based on k-mer vectors of amino acid sequence by using the structural similarity calculated from structural alignment and pairwise kernel in 2010 . However, Hue *et al.* method has the following two problems; (1) it takes large time to perform structural alignment of all combinations to make protein Gram matrix, and (2) when predicting a new protein, it is needed to transform the similarity matrix and learn every time. In this study, we propose a method to calculate pseudo structural similarity by vector obtained from structural alignment with only a small number of representative proteins (a protein library). In the proposed method, it is only necessary to perform structural alignment with the protein library, so the problem (1) can be solved. Furthermore, by using a function with semidefinite property, it becomes possible to predict new proteins without matrix transformation, and the problem (2) can be solved. In addition, the computer experiment based on the two data sets showed that the proposed method has almost the same accuracy and shorter time as the method of Hue *et al.*'s method.

Keywords: protein-protein interaction (PPI), PPI prediction, machine learning, pairwise kernel

[1] 東京工業大学 情報理工学院 情報工学系，
Department of Computer Science, School of Computing, Tokyo Institute of Technology
[a)] akiyama@c.titech.ac.jp

1. 導入

生体内のタンパク質は互いに相互作用しながら機能を

発揮している [1]．この相互作用はタンパク質間相互作用 (protein–protein interaction, PPI) と呼ばれており生命現象の中核を担っている．また，近年では疾患の要因とされる PPI を阻害することで効果を得る PPI 阻害薬の開発も進んでいる [2,3]．しかし多くのタンパク質に対して実験的に PPI を決定するのは金銭的，時間的なコストが高くなるため，計算機による PPI 予測が注目されている．

計算機による PPI 予測には既知の PPI 情報を用いない手法 [4,5] と既知の PPI 情報を用いる手法 [6,7] が存在するが，後者の既知の PPI 情報を用いる手法はまったく新しい PPI を発見することが難しい一方で，予測精度が高いとされている．既知の PPI 情報を用いる手法には，アミノ酸配列情報を用いる手法 [8,9]，遺伝子情報を用いる手法 [10]，立体構造情報を用いる手法 [11,12] が存在する．しかし，タンパク質の相互作用はその立体構造に基づく物理化学現象であり，アミノ酸配列情報のみから捉えることは難しい [13]．実際にアミノ酸配列の類似性が低いもかかわらず立体構造がよく似た複合体も存在する [14,15]．2010 年に Hue らは構造アラインメントから算出される立体構造の類似度と機械学習を用いることでアミノ酸配列の k-mer に基づく手法 [8] に比べて予測精度の向上を実現した [16]．しかし，Hue らの手法には，タンパク質間類似度行列の作成に時間がかかる，新規タンパク質予測時に学習を都度やり直す必要がある，という 2 つの問題点が存在する．

本研究では，あらかじめ用意した少数の代表的なタンパク質構造群（タンパク質ライブラリ）との構造アラインメントから得られるベクトル表現によって，擬似的な構造類似度を計算する手法を提案することで，これらの問題点を解決した PPI 予測手法を開発することを目的とする．

2. 先行研究

2.1 ペアワイズカーネル

PPI 予測の問題は入力タンパク質 (P_1, P_2) のペアに対して，相互作用するかどうかの分類問題と捉えることができる．この問題をカーネル法を用いてアプローチする場合，なんらかの方法をもちいてタンパク質ペア間グラム行列を定義する必要がある．タンパク質間グラム行列からタンパク質ペア間グラム行列を作成する方法が提案されており，ペアワイズカーネルとよぶ．ペアワイズカーネルの具体的な計算方法として，Ben-Hur らは Tensor Product Pairwise Kernel (TPPK) [18] を提案し，Vert らは Metric Learning Pairwise Kernel (MLPK) [19] を提案した．TPPK と MKPK は

$$
\begin{aligned}
&k_{TPPK}((\boldsymbol{x}_1, \boldsymbol{x}_2), (\boldsymbol{x}_3, \boldsymbol{x}_4)) \\
&= k(\boldsymbol{x}_1, \boldsymbol{x}_3)k(\boldsymbol{x}_2, \boldsymbol{x}_4) + k(\boldsymbol{x}_1, \boldsymbol{x}_4)k(\boldsymbol{x}_2, \boldsymbol{x}_3) \\
&k_{MLPK}((\boldsymbol{x}_1, \boldsymbol{x}_2), (\boldsymbol{x}_3, \boldsymbol{x}_4)) \\
&= (k(\boldsymbol{x}_1, \boldsymbol{x}_3) + k(\boldsymbol{x}_2, \boldsymbol{x}_4) - k(\boldsymbol{x}_1, \boldsymbol{x}_4) - k(\boldsymbol{x}_2, \boldsymbol{x}_3))^2
\end{aligned}
$$

と定義される．ただし，$k(\cdot,\cdot)$ はタンパク質間カーネル関数，$\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4$ はタンパク質の特徴ベクトルをあらわす．ペアワイズカーネルは PPI 予測のほか，薬剤の活性予測などにも応用されている [20]．

2.2 構造アラインメントを用いた PPI 予測

Hue らは構造アラインメントを用いて計算されたタンパク質間グラム行列を用いて PPI 予測を行う手法を開発した [16]．以下に具体的な PPI 予測手順を示す．

Hue *et al.* の PPI 予測手法 [16]

Step 1: データセットに含まれる N 個のタンパク質に対して，タンパク質 P_i, P_j の類似度 m_{ij} を構造アラインメントを行ったときの E-value を E_{ij} として，$m_{ij} = \max(20, -\ln E_{ij})$ とおく．

Step 2: m_{ij} を並べた $N \times N$ 行列 $\mathbf{M} = (m_{ij})$ を作成する．

Step 3: 行列 $\mathbf{M}$ に対して固有値分解を行い，$\mathbf{M} = \mathbf{UDU}^\top$ とする．このとき $\mathbf{D}$ は $\mathbf{M}$ の固有値を $\lambda_l (l = 1, ..., N)$ として，$\mathbf{D} = diag(\lambda_1, ..., \lambda_N)$（対称行列）である．

Step 4: 行列 $\mathbf{M}$ を半正定値行列にするために，$\mathbf{D}' = diag(f(\lambda_1), ..., f(\lambda_N))$ を用いて，行列 $\mathbf{M}' = \mathbf{UD'U}^\top$ を作成する．ここで $f(x)$ は $x > 0$ のとき $f(x) = x + 1$，$x \leq 0$ のとき $f(x) = 0$ である．

Step 5: $\mathbf{M}'$ の正規化を行うために，行列 $\mathbf{M}'$ の (i, j) 成分を m'_{ij} として，行列 $\mathbf{G} = (g_{ij})$ を (i, j) 成分 g_{ij} が $g_{ij} = m'_{ij}/\sqrt{m'_{ii}m'_{jj}}$ となるように作成する．

Step 6: 行列 $\mathbf{G} = (g_{ij})$ をタンパク質間グラム行列として，Support Vector Machine (SVM) とペアワイズカーネルを用いて学習をする．このときに正例，負例数に比例した重みをつけて学習を行う．たとえば正例数が n_+, 負例数が n_- のとき，正例には $\frac{n_-}{n_+ + n_-}$, 負例には $\frac{n_+}{n_+ + n_-}$ の重み付けをする．

しかし Hue らの手法には 2 つの問題点が存在する．

問題点 1: タンパク質間類似度行列の作成に時間がかかる

データセットに含まれているタンパク質の数を N 個として，合計で ${}_N\mathrm{C}_2$ 回の構造アラインメントを必要とする．構造アラインメントは比較的時間のかかる計算であり，1 回の計算に約 0.05 秒必要とする．例えば 50,000 件のタンパク質に対して予測を行いたい場合，標準的な 1 CPU コアの逐次処理で約 2 年の時間を要する．

問題点 2: 新規タンパク質予測時に学習を都度やり直す必要がある

Step 3, 4, 5 の行列変換および正規化は SVM で学習を行う際に用いるグラム行列が半正定値性を満たすために必要な操作である．しかし，この手順は学習時のデータセットに

含まれていない新規タンパク質を予測するときに都度行う必要がある．このため，新規タンパク質予測時に新しいグラム行列を作成し，それに基づいてすべてのデータに対する学習も都度やり直す必要がある．

3. 提案手法

3.1 概要

提案手法の PPI 予測の流れについて**図 1** に示す．また，予測フローの詳細を以下に示す．

提案手法による PPI 予測

Step 1: データセットに含まれるタンパク質 $P_i(i = 1, ..., N)$ とタンパク質ライブラリに含まれるタンパク質 $P_j(j = 1, ..., K)$ に対して構造アラインメントを行ったときの E-value を E_{ij} として，$s_{ij} = \max(20, -\ln E_{ij})$ と計算する．

Step 2: s_{ij} を j 方向（タンパク質ライブラリごと）に Z–score 化を行い，$s'_{ij} = \frac{s_{ij}-\mu_j}{\sigma_j}$ とする．ただし μ_j, σ_j は s_{ij} の i に関する平均，標準偏差である．

Step 3: データセットに含まれるタンパク質 P_i の特徴ベクトル $\boldsymbol{v}_i$ を $\boldsymbol{v}_i = (s'_{i1}, ..., s'_{iK})^\top$ と定義する．

Step 4: (i, j) 成分 g_{ij} が $g_{ij} = \mathrm{RBF}(\boldsymbol{v}_i, \boldsymbol{v}_j)$ となる対称行列 $\mathbf{G} = (g_{ij})$ を求める．$\mathbf{G}$ は擬似的なタンパク質類似度行列に相当する．ただし $\mathrm{RBF}(\boldsymbol{v}_i, \boldsymbol{v}_j) = \exp(-\gamma\|\boldsymbol{v}_i - \boldsymbol{v}_j\|^2)$ である．$\mathbf{G}$ は半正定値性を満たす．

Step 5: 行列 $\mathbf{G}$ をタンパク質間グラム行列とし，SVM とペアワイズカーネルを用いて学習をする．このとき Hue らの手法の Step 6 と同様に正例，負例数に比例した重みをつけて学習を行う．

提案手法と Hue らの手法に必要な構造アラインメント回数比較を**表 1** に示す．Hue らの手法は ${}_N\mathrm{C}_2$ 回の構造アラインメントを必要としていたため，たとえば $N = 10{,}000$ かつ $K = 500$ のとき学習時に約 10 倍の構造アラインメント回数の削減が可能である．さらに M 個の新規タンパク質予測時には，たとえば $N = 10{,}000$ かつ $K = 500$ かつ $M = 100$ とすると，約 10 倍の構造アラインメント回数の削減が可能である．これにより**問題点 1** が解決される．さ

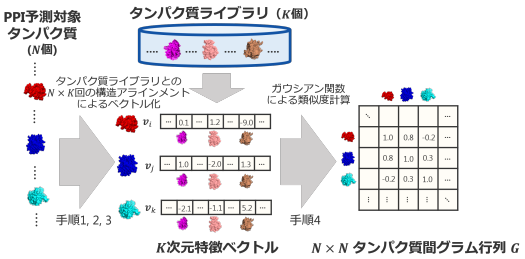

図 1 提案手法におけるタンパク質グラム行列の作成
Fig. 1 The way to make Gram matrix

表 1 既存手法と提案手法のアラインメント回数の比較（K はライブラリに含まれる構造数）
Table 1 Comparison of Hue *et al.* and proposed method

	既存手法	提案手法
N 個のタンパク学習時	$N(N-1)/2$	NK
M 個の新規タンパク質予測時	$M(N+M)/2$	MK

らに，タンパク質間グラム行列作成のときに RBF カーネルを用いることで，半正定値行列を作成することが可能であり，学習データセット中に含まれない新規タンパク質予測時にグラム行列の再計算が不要である．そのため，提案手法では学習済みの学習器を用いて PPI 予測が可能である．これにより**問題点 2** が解決される．

3.2 タンパク質ライブラリの構築方法

Structual Classification of Proteins (SCOP) [21] は 1995 年に Murzin らによって開発された，タンパク質立体構造の分類方法およびデータベースである．本研究では SCOP を用いてタンパク質ライブラリの構築を行う．はじめに，SCOP に記載されているタンパク質のうち，25 残基以下または 1,000 残基以上の構造を除去した．結果として，924 個のフォールドとそれに属する 43,071 個のタンパク質立体構造を得た．これを集合 $\mathcal{SCOP}$ とあらわす．$\mathcal{SCOP}$ から以下の 2 通りの方法を用いてタンパク質ライブラリを構築した．

タンパク質ごとにランダム抽出 (Protein-wise)
$\mathcal{SCOP}$ に含まれるタンパク質立体構造から重複を許さずにランダムに抽出することでタンパク質ライブラリを構築する．

フォールドごとにランダム抽出 (Fold-wise)
$\mathcal{SCOP}$ に含まれるタンパク質立体構造が属するフォールドから，重複を許さずに 900 個ランダムに抽出する．得られたフォールドに属する立体構造を 1 つランダムに抽出することでタンパク質ライブラリを構築する．

それぞれの抽出方法に対してランダムに 900 個の 3 つのライブラリを作成した．さらに各 900 個のライブラリに対して，$\{10, 50, 100, 150, ..., 800, 850\}$ 個のそれぞれ包含関係をもつサブセットからなるライブラリを構築した．よって構築されたライブラリの合計は $3 \times 2 \times 19 = 114$ 個である．各ライブラリを “α_β_γ” と呼び，α, β, γ は以下の対応に従う．

$$\alpha = \begin{cases} \mathrm{P}\ (\text{Protein-wise のとき}) \\ \mathrm{F}\ (\text{Fold-wise のとき}) \end{cases}$$

$$\beta \in \{1, 2, 3\}\ (\text{ランダムに関してのライブラリ番号})$$

$$\gamma \in \{10, 50, 100, 150, ..., 800, 850, 900\}$$

例えば，P_1_100 は Protein-wise にランダム抽出された 1 個目のライブラリのうち，ライブラリに含まれる構造の数

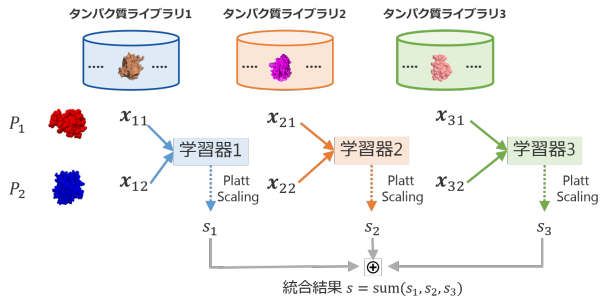

図 2　予測結果の統合プロトコル
Fig. 2　Integration of prediction results

が 100 個のものを指す．また，ランダム抽出には包含関係があり，例えば F_2_200 は F_2_250 のサブセットである．

3.3　複数のライブラリから得られた予測結果の統合

複数のライブラリから構築された学習器は異なる構造的な特徴を学習している可能性が考えられる．そのため，本研究では 3 つの異なるライブラリから構築された学習器の予測結果を統合して，1 つの予測結果とすることを提案する．しかし，一般に SVM から得られるスコアはサポートベクトルの数に依存するため，Platt Scaling [22] を用いてスコアを [0,1] の範囲に正規化する．Platt Scaling では入力特徴量 $\boldsymbol{x}$ について，SVM によって計算された分離超平面からの距離 $d_{\boldsymbol{x}}$ を用いて式 (1) に従って $[0, 1]$ 区間のスコア $s(\boldsymbol{x})$ を計算する．

$$s(\boldsymbol{x}) = \frac{1}{1 + \exp(Ad_{\boldsymbol{x}} + B)} \tag{1}$$

なお，式 (1) において A, B はハイパーパラメータである．**図 2** のように各学習器に関して Platt Scaling を用いて計算されたスコアの和を最終的なスコアとして計算する．

4.　実験

4.1　データセット

使用するデータセットについては，Hue らの研究 [16] で用いられているものをもとに作成した Hue データと，Maheshwari らの研究 [11] で用いられているものをもとに作成した Maheshwari データの 2 つを用いた．データセットの概要を**表 2** に示す．

Hue データ

Hue らの研究 [16] で用いられたデータセットである．PPI データベースの DIP [23] と PSI-BLAST [24] を使用して作成されており，本研究では残基数が 25 より少ない，または 1,000 より多い構造を除去し，さらに負例として正例ペアをランダムに交換したものを追加して用いた．

Maheshwari データ

Maheshwari らの研究 [11] で用いられたデータセットである．PDB の複合体構造をもとに作成されており，負例はホモダイマーのランダムな交換によって作成されている．本研究では，Hue データセットに含まれている構造ペアを PDB ID ベースで除去し，残基数が 25 より少ない，または 1,000 より多い構造を除去した．さらに，正例，負例をそれぞれ 1,000 個ずつランダムにサンプリングして用いた．

4.2　構造アラインメントに要する時間計測実験

Hue データに含まれるタンパク質に対して，すべてのペアについて構造アラインメントを行ったときの時間（Hue らの手法における時間）とタンパク質ライブラリに対しての構造アラインメントを行ったときの時間（提案手法における時間）を計測した．使用するタンパク質ライブラリはライブラリ中で最も AUC が高かったものを用いた．構造アラインメントツールとして，Mammoth [25] を使用した．時間計測は Linux の `time` コマンドを用いて行い，I/O 待ちを考慮した Real time と，I/O 待ちを考慮していない User time と System time の和の 2 種類を求めた．利用した計算機環境を**表 3** に示す．

4.3　予測精度評価実験

Hue らの手法と各タンパク質ライブラリを用いた提案手法に対して，Hue データに対する Cross Validation (CV) と Maheshwari データに対するテストの 2 つの方法で予測精度評価実験を行った．以下に 2 つの詳細を示す．

Hue データに対する CV

Hue データに対して 5-fold CV を行った．ペアワイズカーネルには TPPK，MLPK の 2 種類を使用した．SVM のハイパーパラメータは $C = \{2^{-15}, 2^{-14}, ..., 2^{4}\}$，RBF カーネルのハイパーパラメータは $\gamma = \{2^{-5}, 2^{-4}, ..., 2^{10}\}$ の範囲で探索を行った．なお，Hue らの手法は RBF カーネルを使用しないため，γ の探索は行わない．また，Platt Scaling におけるパラメータ（式 (1) における A, B）は各 fold において学習データをさらに 5-fold CV することで決定した．Platt Scaling のパラメータの決定は SVM の学習とは独立に行われた．

表 2　データセット概要
Table 2　Dataset

データセット名	正例数	負例数	タンパク質構造数
Hue データ	1,356	8,318	6,148
Maheshwari データ	1,000	1,000	3,880

表 3　使用する計算機環境
Table 3　Computing environment

OS	SUSE Linux Enterprise Server 12 SP2
CPU	Intel Xeon E5-2680 V4 Processor (Broadwell-EP, 14 core, 2.4 GHz) × 2
RAM	256 GiB
ストレージ	Intel DC P3500 2 TB (SSD)

表 4　計算時間 [s]．括弧の中は Hue らの手法に対する高速化率）
Table 4　Calculation time (Speed-up rate)

手法	実時間	ユーザー+システム 時間
Hue ら	882,400	143,400
P_1_550	133,800 (×6.6)	129,400 (×6.6)
P_2_350	86,300 (×10.3)	83,900 (×10.1)
P_3_600	164,100 (×5.4)	143,400 (×5.9)

Maheshwari データに対するテスト

Hue データをすべてを訓練データとして用い，Maheshwari データをテストとして評価を行った．ハイパーパラメータは Hue データに対する CV で最も精度が良かった値を使用した．

精度評価指標は Receiver Operating Characteristic Curve-Area Under Curve (ROC-AUC) を用いる（単に AUC という場合，ROC-AUC を指すものとする）．

5. 結果

5.1 時間計測実験結果

予測精度評価実験において特に精度が高い P_1_550, P_2_350, P_3_600 の 3 つのライブラリに対して構造アラインメントに必要な時間を計測した．結果を**表 4** に示す．

データセットに含まれるタンパク質の数を N としてタンパク質ライブラリに含まれるタンパク質の数を K とすると，Hue らの手法では $N(N-1)/2$ 回，提案手法では NK 回である．これに基づけば，$N = 6{,}148$ かつ $K = 550, 350, 600$ のときに計算の高速化率の見積もりはそれぞれ $5.59, 8.78, 5.12$ である．実際の高速化率は実時間でそれぞれ $6.6, 10.3, 5.4$ となっており，見積もりと同等またはそれ以上の高速化を達成した．これは構造アラインメントの時間がタンパク質の大きさに依存し，タンパク質ライブラリに含まれる構造がデータセット中のタンパク質の構造より比較的小さいためであると考えられる．また，実時間と（ユーザー+システム時間）で計算時間の差異が十分に小さいことから，I/O 待ちの影響が計算時間に対して十分に小さいことがわかる．

5.2 各ライブラリ間の類似度

各ライブラリにおける Jaccard 係数を**表 5**，**表 6** に示す．各ライブラリ間の Jaccard 係数は十分に小さいため，予測精度を議論する際にライブラリ間の依存関係は考慮しなくてもかまわない．さらに，Protein-wise な抽出のほうが Fold-wise な抽出にくらべて Jaccard 係数が低くなるのはタンパク質ライブラリの構築時にフォールドの方がタンパク質に比べて数が少ないためであると考えられる．

5.3 Hue データに対する CV

Hue データに対する CV についての実験結果を**図 3** と**図 4** に示す．また，TPPK, MLPK それぞれで AUC が最大となったタンパク質ライブラリを**表 7**，**表 8** に示す．

表 5　Protein-wise ライブラリ間の Jaccard 係数
Table 5　Jaccard coefficient with Protein-wise libraries

	P_1_900	P_2_900	P_3_900
P_1_900	1.0	0.00756	0.00376
P_2_900	0.00756	1.0	0.00376
P_3_900	0.00376	0.00376	1.0

表 6　Fold-wise ライブラリ間の Jaccard 係数
Table 6　Jaccard coefficient with Fold-wise libraries

	F_1_900	F_2_900	F_3_900
F_1_900	1.0	0.170	0.174
F_2_900	0.170	1.0	0.163
F_3_900	0.174	0.163	1.0

表 7　Hue データに対する CV の結果 (TPPK)
Table 7　CV results of Hue data (TPPK)

手法	AUC （標準偏差）	γ	C
Hue らの手法	0.601 (0.00600)	-	2^{-1}
P_1_850	**0.733** (0.0165)	2^{-11}	2^{0}
P_2_700	0.730 (0.0143)	2^{-10}	2^{-1}
P_3_550	0.726 (0.0118)	2^{-10}	2^{-1}
F_1_600	0.714 (0.0105)	2^{-10}	2^{-1}
F_2_250	0.715 (0.00861)	2^{-10}	2^{1}
F_3_550	0.716 (0.0160)	2^{-10}	2^{-2}

表 8　Hue データに対する CV の結果 (MLPK)
Table 8　CV results of Hue data (MLPK)

手法	AUC （標準偏差）	γ	C
Hue らの手法	0.700 (0.0186)	-	2^{-3}
P_1_550	0.735 (0.0135)	2^{-12}	2^{2}
P_2_350	**0.741** (0.0147)	2^{-10}	2^{-1}
P_3_600	0.736 (0.0176)	2^{-15}	2^{6}
F_1_900	0.704 (0.0174)	2^{-15}	2^{6}
F_2_450	0.703 (0.0201)	2^{-10}	2^{-1}
F_3_500	0.700 (0.0138)	2^{-10}	2^{-1}

表 7 と表 8 から最も予測精度が良かったライブラリとペアワイズカーネルの組み合わせは P_2_350, MLPK であった．図 3 と図 4 をみると，タンパク質ライブラリに含まれるタンパク質の数は 200 個以上ではほぼ精度が一定となっていることがわかった．さらに構築方法が同じライブラリ間では予測精度に差が見られないことがわかった．

また，Hue データに対する CV では Hue らの手法とほぼ同等の精度を達成した．

5.4 Maheshwari データに対するテスト

CV によって選ばれた各タンパク質ライブラリについて，Maheshwari データに対するテストを行った結果を**表 9** と**表 10** に示す．Hue データに対する CV の結果と同様に，

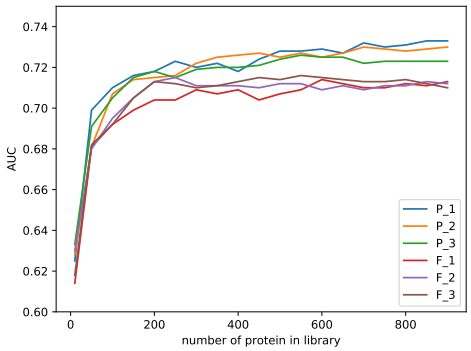

図 3　各ライブラリの CV 結果 (TPPK)
Fig. 3　CV results for each libraries (TTPK)

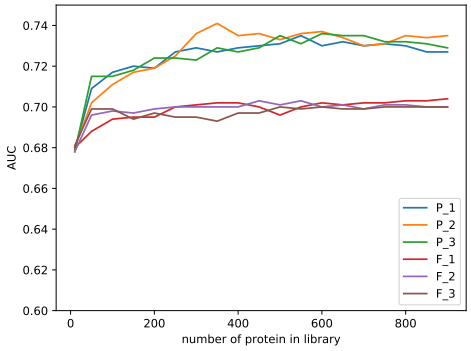

図 4　各ライブラリの CV 結果 (MLPK)
Fig. 4　CV results for each libraries (MLPK)

表 9　TPPK のテスト結果　**表 10**　MLPK のテスト結果
Table 9 Test results (TPPK)**Table 10** Test results (MLPK)

手法	AUC
Hue らの手法	**0.752**
P_1_850	0.744
P_2_700	0.750
P_3_550	0.751
F_1_600	0.745
F_2_250	0.720
F_3_550	0.721

手法	AUC
Hue らの手法	**0.791**
P_1_550	0.766
P_2_350	0.749
P_3_600	0.750
F_1_900	0.744
F_2_450	0.752
F_3_500	0.760

タンパク質ライブラリの構築方法としては Protein-wise の精度が高かった．また，Maheshwari データに対するテストでは提案手法は Hue らの手法に精度がやや劣っていた．これらの Hue データと Maheshwari データでの傾向の差は負例の作成方法の違いが一因である可能性がある．

MLPK における Hue らの手法と P_1_550, P_2_350, P_3_600 の ROC 曲線を図 **5** に示す．提案手法は Hue らの手法に比べて ROC 曲線の立ち上がりは早いものの，AUC の値では劣っている結果になった．False Positive Rate が 0.1 までの AUC を算出したところ，Hue らの手法と P_1_550, P_2_350, P_3_600 の精度はそれぞれ 0.0276, 0.0361, 0.0308, 0.0277 であった．

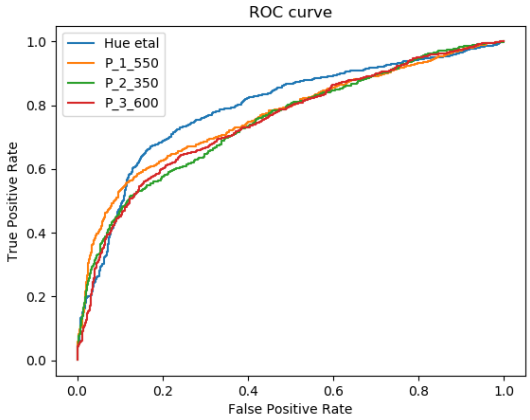

図 5　各手法における ROC 曲線
Fig. 5　ROC curves for each method

5.5　複数ライブラリから得られた予測結果の融合

Hue データに対する CV で高い予測精度を達成した P_1_550, P_2_350, P_3_600 と MLPK を用いて予測結果を融合した．

Hue データに対する CV では，P_1_550, P_2_350, P_3_600 単体の精度が 0.735, 0.741, 0.736 であったが，統合の結果 0.806 (0.00983) となり，精度向上が見られた．Maheshwari データに対するテストでは，P_1_550, P_2_350, P_3_600 単体の精度が 0.766, 0.749, 0.750 であったが，統合の結果は 0.770 であったため，精度向上は見られなかった．これは Hue データと Maheshwari データでの傾向の差は負例の作成方法による違いが一因となっている可能性がある．

6.　考察

6.1　各ライブラリから作成されたタンパク質間グラム行列

本節ではライブラリから作成されたタンパク質間グラム行列の性質について考察する．対象とするタンパク質間グラム行列は P_2_350 と P_3_600 から作成された行列とする．RBF カーネルのハイパーパラメータは Hue データに対する CV で決定されたものを用いる．

たとえば 1MNM_A と 1QWO_A の疑似類似度は表 **11** のようになっている．1MNM_A の構造と 1QWO_A の構造を図 **6** と図 **7** に示す．1MNM_A と 1QWO_A を構造アラインメントした場合の $-\ln E$ は 1.923 である（E は E-value）．これは P_2_350 における $-\ln E$ の平均が 1.631，標準偏差 1.320 であることを考えると，中間的な類似度であることがわかる．このように中間的な類似度をもつタンパク質ペアは各々のタンパク質ライブラリごとに異なる部分に注目した類似度が計算されていることがわかる．

6.2　ライブラリの分割に関する考察

5.5 節で複数のライブラリから構築された学習器の予測結果を融合することで精度が向上する可能性について言及

表 11　1MNM_A と 1QWO_A の提案手法で計算された疑似類似度
Table 11　Psuedo structual simirarity between 1MNM_A and 1QWO_A

タンパク質ライブラリ	疑似類似度
P_1_550	0.833
P_2_350	0.0792
P_3_600	0.980

図 6　1MNM_A の構造
Fig. 6　1MNM_A structure

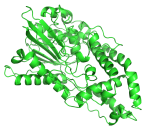

図 7　1QWO_A の構造
Fig. 7　1QWO_A structure

表 12　ライブラリの分割実験結果
Table 12　Library split effect

ライブラリ	CV	テスト
P_3_600	0.736 (0.0176)	0.750
P'_3_200$_1$	0.721 (0.0119)	0.751
P'_3_200$_2$	0.720 (0.0124)	0.725
P'_3_200$_3$	0.717 (0.0170)	0.741
3 つの結果の統合	**0.797** (0.0120)	**0.765**

した．そこで，単一のライブラリに含まれる構造数を増やすのではなく，ライブラリ数を増やすことで精度が向上することが考えられる．本節では P_3_600 のライブラリに対して，200 個からなるサブライブラリ 3 つを構築して 3 つの予測結果を融合した場合の精度の変化について考察する．

P_3_600 を 3 つにランダムに分割したライブラリを作成する．これらのライブラリの名前を P'_3_200$_1$, P'_3_200$_2$, P'_3_200$_3$ として予測精度評価実験を行った結果を**表 12** に示す．ペアワイズカーネルには比較的結果のよかった MLPK を用いた．この結果，Hue データに対する CV, Maheshwari データに対するテストの両方で精度の向上がみられた．これらの結果は，ライブラリに多くのタンパク質を用意して 1 つの学習器を作成するよりも，いくつかのライブラリに分割してそれらの予測結果を融合する方が精度が向上する可能性を示唆している．

6.3　その他の 3 次元記述子との比較

タンパク質の立体構造分類の分野において Canterakis によるゼルニケ記述子 [26] と呼ばれる 3 次元記述子が存在する．ゼルニケ記述子は 121 次元のベクトルで構成され，すべての原子を計算対象とする場合と主鎖のみを計算対象とする場合がある．本研究では主鎖のみから計算されたゼルニケ記述子を用いて Hue データに対する CV と Maheshwari データに対するテストを行った．ハイパーパラメータなどの条件は 4.3 節と同じものを使用した．なお 1Q55_A のみはベクトル変換ができなかったため，データセットから除外した．

PPI 予測精度を求めた結果，MLPK を用いたときに TPPK より精度が良く，CV で AUC=0.704，テストで AUC=0.678 であった．これは Hue らの手法や提案手法よりも予測精度低く，ゼルニケ記述子が相互作用部位と関係ない部分を含めたタンパク質全体の構造の類似度を計算しているためと考えられる．本研究の提案手法ではライブラリの複数の構造群から局所構造を学習できている可能性があり，ゼルニケ記述子等を PPI 予測に適用する場合は例えばタンパク質の局所構造に対して計算したゼルニケ記述子を用いる等の工夫が必要である．

7.　結論

7.1　本研究の結論

本研究では，あらかじめ用意した 100 から 1,000 個程度のタンパク質ライブラリのみとの構造アラインメントから得られるベクトル表現によって擬似的な構造類似度を計算することにより高速に PPI 予測を行う手法を提案した．これにより，従来手法に比べて以下の 2 つの改良に成功した．

構造アラインメントの回数削減
従来はデータセットに含まれているタンパク質の数を N 個として，合計で ${}_N\mathrm{C}_2$ 回の構造アラインメントを必要としていたが，提案手法ではタンパク質ライブラリの数を K 個とすると，構造アラインメントの回数が NK 回に削減された．

タンパク質間グラム行列の変換が不要
提案手法ではタンパク質グラム行列作成のときに半正定値行列への変換が不要となった．学習済みの学習器を用いて，訓練データセットに含まれないタンパク質に対する PPI 予測が可能となった．

先行研究で用いられた 2 つのデータセットを元にした計算機実験によって $K \sim 350$ 程度で予測精度が従来手法と同等程度であることが示され，$N \sim 6{,}000$ のデータセットに対して，約 10 倍の高速化を達成した．タンパク質ライブラリに含まれる構造数は 200 ∼ 600 程度で精度が一定となり，ライブラリの構築方法としては SCOP からランダムにタンパク質を抽出する方法（Protein-wise）が精度が最も高くなった．さらに各々のタンパク質ライブラリが異なる立体的な特徴を捉えていることを確認し，複数のタンパク質ライブラリから学習した判別器を統合することで予測精度が向上する可能性について言及した．

7.2　今後の課題

Hue データの CV と Maheshwari データへのテストにおいて，Hue らの手法と提案手法の予測精度に違いがあった．今回は 2 つのデータセットを扱ったが，他の複数のデータセットに対して予測精度を評価することが必要がある．ま

た 6.3 節において，ライブラリの分割によって予測精度が向上する可能性について言及したが．ライブラリを分割する際の適切な分割方法について検討する必要がある．

謝辞 本研究の一部は，JSPS 科研費 (17H01814, 18K18149)，JST CREST「EBD: 次世代の年 ヨッタバイト処理に向けたエクストリームビッグデータの基盤技術」(JPMJCR1303)，JST リサーチコンプレックス推進プログラム，文部科学省 地域イノベーション・エコシステム形成プログラムの支援を受けて行われた．

参考文献

[1] U. Stelzl, *et al.*, "A human protein-protein interaction network: A resource for annotating the proteome", *Cell*, 122:957–968, 2005.

[2] T. Oltersdorf, *et al.*, "An inhibitor of Bcl–2 family proteins induces regression of solid tumours", *Nature*, 435:677–681, 2005.

[3] G. M. Popowicz, *et al.*, "Structures of low molecular weight inhibitors bound to MDMX and MDM2 reveal new approaches for p53-MDMX/MDM2 antagonist drug discovery", *Cell Cycle*, 9:1104–1111, 2010.

[4] M. Ohue, *et al.*, "MEGADOCK: An All-to-All Protein-Protein Interaction Prediction System Using Tertiary Structure Data", *Protein Pept Lett*, 21:766–778, 2014.

[5] Y. Matsuzaki, *et al.*, "Protein-protein interaction network prediction by using rigid-body docking tools: application to bacterial chemotaxis", *Protein Pept. Lett*, 21:790–798, 2014.

[6] Z. H. You, *et al.*, "An improved sequence-based prediction protocol for protein-protein interactions using amino acids substitution matrix and rotation forest ensemble classifiers", *Neurocomputing*, 228:277–282, 2017.

[7] C. Leslie, *et al.*, "The spectrum kernel: a string kernel for SVM protein classification", *Pac Symp Biocomput*, 575:564–575, 2002.

[8] J. Shen, *et al.*, "Predicting protein-protein interactions based only on sequences information", *Proc Natl Acad Sci USA*, 104:4337–4341, 2007.

[9] N. Zaki, *et al.*, "Protein-protein interaction based on pairwise similarity", *BMC Bioinform*, 10:150, 2009.

[10] S. Bandyopadhyay, *et al.*, "A New Feature Vector Based on Gene Ontology Terms for Protein-Protein Interaction Prediction", *IEEE/ACM Trans Comput Biol Bioinform*, 14:762–770, 2017.

[11] S. Maheshwari, *et al.*, "Across-proteome modeling of dimer structures for the bottom-up assembly of protein-protein interaction networks", *BMC Bioinform*, 18:257, 2017.

[12] J. Qiu, *et al.*, "A structural alignment kernel for protein structures", *Bioinformatics*, 23:1090–1098, 2007.

[13] F. Halakou, *et al.*, "Enriching Traditional Protein-protein Interaction Networks with Alternative Conformations of Proteins", *Sci Rep*, 7:7180, 2017.

[14] P. Aloy, *et al.*, "The relationship between sequence and interaction divergence in proteins", *J Mol Biol*, 332:989–998, 2003.

[15] A. Szilagyi, *et al.*, "Template-based structure modeling of protein-protein interactions", *Curr Opin Struct Biol*, 24:10–23, 2014.

[16] M. Hue, *et al.*, "Large-scale prediction of protein-protein interactions from structures", *BMC Bioinform*, 11:144, 2010.

[17] A. R. Ortiz, *et al.*, "MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison", *Protein Sci*, 11:2606–2621, 2009.

[18] A. Ben-Hur, *et al.*, "Kernel methods for predicting protein-protein interactions", *Bioinformatics*, 21:i38–i46, 2005.

[19] J.-P. Vert, *et al.*, "A new pairwise kernel for biological network inference with support vector machines", *BMC Bioinform*, 8:S8, 2007.

[20] J. L. Faulon, *et al.*, "Genome scale enzyme – Metabolite and drug - Target interaction predictions using the signature molecular descriptor", *Bioinformatics*, 24:225–233, 2008.

[21] A. G. Murzin, *et al.*, "SCOP: A structural classification of proteins database for the investigation of sequences and structures"", *J. Mol. Biol.*, 247:536540, 1995.

[22] H.-T. Lin, *et al.*, "A note on Platt's probabilistic outputs for support vector machines", *Mach Learn*, 68:267–276, 2007.

[23] I. Xenarios, *et al.*, "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions", *Nucleic Acids Res*, 30:303–305, 2002.

[24] S. Altschul, *et al.*, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res*, 25:3389–3402, 1997.

[25] A. R. Ortiz, *et al.*, "MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison", Protein Sci., vol. 11, pp. 2606-2621, 2009.

[26] N. Canterakis, "3D Zernike Moments and Zernike Affine Invariants for 3D Image Analysis and Recognition", *11th Scand Conf Image Anal*, 1999.