Vol.2016-BIO-45 No.9
2016/3/19

タンパク質-化合物相互作用ネットワークのリンクマイニング

山崎 卓朗[1,a)]　大上 雅史[2,b)]　秋山 泰[2,3,c)]

概要：新薬開発にかかる莫大な経済的，時間的コストを削減するために，新薬候補化合物をコンピュータ上で選別する，Virtual Screening という手法が広く用いられている．Virtual Screening の一種として，既知タンパク質-化合物相互作用情報から新規相互作用の予測を行う Chemical Genomics-Based Virtual Screening (CGBVS) があり，これには様々な機械学習やデータマイニングの手法が用いられている．CGBVS は高効率かつ高精度な予測が出来る一方で，相互作用情報が未知の新規化合物に対する予測精度が不十分であるという欠点がある．本研究では，リンクマイニングの観点から，CGBVS の既存法の中で新規化合物に対する予測精度が最も高い Pairwise Kernel Method (PKM) に対しリンク指標を用いた改良方法を提案した．また，AUROC および AUPR を用いて提案法の予測精度の評価を行った．その結果，提案法は従来法に比べて AUPR 値で 0.468 から 0.562 への精度向上を達成した．

キーワード：バーチャルスクリーニング，ペアワイズカーネル，リンクマイニング

Link Mining in Protein–Compound Interaction Network

Takuro Yamazaki[1,a)]　Masahito Ohue[2,b)]　Yutaka Akiyama[2,3,c)]

Abstract: Virtual screening (VS) is widely used in the process of a computational drug discovery for reducing a large amount of cost. Chemical genomics-based virtual screening (CGBVS) which is a kind of VS predicts new protein-compound interactions (PCIs) from known PCIs data using several methods of machine learning or data mining. Although CGBVS provides highly efficient and accurate PCIs prediction, CGBVS has poor performance on prediction for new compound for which PCIs are unknown. Pairwise kernel method (PKM) is one of the state-of-the-art methods of CGBVS, that showed highest accuracy on prediction for new compounds. In this study, from the viewpoint of link mining we improved PKM by combining link indicator and chemical similarity, and evaluated their accuracy. The proposed method obtained AUPR value of 0.562 which is higher than that achieved by using normal PKM (0.468).

Keywords: Virtual Screening, Pairwise Kernel，Link Mining

[1] 東京工業大学 工学部 情報工学科
Department of Computer Science, Faculty of Engineering, Tokyo Institute of Technology
[2] 東京工業大学 大学院情報理工学研究科 計算工学専攻
Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology
[3] 東京工業大学 情報生命博士教育院
Education Academy of Computational Life Sciences, Tokyo Institute of Technology
[a)] yamazaki@bi.cs.titech.ac.jp
[b)] ohue@bi.cs.titech.ac.jp
[c)] akiyama@cs.titech.ac.jp

1. 導入

計算機を用いて薬剤候補化合物を絞り込む Virtual Screening には，特定のタンパク質に対する既知の活性情報を用いる Ligand-Based Virtual Screening，構造情報を用いる Structure-Based Virtual Screening，複数のタンパク質と化合物の既知相互作用情報から予測を行う Chemical Genomics-Based Virtual Screening (CGBVS) という複数のアプローチがある．中でも CGBVS は高精度な予

図 1 未知タンパク質，未知化合物，未知相互作用の予測の概要図．クロスバリデーションでは，それぞれ protein-CV, compound-CV, interaction-CV に対応する．

測が可能であると言われており [1]，サポートベクタマシン (Support Vector Machine, SVM) [2] や制約付きボルツマンマシン [3]，k-最近傍法 [4]，決定木 [5] などの様々な機械学習やデータマイニングの手法が用いられている [6]．Ding らが 2014 年に行った大規模な CGBVS の既存法の比較では，Pairwise Kernel Method (PKM) [7] が最も予測精度が良い方法として示されている [1]．PKM は Jacob らが 2008 年に提案した手法であり [7]，Pairwise Kernel という組で構成されるサンプルに適したカーネルと SVM を組み合わせることで高い予測精度を達成したものである．しかしながら，既にある程度ペアの相互作用情報が得られている場合は高い精度での予測が可能である一方で，相互作用情報が既知の相手が無いケース，特に未知化合物に対する予測精度が不十分であるという問題があった．本研究では，タンパク質-化合物相互作用予測を行う CGBVS を，2 部ネットワークのリンクマイニングによる観点で見たときに，既存法である PKM がネットワークのノード情報しか用いていないことに着目した．リンクマイニング [8] とは，ウェブのハイパーリンクや SNS の知り合い関係などに代表される，ネットワーク構造を持つデータを対象とするデータマイニングのことである．リンクマイニングに用いられる情報はノード情報と構造情報に大別でき，SNS の例ではノード情報はユーザの年齢や住所・趣味などの個人情報，構造情報は誰と誰が友人関係にあるかという情報にあたり，リンク指標とも呼ばれる．PKM では予測の際にタンパク質や化合物の化学的な類似度のみを用いていることから，ネットワークのノード情報のみを用いて予測を行っていると言える．これは SNS の例では，友人の友人は友人である可能性が高いといった，ネットワーク構造から読み取ることの出来る情報を用いずに，共通の趣味や居住地といった個人情報のみから友人関係の予測を行っていることに該当する．そのため，PKM にタンパク質-化合物相互作用ネットワークの構造情報をうまく組み合わせることで，予測精度の向上が期待できる．

そこで本研究では，PKM にタンパク質-化合物相互作用ネットワークから算出されるリンク指標を統合し，予測精度の改良を試みた．

2. 既存法：Pairwise Kernel Method

Pairwise Kernel Method (PKM) は，Jacob らが 2008 年に提案したサポートベクタマシン (Support Vector Machine, SVM) [2] を用いた予測手法である [7]．SVM は主に 2 クラス分類に用いられる教師あり学習アルゴリズムの 1 つであり，カーネル関数を用いて入力値を有限もしくは無限次元の特徴空間へ写像することにより効率的に非線形の決定境界を構成する．

PKM ではカーネル関数の構成について独自のカーネル (Pairwise Kernel) を用いており，タンパク質-化合物間相互作用予測問題に限らず様々な問題に応用されている [9] [10]．Pairwise Kernel では，タンパク質 t と化合物 d のペアの特徴ベクトル $\Phi(t,d)$ を，化合物 d の特徴ベクトル $\Phi_c(d)$ とタンパク質 t の特徴ベクトル $\Phi_p(t)$ のテンソル積を用いるというアイデアに基づく．

$$\Phi(t,d) = \Phi_p(t) \otimes \Phi_c(d)$$

ここでタンパク質のカーネル関数 $K_p(t,t')$ と化合物のカーネル関数 $K_c(d,d')$ は以下のように表される．

$$\begin{aligned} K_p(t,t') &= \Phi_p(t)^\top \Phi_p(t') \\ K_c(d,d') &= \Phi_c(d)^\top \Phi_c(d') \end{aligned}$$

これらを用いて，タンパク質-化合物ペアのカーネル関数 $K_{pair}((t,d),(t',d'))$ が以下のように導出される．

$$\begin{aligned} K_{pair}((t,d),(t',d')) &= \Phi(t,d)^\top \Phi(t',d') \\ &= (\Phi_p(t) \otimes \Phi_c(d))^\top (\Phi_p(t') \otimes \Phi_c(d')) \\ &= \Phi_p(t)^\top \Phi_p(t') \times \Phi_c(d)^\top \Phi_c(d') \\ &= K_p(t,t') \times K_c(d,d') \end{aligned} \quad (1)$$

このとき，$K_p(t,t')$ と $K_c(d,d')$ が正定値カーネルならば $K_{pair}((t,d),(t',d'))$ も正定値カーネルとなる．

カーネル関数は類似度を表しているとみなすことができる．$K_p(t,t')$ はタンパク質 t と t' の類似度であり，$K_c(d,d')$ は化合物 d と d' の類似度を表す．すなわち，タンパク質-化合物ペアのカーネル関数は，タンパク質間類似度と化合物間類似度との積によって求められるということが，Pairwise Kernel のアイデアの重要な点である．

PKM では式 (1) に基づいて算出された Pairwise Kernel を用いて SVM の学習を行い，入力として与えられた化合物-タンパク質ペアに対して活性の有無の予測を行う．Pairwise Kernel の計算は全てのタンパク質-化合物ペアに対して行われるため，カーネル行列のサイズは非常に大きくなる．例えばデータセット中に化合物が 600 個，タンパク質が 500 個含まれる場合はカーネル行列は 600×500 次の正方行列となる．そのため実用上はメインメモリのサイズの制約から，全てのペアを学習に用いることはない．実

際には負例は正例よりも非常に多い（相互作用が疎である）ため，正例と同じ数だけ負例をランダムサンプリングして用いている [1].

3. 提案法：リンク指標の統合

既存の手法では予測の際にタンパク質や化合物の化学的な類似度のみを用いていた．そのため，このノード情報にネットワークの構造情報である，リンク指標を組み合わせることで予測精度の向上が期待できる．本研究では予測に適したリンク指標の探索を行うとともに，化学的な類似度との統合方法を検証した．

3.1 リンク指標

ネットワークの構造情報を数値化するために，リンク指標として以下の 3 つの指標を検討対象とした．

$$\text{Jaccard Index}(u,v) = \frac{|\Gamma(u)\cap\Gamma(v)|}{|\Gamma(u)\cup\Gamma(v)|}$$

$$\text{Cosine Similarity}(u,v) = \frac{|\Gamma(u)\cap\Gamma(v)|}{\sqrt{|\Gamma(u)||\Gamma(v)|}}$$

$$\text{LHN-1}(u,v) = \frac{|\Gamma(u)\cap\Gamma(v)|}{|\Gamma(u)||\Gamma(v)|}$$

ここで u, v はネットワークのノードを表し，$\Gamma(u)$ はノード u に連結したノードの集合を表す．タンパク質-化合物間相互作用予測では，u や v にタンパク質 t_1, t_2 をあてはめたとき，$\Gamma(t_1)$ と $\Gamma(t_2)$ はそれぞれタンパク質 t_1, t_2 と相互作用が知られている化合物群となる．つまり，相互作用相手となっている化合物の種類に基づいて，タンパク質同士の類似度を決定するという指標となる．

このリンク指標の値を，ネットワーク構造に基づくタンパク質間類似度，および化合物間類似度とした．本研究ではリンク指標の算出に，ネットワーク解析用の Python ライブラリである networkx [11] を用いた．

リンクマイニングに用いられるリンク指標は上述した 3 つの他にも，Adamic Adar Index や Graph Distance などさまざまある．しかし，それらの多くは Pairwise Kernel として用いるために必要な正定値性を満たさず，PKM と組み合わせることに適さない．そのため，本実験では正定値性が満たされることが確認された Jaccard Index，Cosine Similarity，LHN-1 の 3 つを用いることとした．

3.2 PKM との統合方法

上述したリンク指標を用いてネットワーク構造に基づくタンパク質間，化合物間類似度を求め，化学的類似度と統合して新たな類似度を定義した．リンク指標と化学的類似度とで値域や分布が異なる場合，同時に扱うときには標準化を行う必要があると考え，リンク指標と化学的類似度の統合方法は重み付き和と z-score で標準化した値の重み付き和の 2 通りを検証した．ここで $L_{i,j}$ はタンパク質あるいは化合物 i, j のリンク指標の値，$C_{i,j}$ は別途与えられる化学的類似度の値 [*1] を表す．w は重みパラメータである．

(1) 重み付き和

$$S_{i,j} = C_{i,j} + w \cdot L_{i,j}$$

(2) 標準化重み付き和

$$S_{i,j} = z\text{-}score(C_{i,j}) + w \cdot z\text{-}score(L_{i,j})$$

本研究では重み w として $\{0.1, 0.3, 0.5, 1\}$ の 4 種類に対して実験を行った．

4. 評価実験

4.1 PKM の実装

本研究では PKM および提案法の実装に際して機械学習用の python ライブラリである scikit-learn [15] を用いた．scikit-learn で利用できる SVM は，多くの研究で用いられている LIBSVM [16] を元に実装されている．また，本研究では先行研究 [1] に準じて，SVM のハイパーパラメータで誤分類の許容を調整するコストパラメータ C を，訓練データに対する 3-fold Cross Validation によって，$\{0.1, 1, 10, 100, 1000\}$ の中から最適な値を選出した．

4.2 データセット

本研究ではタンパク質-化合物相互作用およびタンパク質間類似度，化合物間類似度のデータセットとして，先行研究で挙げた Ding らによるレビュー論文に準じて Yamanishi らが 2008 年の論文 [17] で用いたデータセットを使用した．データセットはタンパク質ファミリーごとに用意され，ここでは核内受容体 (Nuclear Receptor)，イオンチャネル (Ion Channel)，G タンパク質共役受容体 (G Protein-Coupled Receptor, GPCR)，酵素 (Enzyme) の 4 種類が与えられた．以下にデータセットの内訳を示す．

表 1 データセットの内訳

	Nuclear Receptor	GPCR	Ion Channel	Enzyme
化合物数	54	223	210	445
タンパク質数	26	95	204	664
相互作用数	89	634	1475	2925
相互作用の存在割合 (%)	6.4	3.0	3.5	0.99

タンパク質-化合物相互作用情報は，生体関連情報のデータベースである KEGG BRITE [18]，酵素のデータベースである BRENDA [19]，薬剤情報のデータベースである SuperTarget [20] および DrugBank [21] から取得され，タ

[*1] PKM の文献 [7] では化合物の化学的類似度として Chem-CPP [12] の Tanimoto 係数を，タンパク質の化学的類似度として EC number の階層構造に基づく距離情報を用いている．その他，種々の化合物 fingerprint，Pfam domain fingerprint の Tanimoto 係数，Side Effect 情報なども用いられる [13] [14].

ンパク質-化合物の隣接行列の形式をとる．

タンパク質間類似度は遺伝子とタンパク質情報のデータベースである KEGG GENES database [18] から得た標的タンパク質のアミノ酸配列をもとに normalized Smith-Waterman score [22] で算出される．

化合物間類似度は生体関連化合物のデータベースである KEGG LIGAND database [18] から取得された化合物の構造をもとに SIMCOMP score [23] を用いて算出される．

4.3 評価指標

本研究では評価指標として Area Under the Receiver Operating Characteristic curve (AUROC)，Area Under the Precision Recall curve (AUPR) の 2 つを用いた．この節ではこれらの評価指標の概要を述べる．ラベルが正例と負例の 2 クラス分類問題を考える．分類器が出力する結果は以下の 4 つのパターンが考えられる．

- True Positive : 正例のものを正しく正例と予測した
- True Negative : 負例のものを正しく負例と予測した
- False Positive : 負例のものを誤って正例と予測した
- False Negative : 正例のものを誤って負例と予測した

予測モデルがデータセットに対して予測を行ったとき，True Positive となった要素の数を #TP，True Negative となった要素の数を #TN，False Positive となった要素の数を #FP，False Negative となった要素の数を #FN とする．

4.3.1 AUROC

Receiver Operating Characteristic curve (ROC 曲線) は予測モデルの出力の閾値を変化させながら，縦軸に True Positive Rate，横軸に False Positive Rate をとった曲線である．True Positive Rate (真陽性率) は正例のものの中で正しく予測できた割合であり，以下の式で求められる．

$$\text{True Positive Rate} \equiv \frac{\#\text{TP}}{\#\text{TP} + \#\text{FN}}$$

False Positive Rate (偽陽性率）とは，負例のものの中で誤って予測した割合であり，以下の式で求められる．

$$\text{False Positive Rate} \equiv \frac{\#\text{FP}}{\#\text{FP} + \#\text{TN}}$$

ROC 曲線の下側の面積 (area under the ROC curve: AUROC) は予測モデルの性能のよさを表している．理想的な予測モデル，つまり正例と負例を完全に分離できる予測モデルにおける ROC 曲線は，原点から $(0, 1)$ まで上昇し，そこから水平に $(1, 1)$ まで続き，AUROC は 1.0 となる．また，予測をランダムに行う予測モデルでは ROC 曲線は原点と $(1, 1)$ を結ぶ直線となり，AUROC は 0.5 となる．

4.3.2 AUPR

Precision-Recall 曲線は予測モデルの出力の閾値を変化させながら，縦軸に Precision，横軸に Recall をとった曲線である．Precision (適合率) とは，正例と予測されたものの中で正しく予測できた割合であり，以下の式で求められる．

$$\text{precision} \equiv \frac{\#\text{TP}}{\#\text{TP} + \#\text{FP}}$$

Recall (再現率) は True Positive Rate と同値であり，以下の式で求められる．

$$\text{recall} \equiv \text{True Positive Rate} = \frac{\#\text{TP}}{\#\text{TP} + \#\text{FN}}$$

Precision-Recall 曲線は右下がりの曲線を描き，その下側の面積 (area under the precision-recall curve: AUPR) は予測モデルの性能の良さを表している．AUPR は正例の数が負例よりも少ない場合に，AUROC よりも False Positive に対して厳しい評価を行うことができる．また，AUROC に対する最適化が AUPR に対する最適化にもなるとは限らない [24]．Ding らはタンパク質-化合物相互作用は負例よりも正例の数がかなり少ないため，False Positive に対して厳しく評価をすべきだと述べている [1]．なお予測をランダムに行う予測モデルでは，AUPR は全サンプル中の正例の割合に等しくなる．

4.4 実験方法

各実験では Ding らの論文に準じて，与えられたデータセットに対し 10-fold Cross Validation を 5 回行い，得られた値の平均値を予測精度として評価の対象とする．Cross Validation ではタンパク質 (protein)，化合物 (compound)，相互作用 (interaction) を分割の対象とする．各分割の方法は，**図 1** に示したものと対応する．タンパク質に対する Cross Validation では新規タンパク質に対する予測性能，化合物では新規化合物に対する予測性能を評価できる．また．相互作用に対する Cross Validation は既存のタンパク質-化合物間相互作用情報から未知の相互作用を予測する性能の評価を目的としている．

5. 実験結果

本章にて実験結果を示す．もしも予測をランダムに行った場合に得られる各評価値は，AUROC は 0.5，AUPR はデータセットによって異なるが**表 1** の正例の割合を平均して 0.0346 となる．

5.1 リンク指標を単体でカーネルとして用いた場合の検証

表 2 および**表 3** に各指標を化学的類似度と統合せずに，単体で類似度として用いた場合の予測性能を示す．

表 2 類似度指標の比較 (AUROC)

類似度指標	Cross Validation の対象		
	Protein	Compound	Interaction
Jaccard Index	0.601	0.623	0.611
Cosine Similarity	0.631	0.654	0.905
LHN-1	0.564	0.557	0.573

表 3　類似度指標の比較 (AUPR)

類似度指標	Cross Validation の対象		
	Protein	Compound	Interaction
Jaccard Index	0.522	0.545	0.517
Cosine Similarity	0.362	0.393	0.649
LHN-1	0.526	0.536	0.530

これらの結果より，Cosine Similarity を類似度指標として用いた場合に AUROC が最大になり，LHN-1 を類似度指標として用いた場合に AUPR が最も高くなることが平均的には確認された一方で，Cross Validation の対象によって良し悪しが大きく変化することも確認された．

5.2　各リンク指標と化学的類似度の統合結果

前節の結果から各リンク指標を単体で用いた場合の差異を見出すことが困難であったため，化学的類似度と統合して Pairwise Kernel として用いたときの結果を検証した．化学的類似度との統合方法として (1) 重み付き和，(2) 標準化重み付き和の 2 つについて実験を行う．重み w は $\{0.1, 0.3, 0.5, 1\}$ の 4 パターンを比較した結果を以下に示す． 本検証において得られた AUROC，AUPR について，最大のものに下線が引かれている．

表 4　各統合方法に対する比較 (AUROC)

類似度指標および統合方法	重み w			
	0.1	0.3	0.5	1
(1) Jaccard Index	0.898	0.896	0.898	0.884
(1) Cosine Similarity	0.902	0.903	<u>0.906</u>	0.899
(1) LHN-1	0.889	0.894	0.894	0.888
(2) Jaccard Index	0.890	0.894	0.902	0.892
(2) Cosine Similarity	0.902	0.896	<u>0.906</u>	0.901
(2) LHN-1	0.894	0.899	0.895	0.889

表 5　各統合方法に対する比較 (AUPR)

類似度指標および統合方法	重み w			
	0.1	0.3	0.5	1
(1) Jaccard Index	0.492	0.490	0.481	0.441
(1) Cosine Similarity	0.547	0.540	<u>0.562</u>	0.523
(1) LHN-1	0.490	0.502	0.500	0.487
(2) Jaccard Index	0.492	0.505	0.502	0.483
(2) Cosine Similarity	0.537	0.556	0.539	0.543
(2) LHN-1	0.494	0.507	0.487	0.493

表 4，**表 5** より AUROC，AUPR ともに (1) の単純な重み付き和を選択し，Cosine Similarity を重み $w = 0.5$ で付加した場合に予測精度が最大となった．よって，以降では Cosine Similarity を重み 0.5 で付加したものを提案法の類似度として使用する．**表 6** および**表 7** にリンク指標の有無に対する予測精度の比較結果を示す．本実験で得られた提案法の全ての結果は従来法と比較して，統計的に有意な差があることが示された ($\alpha = 0.05$)．

表 6　既存法と提案法の比較 (AUROC)

手法	Cross Validation の対象		
	Protein	Compound	Interaction
既存法 (PKM)	0.873	0.853	0.936
提案法 (PKM+リンク指標)	0.897	0.871	0.950

表 7　既存法と提案法の比較 (AUPR)

手法	Cross Validation の対象		
	Protein	Compound	Interaction
既存法 (PKM)	0.496	0.360	0.550
提案法 (PKM+リンク指標)	0.580	0.381	0.725

6.　考察

6.1　精度の評価

本実験においては，リンク指標として Cosine Similarity を重み 0.5 で付加した場合に精度が最大となった．表 6 および表 7 より AUROC および AUPR はリンク指標を付加した場合に，タンパク質，化合物，相互作用のいずれを対象とした Cross Validation に対しても有意に向上することが確認できた．

Cosine Similarity が良い結果を示した理由の考察のため，各リンク指標を用いて算出したタンパク質間類似度の分布を**図 2** に示す．図 2 より，Jaccard Index と LHN-1 は類似度の分布が 0 と 1 にはっきりと分かれているのに対し，Cosine Similarity は 0 から 1 の間にもまばらに分布していることが見て取れる．また，類似度が 1 となる要素の数も Cosine Similarity が最も多くなっている．このため，各リンク指標を単体で評価した 5.1 節の結果では類似度が 1 となる数が最も多い Cosine Similarity は True Positive Rate が高くなり，結果として AUROC が高くなったと考えられる．また，逆に類似度が 1 となる数が最も少ない LHN-1 については precision が高くなり，結果として AUPR が高くなったと考えられる．しかし，化学的類似度と組み合わせた 5.2 節の結果では Cosine Similarity が最も高い予測精度となったことから，0 か 1 かの極端な分布より，まばらな分布のほうがネットワークの構造情報としてより効果が高いと考えられる．

6.2　計算量の評価

訓練データに含まれるタンパク質数を m，化合物数を n とすると，既存法である PKM の予測モデル構築にかかる計算量は $O(m^3n^3)$ である．本実験で用いた 3 つのリンク指標を算出するための計算量は $O(mn(m+n))$ であり，提案法の計算量は $O(m^3n^3 + mn(m+n)) = O(m^3n^3)$ となるため，既存法と同じ計算量であることがわかる．実際に実行時にかかった計算時間を**表 8** に示す．極端に実行時間が短い Nuclrear Receptor を除けば，提案法は既存法に比べて計算時間の増加分は数%に留まっていた．また，相互

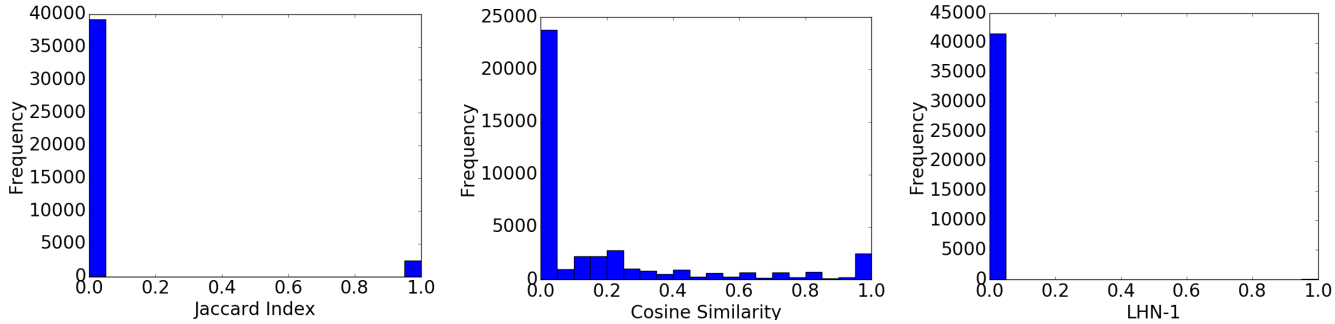

図 2 各リンク指標の分布 (Ion channel データセットでのタンパク質間類似度を示した)

表 8 実行時間の比較 (10-fold Cross Validation における 10 回の実行の内，1 回分の平均実行時間)

	Nuclear Receptor	GPCR	Ion Channel	Enzyme
既存法 （秒）	0.0680	4.86	24.1	232
提案法 （秒）	0.0850	5.17	24.8	239
実行時間の増加量（%）	25	6.4	2.9	3.3

作用が疎であるほど実行時間の増加は少ない傾向も見られた．実際のタンパク質や化合物のデータは数千から数万規模になることも多く，実用上は相互作用行列が疎になることが多い．そのため，リンク指標を付加することによる実行時間の増加はわずかであり，実用上は従来法と同等程度の時間で計算可能であるといえる．

6.3 CGBVS におけるその他の問題点の検討

本研究では，既存法がネットワークのノード情報のみを用いて予測を行っているという問題点に焦点を当てた．この問題点以外にも既存法には，実験的に負例と示されているデータがほとんどないため，正例か負例か不明なものを便宜的に負例として扱っているという問題点がある．この問題に対して，データセットの中からより負例らしい負例を抽出し，新たなデータセットを構築することで精度の向上が期待できる．このような負例構築の先行研究として Liu らの研究 [14] が挙げられる．Liu らはタンパク質の特徴量と化合物の特徴量からタンパク質-化合物ペアの距離を提案し，この距離に従って高信頼の負例を抽出した．我々はレコメンドシステムなどに用いられる協調フィルタリングを応用した，計算効率の良い高信頼負例構築の方法を検討し，一定の精度向上を見たが，ランダムサンプリングの場合と比べて統計的有意差は確認できなかった．一方で，CGBVS の手法に対してリンクマイニングの方法論は有効であると考えており，既存法との組み合わせ方法など検討する価値があると思われる．

7. まとめ

7.1 結論

本研究では，CGBVS の既存法の中で最も新規化合物に対する予測精度が高い PKM に対し，リンクマイニングの観点から，リンク指標を用いた予測精度の改良を試みた．

本研究では，指標として Cosine Similarity を用い，化学的類似度に対し重み 0.5 とした単純重み付き和が新規の類似度として最もよい予測精度を示した．結果として，AUROC および AUPR のいずれも統計的に有意な精度の向上が確認された．

7.2 今後の課題

本研究の今後の課題として以下が挙げられる．

- PKM 以外の予測手法への応用
  本研究では精度向上の対象として PKM のみに限って実験を行ったが，本研究の提案法は他の既存法に対しても容易に適用可能であるため，現在提案されている他の予測モデル (制約付きボルツマンマシン等) を用いることで PKM と組み合わせた場合よりもさらに大きな予測精度の向上を達成できる可能性がある．
- 統合方法の更なる工夫
  本実験では重み付き和と z-score による標準化重み付き和の 2 通りの統合方法を検討したが，それ以外にも最大値や相乗平均，調和平均などが考えられる．また，リンク指標は使用する類似度指標に応じて表現する値が異なるので，複数のリンク指標を用いる方法も検討する価値があると考えられる．

謝辞 本研究の一部は JSPS 科研費 基盤研究 (A) (24240044)，若手研究 (B) (15K16081)，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」の支援によって行われた．

参考文献

[1] Ding H., Takigawa I., Mamitsuka H., *et al.*: Similarity-based machine learning methods for predicting drug-target interactions: a brief review, *Brief Bioinform*, 15(5), 734–747 (2013)

[2] Cortes C., Vapnik V.: Support-vector networks, *Mach*

Learn, 20(3), 273–297 (1995)

[3] Smolensky P.: *Information processing in dynamical systems: foundations of harmony theory*, MIT Press, Cambridge, 194–281 (1986)

[4] Belur V.: *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*, IEEE Computer Society Press (1991)

[5] Menzies T., Hu Y.: Data Mining for Very Busy People, *IEEE Comput Mag*, 36(11), 18–25 (2003)

[6] Kubinyi H.: *Chemogenomics in drug discovery*, Springer Berlin Heidelberg, 1–19 (2006)

[7] Jacob L., Vert J. P.: Proteinligand interaction prediction: an improved chemogenomics approach, *Bioinformatics*, 24(19), 2149–2156 (2008)

[8] Kashima H.: ネットワーク構造予測, 人工知能学会誌, 22(3), 344–351 (2007)

[9] Brunner C., Fischer A., Luig K., *et al.*: Pairwise support vector machines and their application to large scale problems, *J Mach Learn Res*, 13(1), 2279–2292 (2012)

[10] Ben-Hur A., Noble W. S.: Kernel methods for predicting proteinprotein interactions, *Bioinformatics*, 21(1), 38–46 (2005)

[11] Hagberg A. A., Schult D. A., Swart P. J.: Exploring Network Structure, Dynamics, and Function using NetworkX, In *Proc. of the 7th Python in Science Conference*, Pasadena, CA, 11–16 (2008)

[12] Perret J., Mahe P., Vert J.: Chemcpp: an open source c++ toolbox for kernel functions on chemical compounds, http://chemcpp.sourceforge.net

[13] Chen B., Ding Y., Wild D. J.: Assessing Drug Target Association Using Semantic Linked Data, *PLoS Comput Biol*, 8(7), 1–10 (2012)

[14] Liu H., Sun J., Guan J., *et al.*: Improving compoundprotein interaction prediction by building up highly credible negative samples, *Bioinformatics*, 31(12), 221–229 (2015)

[15] Pedregosa F., Varoquaux G., Gramfort A., *et al.*: Scikit-learn: Machine Learning in Python, *J Mach Learn Res*, 12, 2825–2830 (2011)

[16] Chang C., Lin C.: LIBSVM: a library for support vector machines, *ACM T Intell Sys Tech.*, 2(27), 1–27 (2011)

[17] Yamanishi Y., Araki M., Gutteridge A., *et al.*: Prediction of drug-target interaction networks from the integration of chemical and genomic spaces, *Bioinformatics*, 24(13), 232–240 (2008)

[18] Kanehisa M., Goto S., Hattori M., *et al.*: From genomics to chemical genomics: new developments in KEGG, *Nucleic Acids Res*, 34(1), 354–357 (2006)

[19] Schomburg I., Chang A., Ebeling C., *et al.*: Brenda, the enzyme database: updates and major new developments, *Nucleic Acids Res*, 32(1), 431–433 (2004)

[20] Gunther S., Kuhn M., Dunkel M., *et al.*: Supertarget and matador: resources for exploring drugtarget relationships, *Nucleic Acids Res*, 36(1), 919–922 (2008)

[21] Wishart D. S., Knox C., Guo A. C., *et al.*: Drugbank: a knowledgebase for drugs, drug actions and drug targets, *Nucleic Acids Res*, 36(1), 901–906 (2008)

[22] Smith T. F., Waterman M.: Identification of common molecular subsequences, *J Mol Biol*, 147(1), 195–197 (1981)

[23] Hattori M., Okuno Y., Goto S., *et al.*: Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways, *J Am Chem Soc*, 125(39), 11853–11865 (2003)

[24] Davis J., Goadrich M.: The relationship between Precision-Recall and ROC curves, In *Proc. of ICML2006*, 233–240 (2006)