特集　情報システム

原著論文

インタラクトームレベルのデータセットを用いたタンパク質間相互作用予測とその応用

村上洋一*・水口賢司**

要旨：生体内のタンパク質間相互作用（PPI）の全体、タンパク質のインタラクトームを明らかにすることは、生物学的なパスウェイやタンパク質の機能を理解するために重要である。それを明らかにするために、現在の実験技術の限界を解決する形で、その相互作用を予測する計算科学的な手法がこれまでに数多く提案されてきた。筆者らは、近年、インタラクトームレベルのヒトの学習データセットを用いて、以前に開発したPSOPIAの高性能化に成功した。本研究では、新しいPSOPIAの有効性を検証するために、現在最も予測性能が高いと報告されている別の予測法との性能比較を行った。その結果、PSOPIAはより多くの信頼性の高いPPIを予測できることが示された。また、PSOPIAは、マウスやラットのPPI予測にも有効であることが示された。さらに、以上の結果を含めて、PSOPIAのさらなる高性能化や応用について議論を行うものである。

キーワード：タンパク質間相互作用予測，タンパク質間相互作用ネットワーク，相同相互作用，ビッグデータ，機械学習

Prediction of Protein-Protein Interactions with an Interactome-level Dataset and Its Applications

Yoichi MURAKAMI* and Kenji MIZUGUCHI**

Abstract: Identification of protein interactome, the whole set of protein-protein interactions（PPI）in vivo, is important to understand biological pathways and functions of many proteins. Many computational methods to predict PPIs have so far been proposed in order to make up for limitations of current experimental techniques for identifying PPIs. We have recently improved the performance of our PPI prediction method, PSOPIA, using a human training dataset at the interactome-level. In this study, the new PSOPIA was compared with a method that has recently been developed and reported to have the highest performance of the currently available methods, in order to evaluate the predictability of the PSOPIA. As a result, it could predict more PPIs with high-confidence than the reported method. Also, it was shown that the PSOPIA could predict PPIs in mouse and rat. Furthermore, from these results, we discuss the further improvement of the PSOPIA and its applications.

Keywords: Prediction of protein-protein interactions, Protein-protein interaction networks, homologous interactions, Big data, Machine Learning

*東京情報大学　総合情報学部
Faculty of Informatics, Tokyo University of Information Sciences
**国立研究開発法人医薬基盤・健康・栄養研究所　バイオインフォマティクスプロジェクト
Bioinformatics Project, National Institutes of Biomedical Innovation, Health and Nutrition

2018年５月16日受付
2018年８月29日受理

1．はじめに

生体内で起きているタンパク間の相互作用（Protein-Protein Interactions; 以下「PPI」と称する）の全体像、タンパク質のインタラクトーム（以下単に「インタラクトーム」と称する）を明らかにすることは、多くのタンパク質の生物学的機能の理解や様々な生化学的パスウェイを解明するために重要である。また、PPIの高い特異性が故に、合理的創薬、すなわち、遺伝子やタンパク質などに関する知見から特定した病因となるPPIを標的として医薬品を理論的にデザインする創薬において、PPIは有望な標的として期待されている[1]-[3]。このように重要なPPIの同定には、様々なハイスループットな実験技術、例えば、酵母ツーハイブリッド法や質量分析法に基づく方法などが利用されている。しかしながら、様々な物理化学的な要因、例えば、翻訳後修飾[4][5]、天然変性タンパク質の過渡的な構造形成[6]-[9]、また異なる生理学的条件などの要因によって、PPIを実験的に同定することは依然として困難な作業である。加えて、異なる細胞に局在する２つのタンパク質は、生体内では決して相互作用しないが、生体外では原理的に相互作用してしまうことが起こり得るかもしれない。これらの理由から実験によって誤って同定されてしまったPPI（偽陽性）の情報が実験結果に混じってしまう可能性がある。

このような実験的な限界を解決するために、既知のPPI（真陽性）から得られた特徴情報に基づいて新規のPPIを予測する計算科学的な手法が数多く提案されている[10]-[19]。例えば、タンパク質のアミノ酸配列上の隣り合う３つの残基の出現頻度[10][15][18]、部分的なアミノ酸配列の組み合わせパターン[12]やその出現頻度を標準化した値[17]、さらに各アミノ酸残基の異なる物理化学的指標に基づく数値の自己共分散[16][20]などの特徴情報を利用した手法が提案されている。また、「種Ａにおいて２つのタンパク質が相互作用することが知られており、かつ別の種Ｂにおいてそれらのタンパク質が保存されている場合、それらのタンパク質は種Ｂにおいても相互作用する可能性が高い」という考え（インターログ）に基づく方法も提案されている[21][22]。Yuらは、未だ相互作用することが知られていない２つのタンパク質（P_A, P_B）と、相互作用することが知られている近縁種のタンパク質ペアとの配列類似度の幾何平均が80％以上、あるいはそれらのBLASTの期待値の幾何平均が10^{-70}以下である場合、P_AとP_Bは同様に相互作用する可能性が高いことを統計的に明らかにした[22]。このようなインターログに基づいて、例えば、Wilesらは５種に存在する既知のPPIから新規のPPIを予測する手法、InterologFinderを開発した[23]。この方法は、与えられたタンパク質ペアに対して、既に実験結果がある場合はその情報を返し、それがない場合は予測結果を返すウェブサーバである。Chenらは、576種に存在する既知のPPIデータを統合したデータベースから、与えられた２つのタンパク質と相同なPPIを検索するウェブサーバ、PPISearchを開発した[24]。さらに、Garciaらは、複数の種のインターログ情報だけでなく、与えられた２つのタンパク質のドメインや遺伝子オントロジーの情報も利用して、新規のPPIを予測するウェブサーバ、BIPSを開発した[25]。これらの予測法は、複数の種から可能な限り多くのオーソログなPPIデータを集めることによって、予測性能を上げることができる。しかしながら、オーソログなPPIデータの存在に依存しており、それが存在していない場合は、新規のPPIを予測することは難しいという問題がある。

一方、筆者らは、Averaged One-Dependence Estimators（AODE）[26]という機械学習法を用いて、次の３つの特徴情報に基づく予測法、PSOPIAを開発した[27]。⑴既知のPPIとの配列相同性（F_{Seq}）（図１ａ）、⑵ドメインペアが既知のPPIに出現する傾向値の平均（F_{Dom}）（図１ｂ）、⑶PPIネットワークにおける相同なタンパク質間の最短距離（F_{Net}）（図１ｃ）。⑶は、「２つのタンパク質と相同な２つのタンパク質が既知のPPIネットワークにおいて近接して存在しているならば、相互作用する可能性が高い」という仮説に基づいている。近接する２つのタンパク質は、必ずしも直接的に相互作用していなくても、他の近接するタンパク質と複合体構造を形成する可能性があり、また共通した細胞内局在に存在することで相互作用する可能性が高まると考えられる。筆者らは、PPIネットワーク上にある２つのタンパク質間の最短距離が２以下のとき、62.3％の割合で相互作用する可能性が高いことを統計的に明らかにした[27]。

機械学習モデルの構築と評価に必要な学習データ

セットやテストデータセットの選択は、PPI予測にとって重要な問題である。ParkとMarcotteらは、学習モデルへの入力として２つのタンパク質を必要とする方法では、テストで使われる２つのタンパク質の両方が学習データセットでも使われている場合、片方だけが使われている場合に比べて、より高い予測性能を示す傾向があること明らかにした［28］。また彼らは、これまでに開発されたPPI予測法に関して報告されている性能は、テストデータへのバイアスがあることを指摘した［28］。さらに彼らは、テストで使われる２つのタンパク質を次の３つのクラスに分類した；（C1）両方が学習データセットと共有されている、（C2）片方だけが学習データセットと共有されている、（C3）どちらも学習データセットと共有されていない。彼らは、クラスC3に対する予測性能、すなわち、２つのタンパク質の事前知識がない予測性能は、クラスC1に対する予測、すなわち、２つのタンパク質の事前知識がある予測性能よりも、より難しいことを証明し、そしてPPIの予測法は各クラスに対して評価されるべきであると主張した［28］。

そこで、近年、筆者らは、ParkとMarcotteらのクラスC1-3のベンチマークセット［28］を用いて、彼らの評価方法に従ってPSOPIAの評価を行い、彼らのベンチマークテストで高い予測性能を示した２つの予測法（M2とM6）との性能比較を行った［29］。M2はサポートベクターマシン（SVM）に基づく予測法［30］であり、M6は学習する際に非PPIデータは必要としない、部分配列の共起頻度に基づく予測法［14］である。その結果として、確かに彼らが報告しているように、クラスC3の性能はクラスC1とC3に比べて明らかに低くなったが、全てのクラスC1-3において、F_{Seq}とF_{Net}のみに基づくPSOPIA'は、M2とM6よりもより高いAUCと$pAUC_{0.5\%}$を示した。しかしながら、F_{Dom}も含めた従来のPSOPIAでは、クラスC1-3のいずれにおいても、AUCと$pAUC_{0.5\%}$に関してPSOPIA'を超えることができなった。この理由は、彼らのベンチマークデータセットは、PPIと非PPIの数が約13,000個程度の小規模かつ数に偏りがない均衡なデータセット（１：１）、つまり生体内の実際のインタラクトームとは本質的に異なるので、彼らのデータセットからPPIと非PPIに出現するドメインペアの出現傾向値を正しく計算できず、従来のPSOPIAで使われているF_{Dom}は予測性能の向上に寄与しなかったためであると考えられる。

機械学習モデルの性能は、学習で使われるデータセットのサイズや多様性に依存している。一般的に、学習データセットを大規模化することによって、それに含まれているかもしれない偽陽性の影響を減らし、より代表的ものにすることができると考えられているが、大規模なデータセットを扱うにはコンピュータの処理能力に依存してしまう。しかしながら、近年のコンピュータ性能の向上に伴い、大規模な学習データ（ビッグデータ）が多方面で使われるようになっている。例えば、人間のプロ囲碁棋士を破ったDeepMindによって開発された囲碁プログラムであるAlphaGoは、ディープニューラルネットワークという機械学習法を用いてビッグデータを学習し、何百万あるいは何億ものパラメータを推論している［31］。

大規模なデータセットの有効性を検証するために、近年、筆者らは、ヒトの大規模かつ不均衡なPPIデータセット（PPIと非PPIの割合は１：769）を作成して、PSOPIAの予測モデルを再構築し、３分割の交差検定によって性能評価を行った［29］。その結果、以前に作成した小規模かつ不均衡な（１：400）データセットで学習した従来のPSOPIA［27］に比べて、大規模なデータセットで学習をすることによって、AUCと$pAUC_{0.5\%}$をそれぞれ＋0.10と＋0.08改善することができた。AUCと$pAUC_{0.5\%}$は、0.89と0.24であった。また、上述したParkとMarcotteらが作成した小規模かつ均衡な（１：１）ベンチマークデータセット［28］を用いた場合、F_{Dom}を用いた従来のPSOPIAの性能が低下してしまったが、大規模なデータセットでは、F_{Dom}は性能が向上に貢献していることが確認された［29］。

本研究では、大規模かつ不均衡なデータセット、すなわちインタラクトームレベルの学習データセットを用いることで性能が向上した新しいPSOPIA［29］の有効性を検証するために、ParkとMarcotteらのベンチマークテスト［28］以降に開発され、現在最も予測性能が高いと報告されている予測法との性能比較を行う。また、新しいPSOPIAは、ヒトのPPIデータに基づいて開発しているが、創薬等の実験モデル動物として利用されるマウスやラットなどのオーソ

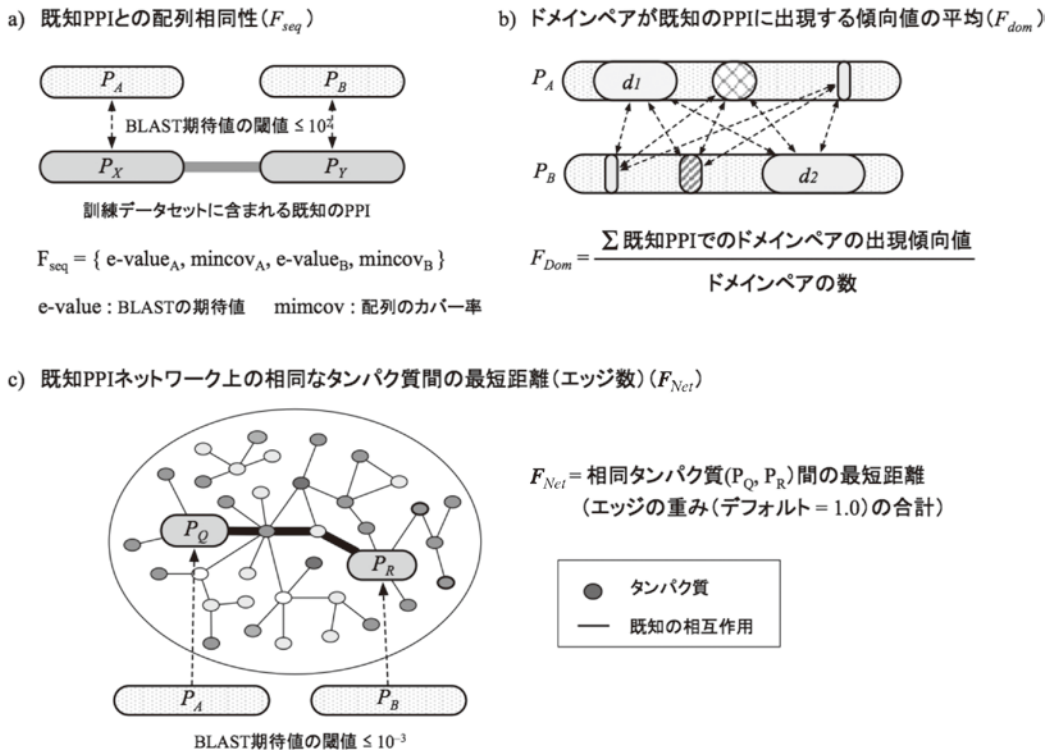

図１　PSOPIAで用いられる特徴情報（文献[27]の図１より改変）

ログなPPI予測への可能性を評価する。さらに、以上の性能比較や評価の結果を含めて、PSOPIAのさらなる高性能化や応用について議論を行う。

２．研究方法

２．１　AODEを用いたPPI予測法：PSOPIA

AODEは、単純ベイズ分類器の独立性仮定を緩和して、１つの特徴との依存を許した確率分類器である[26]。これまでにこの機械学習法は、複数の予測法の出力結果を合成することに応用され、計算量を増大させることなく、大規模かつ不均衡なデータセットを用いて予測モデルを構築できることが報告されている[26][32][33]。またAODEは、モデル選択やパラメータの最適化が必要ないため、大規模なデータセットを用いて予測モデルの再構築を容易にできるという利点もある。AODEの詳細については文献[26]を参照。

PSOPIAは、第１章で説明したタンパク質ペアの３つの特徴情報（F_{Seq}, F_{Dom}, F_{Net}）を用いて学習したAODEモデルである（図１）。F_{Dom}とF_{Net}については常に１つの特徴値が決まるが、タンパク質（A, B）に対して２通りの並び順（A-B, B-A）があるため、F_{Seq}については少なくとも２つの特徴値の並び順、つまり２つの特徴ベクトルがある；F_{Seq}={e-value$_A$, mincov$_A$, e-value$_B$, mincov$_B$}とF_{Seq}'={e-value$_B$, mincov$_B$, e-value$_A$, mincov$_A$}。e-valueとmincovは、与えられた２つのタンパク質のどちらか片方のタンパク質と、学習データセットの中にある相互作用するタンパク質ペアのどちらか片方のタンパク質との、期待値と長い配列に対するカバー率である。そこで、高次元の特徴ベクトル空間では、F_{Seq}とF_{Seq}'はどちらか一方の半空間に存在することから、片方の半空間にある特徴ベクトルのみを用いて予測モデルの構築を行った。半空間にある特徴ベクトルを決定する詳細な方法については文献[27]を参照。

２．２　インタラクトームレベルの大規模なデータセット

非PPIの数は、実際、PPIの数よりも圧倒的に多い。網羅的にキュレーションしたPPIデータを登録しているBioGridデータベース[34]には、2018年５月現在、22,514個のヒトの遺伝子から構成される322,610個のPPIデータが登録されている。この数から、253,428,841（＝22,514×(22,514－1)÷2）個の可能なタンパク質ペア（相同なタンパク質のペアを考慮しない）があり、253,106,231（＝253,428,841－322,610）個の非PPIがあると推定できる。すなわ

ち、現時点で、ヒトのインタラクトームにおけるPPIと非PPIとの割合は、１：785であると推定される。このBioGridで見られるようなヒトのインタラクトームを反映するような大規模かつ不均衡な非冗長なデータセットを、次の⑴から⑶の手順に従って作成した。⑴２つ以上の実験で相互作用が確認されている、あるいは２つ以上の学術論文で報告されている信頼性が高い直接的かつ物理的に相互作用するPPIデータセット（High Confidential Direct Physical PPI; HCDP）をTargetMine[35]から取得する。このデータセットには、17,652個のタンパク質から構成される152,562相互作用が含まれており、それらは145個の分離したPPIネットワークを構成している。⑵取得したデータセットを構成する全てのタンパク質のアミノ酸配列を、CD-HITプログラム[36]を用いて、配列類似性が40％以上になるようにクラスタリングする。⑶全ての可能なクラスタペアから次の３つの方法によりPPIあるいは非PPIを選択する; クラスタペアあるいはシングルクラスタにおいて、（ｉ）１個のPPIを含む場合、そのPPIを新規データセットに加える（図２- ｉ）。（ii）２つ以上のPPIを含む場合、クラスタの代表タンパク質から構成されるPPIを新規データセットに加える。代表タンパク質のみから構成されていない場合は、各PPIの各タンパク質が属するクラスタの代表タンパク質に対する配列類似度を計算し、２つの配列類似性に基づく合成ベクトル（Resultant Vector; 以下、RVと称する）を計算する。そして、最も大きなRVを持つPPIを選択して新規データセットに加える（図２- ii）。（iii）PPIを含まない場合、各クラスタの代表タンパク質ペアは非PPIデータとして新規データセットに加えられる（図２- iii）。結果として、43,060個のPPIデータが新規データセットに保持され、33,098,951個の非PPIデータが生成された[29]。新規データセットにおける、PPIと非PPIの割合は１：769である。これはBioGridで推定された割合（１：785）に近似している。

２.３ 予測性能の評価指標

予測モデルの性能は、ROC曲線下面積（Area Under the Curve; 以下、「AUC」と称する）によって評価する。ROC曲線は、予測モデルの比較に最もよく使われる評価指標であり、横軸に偽陽性率（１－真陰性率）、縦軸に真陽性率をプロットしたときにできる曲線である。AUCはその曲線下の面積である。この値が1.0の場合、モデルは理想的であると評価され、一方、0.5の場合は、ランダムに作成されたモデルであると評価される。AUCは、データセットの不均衡性の影響を受けることなくモデル

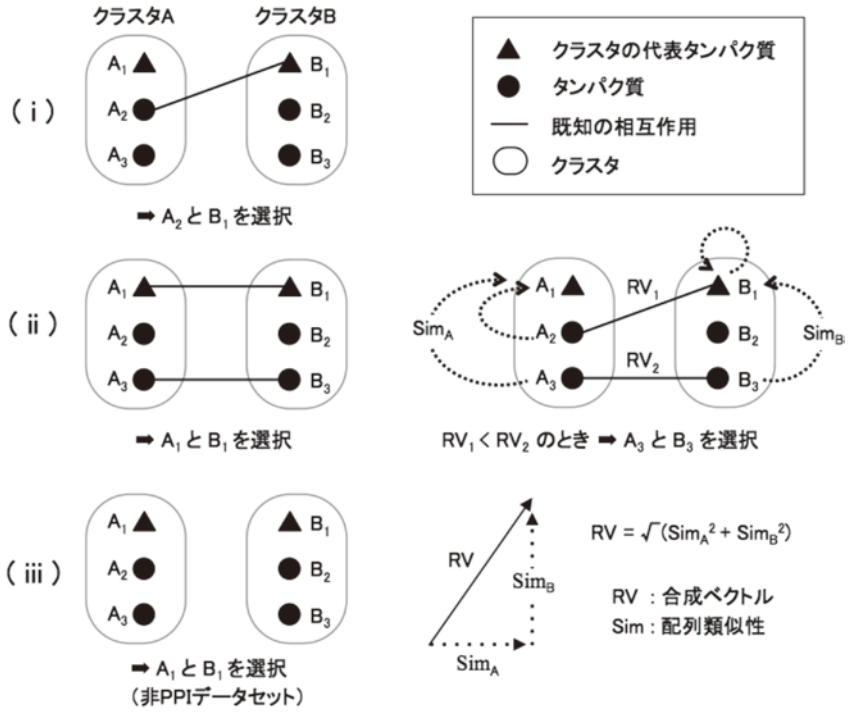

図２ アミノ酸配列クラスタペアからのPPIの選択と非PPIの生成（文献[29]の図２より改変）

を評価できることが知られており、信頼のある性能評価指標であると考えられる[37]。加えて、偽陽性率がx%以下までの標準化された部分的なAUC（$pAUC_{x\%}$）によっても評価する[10][19]。本研究では、xは0.5%とする。より高いpAUCを持つモデルは、偽陽性を抑えながら、より多くの真陽性データを予測できることを示している[10]。また、ある閾値に対して陽性と陰性に分類した後、陽性と判定されたうち真陽性である確率、つまり陽性的的中率（＝真陽性÷（真陽性＋偽陽性））も評価の指標として利用する。

3．結　果

3．1　近年開発された予測法との性能比較

ParkとMarcotteらのクラスC1-3のベンチマークテスト[28]以降、タンパク質の進化的なプロフィールをSVMを用いて学習してPPIを予測する方法、profppikernel、が開発された[38]。この予測法は、各タンパク質を、20個のアミノ酸残基のk乗（20^k）の特徴値からなる特徴ベクトルとして表している。各特徴値は、k個の残基からなる特定の部分配列（k-mer）が、タンパク質の進化的なプロフィールに出現する回数である。例えば、k＝3のとき、各タンパク質は20^3＝8,000個の異なる部分配列パターンの回数からなる特徴ベクトルになる。Profppikernelでは、k-mer回数からなる2つの特徴ベクトルのドット積に基づくプロフィールカーネルを利用してPPI予測を行っている[38]。HampとRostらは、実験で得られたタンパク質と十分な配列類似性がないタンパク質に対するPPI予測の精度を改善したと報告している[38]。また、profppikenelを用いて、ヒトの既知のPPIではない、クラスC2とC3のタンパク質ペア全てに対して網羅的に予測を行い、予測スコアが高い上位1万個のタンパク質ペアをウェブ上で公開している。

そこで、HIPPIE（Human Integrated Protein-Protein Interaction Reference）と呼ばれる、ヒトの各PPIデータに対して、機能的なアノテーションだけでなく、信頼性スコア（以下、HIPPIEスコアと称する）を付与しているデータベース[39]と、HampとRostらが公開しているクラスC2とC3のデータを用いて、profppikernelと新しいPSOPIA[29]との性能比較を実施した。各PPIに対するHIPPIEスコアは、次の3つの数に基づく各スコアを重み付けして合計した値として定義されている；① そのPPIを検出した調査の数、② そのPPIを検出するために使われた異なる実験技術の数、③ それと同様のPPIが確認されたヒト以外の種の数。また、このスコアが、0.63以上0.73未満のときデータの信頼性は中程度（medium confidence）であり、0.73以上のときデータの信頼性が高い（high confidence）と定義されている[39]。図3は、HampとRostらが公開しているクラスC2とC3のPPI予測データの中から、HIPPIEスコアが計算されているデータのみを取得し、またPSOPIAで用いられている大規模な学習データセット（2．2節参照、また予測性能の詳細は文献[29]を参照）との関係においてもクラスC2またはC3の関係にあるデータのみを対象にして、profppikernelとPSOPIAの予測性能を比較した結果を示している。

クラスC2のPPI予測データにおいて、HIPPIEスコアがmediumまたはhigh confidenceの場合、28個と6個のデータが取得され、またPSOPIAとprofppikernelのスコア（0 - 1の範囲）の相関係数は0.04と0.92であった（図3 - 1）。high confidenceの場合は、PSOPIAとprofppikernelのスコアに高い相関関係が見られた。各予測法のスコアが0.8以上の場合、HIPPIEスコアがmediumとhigh confidenceのそれぞれにおいて、PSOPIAが53.6%と50.0%であるのに対してprofppikernelは42.9%と33.3%であった。すなわち、PSOPIAはHIPPIEスコアがmedium confidence以上のタンパク質ペアをprofppikernelよりもより多く予測できた。クラスC3のPPI予測データにおいて、HIPPIEスコアがmediumまたはhigh confidenceの場合、44個と7個のデータが取得され、PSOPIAとprofppikernelのスコアには相関係数は0.15と－0.26であった（図3 - 2）。いずれのconfidenceにおいても、相関関係は見られなかった。また、各予測法のスコアが0.8以上の場合は、mediumとhigh confidenceのそれぞれにおいて、PSOPIAが20.5%と14.3%であるのに対してprofppikernelは11.4%と0%であった。以上のことから、クラスC3の場合、高いHIPPIEスコアを持つタンパク質ペアをクラスC2の場合と同様若しくはそれよりも多く予測することができなかったが、新しいPSOPIAは、いずれのクラスにおいても、profppikernelよりもより多くの高いHIPPIEスコアを持つタンパク質ペアを予測できた。

ProfppikernelとPSOPIAの予測モデルの構築に要する処理時間（学習時間）と予測に要する処理時間（予測時間）の比較をするために、profppikernelをPSOPIAと同じ計算環境（OS：Red Hat Enterprise Linux Server release 6.1，CPU：Intel Xeon E5-2670 2.60GHz，メモリ：64GB）に実装し、同じデータセットを用いて１CPUで学習とテストを実行した際に要する処理時間の比較を行った。データセットは、ParkとMarcotteらのベンチマークデータのサブデータセットである、13,887個のPPIデータと同じ個数の非PPIデータからなる訓練データセット（5,272個のタンパク質）と、1,542個のPPIデータと同じ個数の非PPIデータからなるテストデータセット（3,302個のタンパク質）を用いた。その結果として、学習時間は、profppikernelが約12時間26分であるのに対して、PSOPIAは約17時間37秒であった。PSOPIAの学習時間の大部分は、３つの特徴値を算出して特徴ベクトルを生成する前処理で、AODEの確率モデルを生成する後処理は約６分であった。また予測時間は、profppikernelが約１時間28分であるのに対して、PSOPIAは約９分であった。すなわち、PSOPIAの学習時間はprofppikernelの約1.4倍を要するが、予測時間はprofppikernelの約1/10であることが示された。

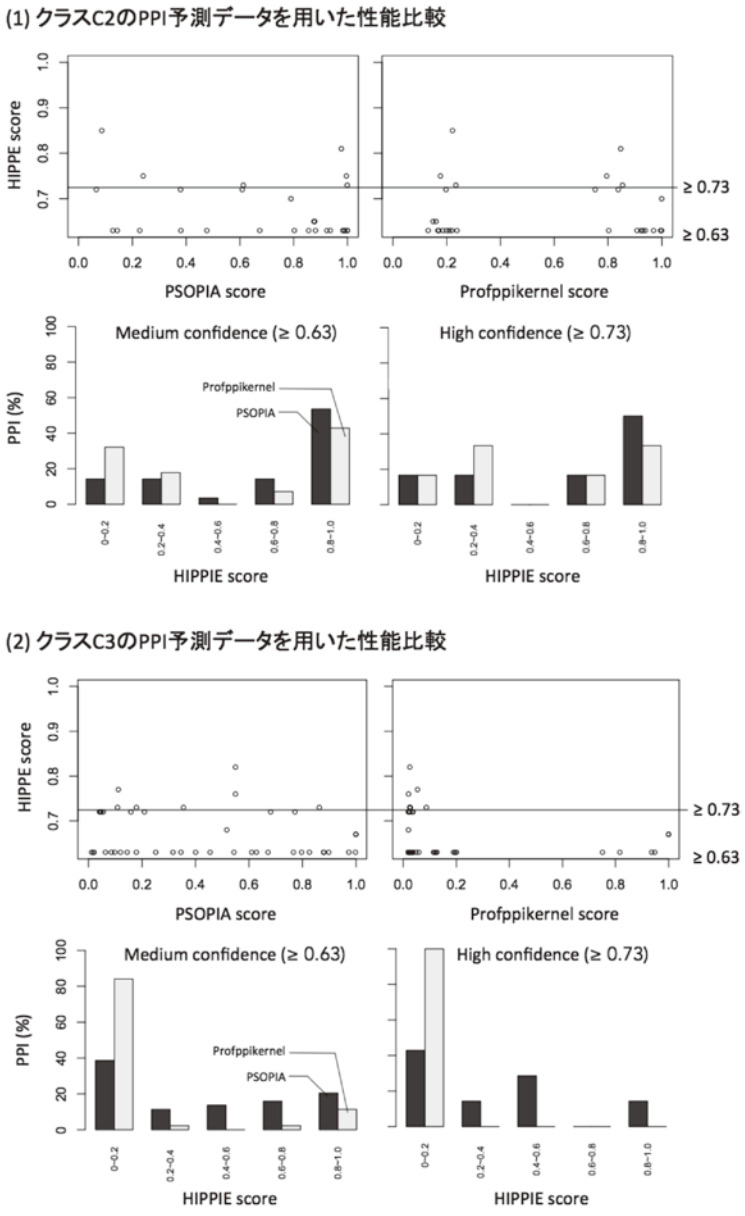

図３　HIPPIEスコアが与えられたクラスC2・C3のPPI予測データを用いた性能比較

3.2　PSOPIAを用いたオーソログなPPI予測への可能性

PSOPIAは、ヒトのPPIデータのみを用いて開発しているが、オーソログなPPI予測への応用の可能性について検証するために、創薬等の実験モデル動物として利用されるマウスやラットのテストデータセットを準備して性能評価を行った。そのような予測が可能であれば、これらの種のPPIのアノテーション情報を増やすことができ、またタンパク質の機能解析に貢献できると考えられる。テストデータセットは、PPIと非PPIの割合が均衡（1：1）とし、種ごとに次の手順に従って作成した；① TargetMine[35]から信頼性が高いPPIを取得して、PPIデータセットとする。② 追実験や異なる実験解析が実施されていないため信頼性が低いと見なされているPPIを取得する。③ それらのPPIを構成するタンパク質をクラスタリングし、PSOPIAの学習データセットの作成方法（図2）に基づいて非冗長な非PPIデータセットを作成する。結果として、マウスにおいては、3,094個のPPIデータセット（1,713個のタンパク質）と、3,094個の非冗長な非PPIデータセット（3,564個のタンパク質）を得ることができた。ラットにおいては、396個のPPIデータセット（390個のタンパク質）と396個の非冗長な非PPIデータセット（641個のタンパク質）を得ることができた。

表1は、閾値[29]ごとの陽性率、偽陽性率及び陽性的中率を示している。マウスでは、閾値が0.975のとき、陽性と判別された62.3%のタンパク質ペアうち、98.6%が真陽性であった。ラットでは、同じ閾値のとき、陽性と判別された39.6%のタンパク質ペアうち、95.9%が真陽性であった。図4は、ROC曲線、AUC及び$pAUC_{0.5\%}$を示している。マウスでは、偽陽性率が0.5%以下のときの$pAUC_{0.5\%}$は0.412であり、ラットでは、$pAUC_{0.5\%}$は0.062であった。以上の結果から、PSOPIAは、ヒトゲノムと極めて相同性が高いゲノムを持つマウスのオーソログなPPIを特異的に予測することができ、またマウスと近縁関係にあるラットにおいても高い精度で予測ができることが示された。

表1　閾値ごとのPSOPIAを用いたマウスとラットのオーソログPPI予測性能

種	閾値	陽性率（%）	偽陽性率（%）	陽性的中率（%）
マウス	0.975	62.3	1.13	98.6
	0.992	53.0	0.56	99.1
	0.996	47.9	0.27	99.6
	0.997	45.6	0.23	99.6
ラット	0.975	39.6	2.04	95.9
	0.992	35.4	2.04	95.5
	0.996	30.9	1.70	95.7
	0.997	28.1	1.02	97.1

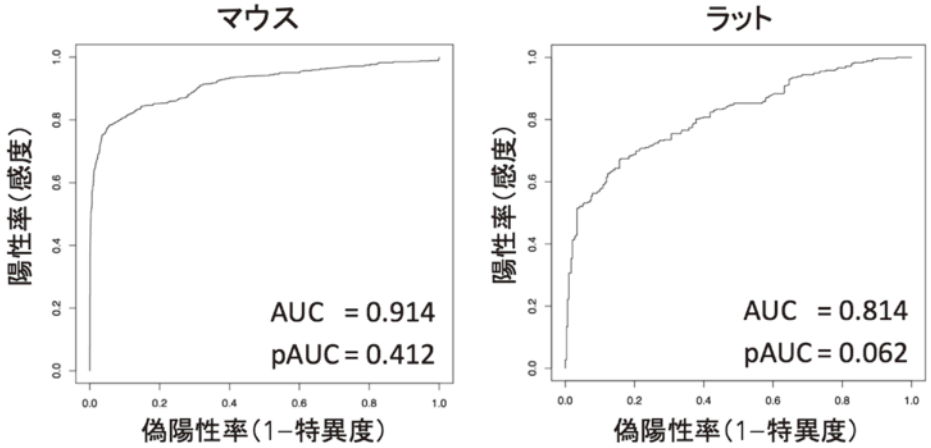

図4　PSOPIAを用いたマウスとラットのオーソログPPI予測の検証結果：AUCと$pAUC_{0.5\%}$

４．考　察

計算科学的な手法によるPPI予測は、実験によるPPI同定の限界を補完し、膨大な数のPPIの生物学的な機能を効率的に解明するために必要であり、またPPIネットワーク解析から得られた生物学的な視点に説得力を与えるためにも必要である。タンパク質間の複合体構造予測（Protein-Protein Docking; 以下、PPDと称する）の近年の進歩は、信頼性のある複合体モデルの生成を可能にしつつあるが、構造が不明あるいは不正確なタンパク質に対してPPDは有効な複合体モデルを生成することは難しく、構造変化を伴うPPDは未だ難しい[40][41]。そのような理由から、複合体構造の有無に依存しない、配列情報のみからPPIを予測する筆者らが開発するPSOPIAは有用である。

機械学習モデルの性能は、データセットのサイズと多様性に依存している。そこでPPIと非PPIを分離する明確な特徴の違いを見出すために、十分な数の信頼性の高いPPIや非PPIを準備したいところだが、実際、非PPIの数は限られている。例えば、Negatomeデータベース2.0のCombinedデータセット（ManualとPDBデータセット）では、3,475個のタンパク質から成る6,542個の非PPIしか利用できない[42]。実際は、非PPIの数はPPIの数に比べて圧倒的に多い。そのような現状において、一般的に、相互作用することが知られていない２つタンパク質をランダムにサンプリングして、それらを非PPIとしてみなして利用している。しかしながら、その数は膨大であり、多くの予測法では、コンピュータメモリやCPUタイムの必要性から、それらのデータをそのまま取り扱うことが難しいため、計算処理可能な数の非PPIデータセットを作成して利用している。一方、PSOPIAで採用しているAODEは、計算コストを増大することなしに大規模なPPIデータを扱うことができるため、ヒトのインタラクトームレベルのデータセットを用いて、予測モデルを構築することができる。インタラクトームレベルのデータセットを用いて学習した新しいPSOPIAは、以前に構築した比較的に小規模かつ不均衡なヒトのPPIデータセットを用いて学習された従来のPSOPIAよりもより高い予測性能を達成することができた[29]。また３．１節において、近年に提案されたタンパク質の進化的なプロフィールをSVMを用いて学習してPPIを予測するprofppikernelと比べて、新しいPSOPIAはより多くの高いHIPPIEスコアを持つクラスC2またはC3のタンパク質ペアを予測できることがわかった。加えて、PSOPIAの学習時間はprofppikernelの約1.4倍であり、予測時間は約1/10であることが示された。予測モデルは、並列計算によって一度構築してしまえばよく、利用者にとっては予測が高速であることが重要であると考える。以上のことから、AODEを用いたインタラクトームレベルのデータセットに基づくPSOPIAは新規のPPIの同定に有用であると考える。

PPIを実験的に同定するには物理化学的な要因による限界があるため、実際には生体内で相互作用しないにもかかわらず、実験的に相互作用が確認されたタンパク質ペアが存在している可能性がある。そのため、一度の実験だけでは相互作用の信頼性は担保されず、３．１節で述べたHIPPIEスコアのように、追実験や異なる実験技術による検証などを実施することにより、データの信頼性を高めることができる。しかしながら、そのような実験は時間やコスト的な問題があるため、未だ誤ったPPI情報が紛れ込んでしまっている可能性もある。一方、実際に生体内で相互作用するにもかかわらずそのような実験が実施されていないために信頼性の低い情報して扱われてしまっているものもある。そこで、新しいPSOPIAは、実験で同定されたPPIデータから、偽陽性を取り除くだけでなく、各PPIデータにPSOPIAスコアを付与することにより、真陽性である可能性が高いか否かを示す指標として利用することができると考える。筆者らは、高い陽性的中率が期待される閾値を決定しており[29]、これらの閾値を用いて、偽陽性と真陽性をある程度確率的に判別することができる。例えば、TargetMineに登録されている信頼性の低いと見なされているPPIデータのうち、閾値が0.996のとき、真陽性率が16.1％で、57,882個のPPIデータ（全体の8.0％）は相互作用している可能性が高いことを示した[29]。今後、各PPIデータに対して計算されたPSOPIAスコアをTargetMineのようなデータウェアハウスに統合し、HIPPIEスコアと同様に、PPIの信頼性を評価する指標として用いることにより、パスウェイ解析やネットワーク解析などにより説得力のある説明を与えることができるようになると期待される。

インターログを用いて、与えられたタンパク質と、データベースに登録されている他のタンパク質との相互作用を予測（相互作用パートナー予測）するウェブサーバが幾つか存在する[23][25][43]。これらのサーバは機械学習を採用しておらず、オーソログなPPIの存在に依存しており、その情報がない場合は新規PPIを発見することは難しい。PSOPIAでも、特徴情報F_{Seq}は、相同なPPI情報に依存しており、そのような情報が得られない場合、残りの特徴情報F_{Dom}とF_{Net}で予測できるように設計されている。しかしながら、そのような状況の場合、他の種のオーソログなPPI情報があれば、それは有効な特徴情報になり得る。３．２節では、ヒトのPPIデータのみに基づくPSOPIAを用いて、マウスやラットのオーソログなPPIの予測を行った。その結果、PSOPIAは、ヒトゲノムと極めて相同性が高いゲノムを持つPPIを特異的に予測することができ、またマウスと近縁関係にあるラットにおいても高い精度で予測が可能であることがわかった。以上のことから、逆に、マウスやラットのPPIデータに基づいたAODEモデルを用いて、ヒトのPPI予測の可能性は十分に期待できることから、今後は同一種のだけでなくオーソログなPPI情報を含めることで、さらなる高性能化が目指せるのではないかと考える。

５．結　論

ヒトのインタラクトームレベルの大規模かつ不均衡なデータセット（１：769）を用いて開発した新しいPSOPIA[29]と、ParkとMarcotteらのクラスC1-3のベンチマークテスト[28]以降に開発され、現在最も予測性能が高いと報告されている予測法、profppikernelとの性能比較を行った。その結果、クラスC3の場合、PSOPIAとprofppikernelのどちらも高いHIPPIEスコアを持つタンパク質ペアの予測は困難であったが、新しいPSOPIAは、いずれのクラスにおいても、profppikernelよりもより多くの高いHIPPIEスコアを持つタンパク質ペアを予測することができた。またPSOPIAは、ヒトのPPIデータのみを用いて開発しているが、他の種のオーソログなPPI予測への応用の可能性を評価するために、マウスやラットのテストデータセットを準備して性能評価を行った。その結果、PSOPIAは、ヒトゲノムと極めて相同性が高いゲノムを持つマウスや、それと近縁関係にあるラットのPPI予測に有効であることが示された。新しいPSOPIAは、今後、外部のデータウェアハウスなどと統合されることでPPIネットワークやパスウェイ解析への貢献が期待される。また実験によって決定されたが追実験や異なる実験解析によって検証されていないため、未だ信頼性が担保されていないPPIデータから、偽陽性を取り除き、また信頼性を示す指標としても利用されることが期待される。

【引用文献】

[１] Jubb, H., Blundell, T. L. and Ascher, D. B.: Flexibility and small pockets at protein-protein interfaces: New insights into druggability, *Prog Biophys Mol Biol*, Vol.119, pp.2-9 (2015).

[２] Prathipati, P. and Mizuguchi, K.: Systems Biology Approaches to a Rational Drug Discovery Paradigm, *Curr Top Med Chem*, Vol.16, pp.1009-1025 (2016).

[３] Wells, J. A. and McClendon, C. L.: Reaching for high-hanging fruit in drug discovery at protein-protein interfaces, *Nature*, Vol.450, pp.1001-1009 (2007).

[４] Duan, G. and Walther, D.: The roles of post-translational modifications in the context of protein interaction networks, *PLoS Comput Biol*, Vol.11, p.e1004049 (2015).

[５] Seet, B. T., Dikic, I., Zhou, M. M. and Pawson, T.: Reading protein modifications with interaction domains, *Nat Rev Mol Cell Biol*, Vol.7, pp.473-483 (2006).

[６] Acuner-Ozbabacan, S. E., Engin, H. B., Gursoy, A. and Keskin, O.: Transient protein-protein interactions, *Protein Eng Des Sel*, Vol.24, pp.635-648 (2011).

[７] Babu, M. M., Kriwacki, R. W. and Pappu, R. V.: Structural biology. Versatility from protein disorder, *Science*, Vol.337, pp.1460-1461 (2012).

[８] Lua, R. C., Marciano, D. C., Katsonis, P., Adikesavan, A. K. Wilkins, A. D. and Lichtarge, O.: Prediction and redesign of protein-protein interactions, *Prog Biophys Mol Biol*, Vol.116, pp.194-202 (2014).

[９] Meszaros, B., Simon, I. and Dosztanyi, Z.: Prediction of protein binding regions in disordered proteins, *PLoS Comput Biol*, Vol.5, p.e1000376 (2009).

[10] Ben-Hur, A. and Noble, W. S.: Kernel methods for predicting protein-protein interactions, *Bioinformatics*, Vol.21 Suppl 1, pp.i38-46 (2005).

[11] Bock, J. R. and Gough, D. A.: Predicting protein--protein interactions from primary structure, *Bioinformatics*, Vol.17, pp.455-460 (2001).

[12] Martin, S., Roe, D. and Faulon, J. L.: Predicting protein-

protein interactions using signature products, *Bioinformatics*, Vol.21, pp.218-226（2005）.

［13］ Sprinzak, E. and Margalit, H.: Correlated sequence-signatures as markers of protein-protein interaction, *J Mol Biol*, Vol.311, pp.681-692（2001）.

［14］ Pitre, S., Dehne, F., Chan, A., Cheetham, J., Duong, A., Emili, A., Gebbia, M., Greenblatt, J., Jessulat, M., Krogan, N., Luo, X. and Golshani, A.: PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs, *BMC Bioinformatics*, Vol.7, p.365（2006）.

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

［16］ Guo, Y., Yu, L., Wen, Z. and Li, M.: Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences, *Nucleic Acids Res*, Vol.36, pp.3025-3030（2008）.

［17］ Roy, S., Martinez, D., Platero, H., Lane, T. and Werner-Washburne, M.: Exploiting amino acid composition for predicting protein-protein interactions, *PLoS One*, Vol.4, p.e7813（2009）.

［18］ Yu, C. Y., Chou, L. C. and Chang, D. T.: Predicting protein-protein interactions in unbalanced data using the primary structure of proteins, *BMC Bioinformatics*, Vol.11, p.167,（2010）.

［19］ Yu, J., Guo, M., Needham, C. J., Huang, Y., Cai, L. and Westhead, D. R.: Simple sequence-based kernels do not predict protein-protein interactions, *Bioinformatics*, Vol.26, pp.2610-2604,（2010）.

［20］ Guo, Y., Li, M., Pu, X., Li, G., Guang, X., Xiong, W. and Li, J.: PRED_PPI: a server for predicting protein-protein interactions based on sequence data with probability assignment, *BMC Res Notes*, Vol.3, p.145（2010）.

［21］ Matthews, L. R., Vaglio, P., Reboul, J., Ge, H., Davis, B. P., Garrels, J., Vincent, S. and Vidal, M.: Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs", *Genome Res*, Vol.11, pp.2120-2126（2001）.

［22］ Yu. H., Luscombe, N. M., Lu, H. X., Zhu, X., Xia, Y., Han, J. D., Bertin, N., Chung, S., Vidal, M. and Gerstein, M.: Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs, *Genome Res*, Vol.14, pp.1107-1118（2004）.

［23］ Wiles, A. M., Doderer, M., Ruan, J., Gu, T. T., Ravi, D., Blackman, B. and Bishop, A. J.: Building and analyzing protein interactome networks by cross-species comparisons," *BMC Syst Biol*, Vol.4, p.36（2010）.

［24］ Chen, C. C., Lin, C. Y., Lo, Y. S. and Yang, J. M.: PPISearch: a web server for searching homologous protein-protein interactions across multiple species, *Nucleic Acids Res*, Vol.37, pp.W369-375（2009）.

［25］ Garcia-Garcia, J., Schleker, S., Klein-Seetharaman, J. and Oliva, B.: BIPS: BIANA Interolog Prediction Server. A tool for protein-protein interaction inference, *Nucleic Acids Res*, Vol.40, pp.W147-154（2012）.

［26］ Webb, G. I., Boughton, J. R. and Wang, Z.: Not So Naive Bayes: Aggregating One-Dependence Estimators, *Mach. Learn.*, Vol.58, pp.5-24（2005）.

［27］ Murakami, Y. and Mizuguchi, K.: Homology-based prediction of interactions between proteins using Averaged One-Dependence Estimators, *BMC Bioinformatics*, Vol.15, p.213,（2014）.

［28］ Park, Y. and Marcotte, E. M.: Flaws in evaluation schemes for pair-input computational predictions, *Nat Methods*, Vol.9, pp.1134-1136（2012）.

［29］ Murakami, Y. and Mizuguchi, K.: PSOPIA: Toward more reliable protein-protein interaction prediction from sequence information, in *2017 International Conference on Intelligent Informatics and Biomedical Sciences（ICIIBMS）*, Okinawa, Japan（2017）.

［30］ Vert, J. P., Qiu, J. and Noble, W. S.: A new pairwise kernel for biological network inference with support vector machines, *BMC Bioinformatics*, Vol.8 Suppl 10, p.S8,（2007）.

［31］ Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Driessche, G. van den., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T. and Hassabis, D.: Mastering the game of Go with deep neural networks and tree search, *Nature*, Vol.529, pp.484-489（2016）.

［32］ Garcia-Jimenez, B., Juan, D., Ezkurdia, I., Andres-Leon, E. and Valencia, A.: Inference of functional relations in predicted protein networks with a machine learning approach, *PLoS One*, Vol.5, p.e9969（2010）.

［33］ Webb, G. I., Boughton, J. R., Zheng, F. K., Ting, M. and Salem, H.: Learning by extrapolation from marginal to full-multivariate probability distributions: decreasingly naive Bayesian classification, *Machine Learning*, Vol.86, pp.233-272（2012）.

［34］ Chatr-aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A., Stark, C., Breitkreutz, B. J., Dolinski, K. and Tyers, M.: The BioGRID interaction database: 2017 update, *Nucleic Acids Res*, Vol.45, pp.D369-D379（2017）.

［35］ Chen, Y. A., Tripathi, L. P. and Mizuguchi, K.: An

integrative data analysis platform for gene set analysis and knowledge discovery in a data warehouse framework, *Database (Oxford)*, Vol.2016, pp.1-14 (2016).

[36] Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W.: CD-HIT: accelerated for clustering the next-generation sequencing data, *Bioinformatics*, vol.28, pp.3150-3152 (2012).

[37] Fawcett, T.: An introduction to ROC analysis, *Pattern Recogn. Lett.*, Vol.27, pp.861-874 (2006).

[38] Hamp, T. and Rost, B.: Evolutionary profiles improve protein-protein interaction prediction from sequence, *Bioinformatics*, Vol.31, pp.1945-1950 (2015).

[39] Schaefer, M. H., Fontaine, J. F., Vinayagam, A. Porras, P. Wanker, E. E. and Andrade-Navarro, M. A.: HIPPIE: Integrating protein interaction networks with experiment based quality scores, *PLoS One*, Vol.7, p.e31826 (2012).

[40] Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J., Vajda, S., Vakser, I. and Wodak, S.J.: CAPRI: a Critical Assessment of Predicted Interactions, *Proteins*, Vol.52, pp.2-9 (2003).

[41] Janin, J. and Wodak, S.: The third CAPRI assessment meeting Toronto, Canada, April 20-21, 2007, *Structure*, Vol.15, Issue7, pp.755-759 (2007).

[42] Blohm, P., Frishman, G., Smialowski, P., Goebels, F., Wachinger, B., Ruepp, A. and Frishman, D.: Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis, *Nucleic Acids Res,* Vol.42, pp.D396-400 (2014).

[43] Brown, K. R. and Jurisica, I.: Online predicted human interaction database, *Bioinformatics*, Vol.21, Issue9, pp.2076-2082 (2005).