Vol.2012-MPS-91 No.14
Vol.2012-BIO-32 No.14
2012/12/6

構造情報に基づく タンパク質間相互作用ネットワーク予測精度の改善

山本 航平[1]　大上 雅史[1,2]　石田 貴士[1]　秋山 泰[1]

概要：我々はタンパク質間相互作用ネットワークを予測することを目的としてタンパク質間ドッキング計算に基づいたタンパク質間相互作用予測手法を開発している．従来相互作用予測を行う際は，予測対象のタンパク質ペアの情報のみを用いて相互作用評価値の計算を行っていたが，本研究ではよりネットワーク予測に適した手法として，対象ペアが，他の相互作用候補ペアと比べてどの程度相互作用し得るかを相対的に評価することで予測精度を向上させることに成功した．

キーワード：タンパク質間ドッキング, タンパク質間相互作用，相互作用ネットワーク予測

Improvement of the Accuracy for Protein-Protein Interaction Network Prediction Based on Tertiary Structural Information

Yamamoto Kohei[1]　Ohue Masahito[1,2]　Ishida Takashi[1]　Akiyama Yutaka[1]

Abstract: We have developed a method to predict protein-protein interaction based on docking calculation for getting information of protein-protein interaction networks. In this study, we propose a new method to predict a protein-protein interaction network. Our proposed method uses not only the docking results of a pair but also the results of the other protein pairs including a protein of the target pair. With the new method, we succeeded to improve the accuracy of the prediction of protein-protein interaction networks.

Keywords: Protein-Protein Docking, Protein-Protein Interaction, Prediction of Protein-Protein interaction Network

1. はじめに

タンパク質は生体内で複数が相互作用することでその機能を発揮することが知られている．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，生命現象の中心的役割を果たしており，PPI を解明することで生命現象の理解が進むことが期待されている．

図 1 はヒトアポトーシスパスウェイに関わるタンパク質群を示しているが．図 1 に示されたタンパク質群は，互いに相互作用することでアポトーシスを制御する．この例に見られるように，多数のタンパク質が生体内の機能の制御に相互に関わっており，その制御構造を理解するにはネットワークとしての相互作用関係の理解が必要である．したがって，タンパク質間相互作用から生命現象を理解しようとしたとき，より理解を進めるためにはタンパク質ペアひとつひとつを個別に考えるのではなく，複数の相互作用関係をネットワークとして捉えて理解することが重要となる．最近では，実際にタンパク質の相互作用ネットワーク（PPI ネットワーク）を実験的に同定しようと試みる研究も行われている [2], [3].

実験的な相互作用予測に関する研究が行われている一方，近年の計算機性能の向上や，タンパク質の立体構造情報の増加に伴い，立体構造情報を用いて計算機上で PPI 予測を

[1] 東京工業大学 大学院情報理工学研究科 計算工学専攻
Graduate School of Information Science and Engineering, Tokyo Institute of Technology
[2] 日本学術振興会 特別研究員
Research Fellow of the Japan Society for the Promotion of Science

Vol.2012-MPS-91 No.14
Vol.2012-BIO-32 No.14
2012/12/6

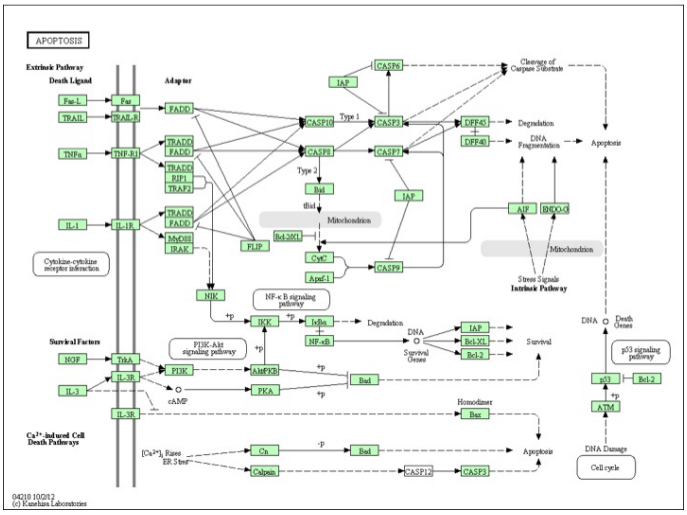

図 1 PPI ネットワークの例.（[1] より引用.）
Fig. 1 An example of a PPI network. (from [1].)

行う手法が開発されている [4], [5], [6], [7]．これらの手法は，立体構造情報から，タンパク質間ドッキング計算などにより複合体候補構造を作成し，得られた候補構造に対して何らかの処理を行い，相互作用するかどうかの判定を行うものである．これらは，1 対 1 のタンパク質間相互作用を予測することを目的としているが，前述の通り PPI は個別のタンパク質ペアごとに理解するよりも，PPI のネットワークとして理解することが重要である．

そこで，我々は大量のタンパク質を入力として，それらのタンパク質が構成する PPI ネットワークの予測を行うことを目的とする．PPI ネットワークの予測のためにはネットワークに含まれるタンパク質全ての間の網羅的な PPI 予測が必要となり，例えば 100 個のタンパク質が含まれるデータセットの場合，${}_{100}C_2 = 4950$ 通りの大量の相互作用予測を行う必要がある．この問題に対し，我々は高速なタンパク質間ドッキング計算に基づいた PPI 予測を行うソフトウェア MEGADOCK[8] の開発を行っており，現実的な計算時間での PPI のネットワーク予測を可能としている．

しかし，このシステムはあくまでネットワーク内の PPI を個別に予測しているだけであり，PPI ネットワークを直接予測してはいない．そこで，本研究では PPI ネットワーク予測の問題を単なる多数の PPI 予測問題の集合と捉えるのではなく，ネットワークそのものを予測する問題と捉え，予測に新たな情報を取り入れることを試みた．

本研究ではタンパク質の持つ相互作用に関する特異性に着目した．タンパク質は一部の例外を除いて，特定の相手と特異的に相互作用することが知られている．そのため大きな PPI ネットワークをマトリックスに置き換えた場合，そのマトリックスは疎になる性質を持つ（**図 2**）．この性質を新たに相互作用予測に利用することで，PPI ネットワーク予測の精度を向上させられると考えられる．

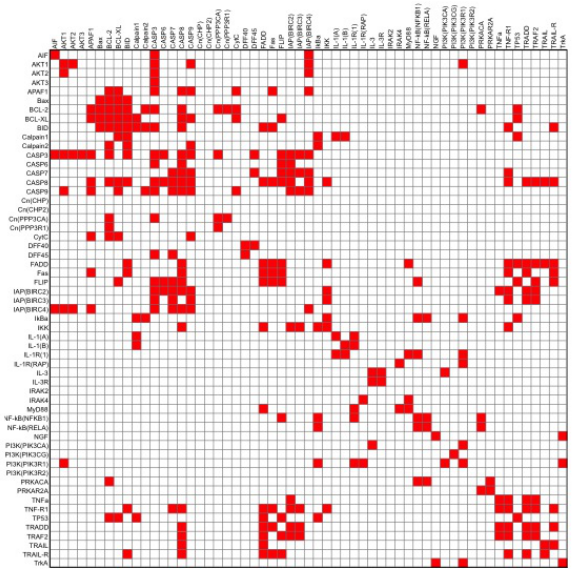

図 2 アポトーシスの PPI ネットワークをマトリックスに変換した例.
Fig. 2 An example of a PPI network matrix obtained from apotosis PPI network.

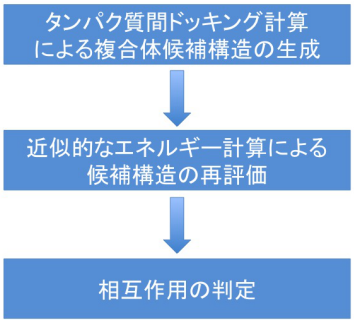

図 3 タンパク質間ドッキング計算を用いた PPI 予測手法の流れ.
Fig. 3 A flow of PPI prediction based on protein-protein docking calculation.

2. タンパク質間ドッキング計算を用いた PPI 予測手法

ここでは，我々がこれまでに開発したタンパク質間ドッキング計算を用いた PPI 予測手法 [8] について述べる．大まかな手順を**図 3** に示した．以後，図のそれぞれのステップについて順に説明していく．

2.1 タンパク質間ドッキング計算による複合体候補構造の生成

PPI 予測を行うには，まず予測したい 2 つのタンパク質単体構造を入力にとり，それらに対してドッキング計算を行い，多数（我々の手法ではしばしば 6000 個）の複合体候補構造群を生成する．ドッキング計算には MEGADOCK を用いることを想定しているが，ZDOCK[9], Hex[10] などの他のドッキングソフトウェアを利用することも可能である．

Vol.2012-MPS-91 No.14
Vol.2012-BIO-32 No.14
2012/12/6

2.2 近似的なエネルギー計算による複合体候補構造の再評価

ドッキング計算で生成された複合体候補構造に対し，より精密なエネルギー計算を行うことで複合体候補構造を再評価する．タンパク質間ドッキング計算から得た複合体候補構造は，タンパク質をボクセル化した “粗い” 計算によって生成されており，ボクセル上のドッキングスコアは高いものの，原子レベルの解像度ではエネルギー的には不安定な複合体候補構造が存在する．そのため，複合体候補構造群の中でエネルギー的により安定な構造を選び出す操作を行うことで PPI 予測精度の向上が期待できる．本手法ではエネルギー計算には ZRANK[11] を用いる．ZRANK は複合体の相互作用面の原子の情報から，近似的な複合体のエネルギー計算を高速に行うソフトウェアである．ドッキング計算で得た複合体候補構造それぞれに関して ZRANK によるエネルギースコア（ZRANK スコア）を計算する．

2.3 相互作用の判定

ZRANK スコアが最も高い複合体候補構造のドッキングスコアをもとに各タンパク質ペアについて相互作用を判定する評価値を計算する．評価値 E と相互作用するかどうかの判定は以下の式で与えられる．

$$E = \frac{S_1 - \mu}{\sigma}$$

$$\mathrm{PPI}(E) = \begin{cases} \text{True} & \text{if } E > E^* \\ \text{False} & \text{otherwise} \end{cases}$$

ただし，S_1 は ZRANK スコアトップの複合体候補構造のドッキングスコア，μ, σ はそれぞれ，全ての複合体候補構造のドッキングスコアの平均と標準偏差を表す．この評価値 E は，複合体候補構造群の中で，ZRANK スコアトップの構造のドッキングスコアが，他の構造のドッキングスコアと比べて、どの程度 “飛び抜けて良いか” を表している．E の値がある閾値 E^* を越えていた場合，そのタンパク質ペアは相互作用すると判定する．

3. PPI ネットワーク予測手法（提案手法）

3.1 従来手法の問題点

我々は PPI ネットワーク予測を行うことを目的としており，そのために入力タンパク質に対して網羅的な PPI 予測を行うことが必要となる．ここで，100 個のタンパク質を含む PPI ネットワークを予測することを考えた時，あるタンパク質が従来手法の評価値によって 100 個全てと相互作用すると評価され，別のタンパク質はただ一つのタンパク質と相互作用すると評価された場合，前者の予測結果には多数の偽陽性が含まれると考えられる．なぜなら，一般にタンパク質間相互作用は一部のタンパク質を除いて特異的であり，図 1 の例からもわかるように，大規模な PPI ネットワークにおいて，系に含まれるタンパク質全てと相互作用するようなタンパク質はまずあり得ないためである．したがって後者の例のように，特異的な相互作用を示すタンパク質ペアがより真の相互作用である可能性が高いと考えられる（**図 4**）．

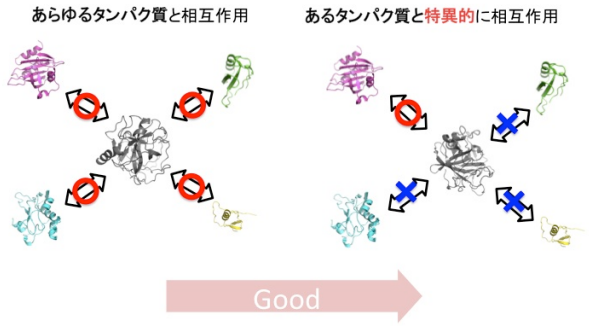

図 4 提案手法の考え方．非特異的な相互作用を示すタンパク質については閾値を厳しく評価する．

Fig. 4 Basic idea of our proposed method. The method trusts specific interactions shown in the right figure.

ところが，従来手法における網羅的な PPI 予測手法は，1 対 1 の予測を繰り返すことで実現されており，PPI ネットワークの持つ性質を十分考慮していない．例として従来手法を用いて，あるデータセットに対して網羅的な PPI 予測を行った結果を**図 5** 示す．図の各セルはタンパク質ペアひとつに対応し，対角線のセルは正例ペアに相当する．また赤く塗りつぶされているセルは “相互作用する” と判定されたペアである．図 5 では対角線以外で赤く塗りつぶされたセルが目立ち，従来手法による予測では，偽陽性が多く発生することがわかる．

そこで，我々はネットワーク予測に適した新しい相互作用の評価値を提案する．新しい評価値は “他のタンパク質ペアと比べて，対象のタンパク質ペアはどの程度特異的に相互作用しているか” についても評価するように設計されている．これは，先ほど述べた PPI ネットワークにおける相互作用の特異性を表現するものであり，この評価値によって予測精度の向上が期待される．

3.2 新しい相互作用評価値の計算

以下の式で各タンパク質ペアについて評価値（E_{network}）を計算する．

$$E_{\mathrm{network}} = \frac{E - \mu_{\mathrm{all}}}{\sigma_{\mathrm{all}}}$$

ただし，E は対象のペアの従来手法によって計算される相互作用評価値，$\mu_{\mathrm{all}}, \sigma_{\mathrm{all}}$ は，対象ペアのレセプター（またはリガンド）とそれ以外のリガンド（またはレセプター）とのペアから得られる従来手法の評価値全てから計算される平均と標準偏差である（**図 6**，**図 7**）．ここで得られた E_{network} の値をもとに，各ペアが相互作用するかどうか判定する．この評価値は，従来手法では様々なタンパク質相

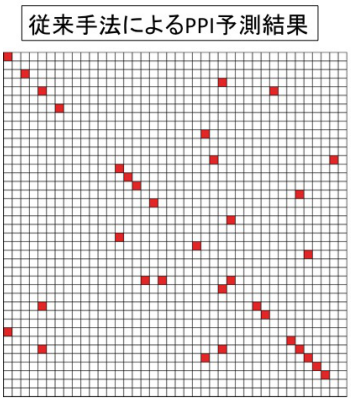

図 5 従来手法による網羅的 PPI 予測結果の例．各セルはタンパク質ペアひとつと対応しており，“正例” ペアは図の対角線に位置する．“相互作用” すると予測されたペアのセルは赤く塗りつぶされている．

Fig. 5 A result of all-to-all PPI prediction with the previous method. Each cell means each protein pair and diagonal cells are “true” interaction pairs. Colored cells are predicted as “positive”.

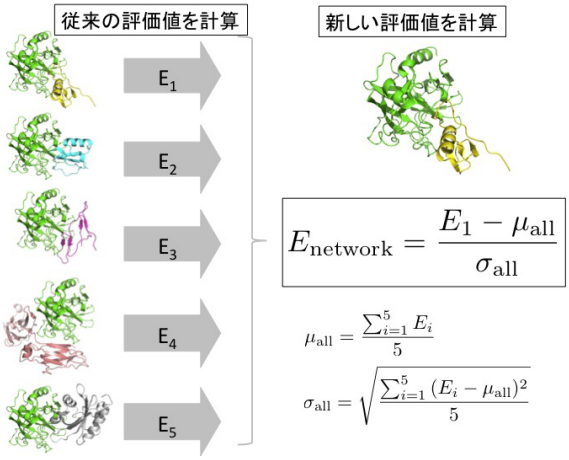

図 7 提案手法における相互作用評価値の計算方法．まず各ペアの従来手法における評価値を計算し，その結果を利用して新しい評価値を計算する．

Fig. 7 Calculation procedure of the proposed method. PPI evaluation values for each pair are calculated by the previous way, then new evaluation value is calculated based on those values.

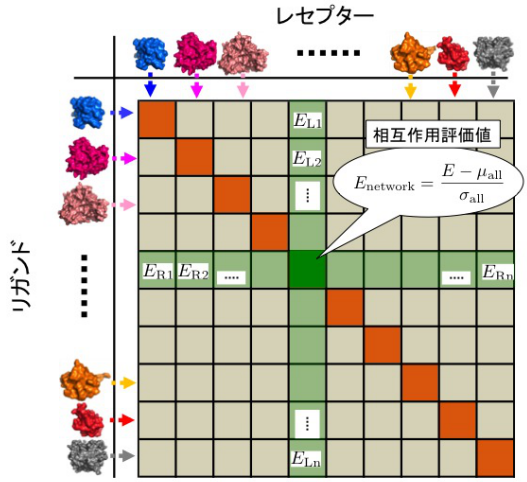

図 6 提案手法における相互作用評価値の求め方．あるペアの評価値を計算するために，レセプター（またはリガンド）と他のタンパク質間の従来手法の評価値を利用する．

Fig. 6 Calculation method for proposed evaluation value. The method uses the docking results of any pair including the receptor or ligand of a target pair.

手に相互作用すると判定されていたようなタンパク質について，正しく相互作用する相手だけに対して “相互作用する” と評価することが可能になると考えられる．

4. 実験

評価実験では，まずタンパク質間ドッキングベンチマークデータセットに対して，ドッキング計算を網羅的に行い，従来手法と提案手法それぞれによる予測精度を評価する．また，Matsuzaki らが利用した実データセット [4] に対して提案手法を用いて予測を行い，従来手法との間で予測精度の評価を行った．ベンチマークデータではじめに評価実験を行ったのは，立体構造情報が判明していてかつ，どのタンパク質同士が相互作用するかについての情報がある大規模な PPI ネットワークのデータが存在せず，提案手法の性能評価が難しかったためである．

4.1 ベンチマークデータセットの作成

提案手法の評価実験に用いるデータセットを以下の手順で作成した．

(1) 1 対 1 で相互作用しているタンパク質ペアについてのタンパク質間ドッキングベンチマークである，protein-protein docking benchmark4.0[12] の bound 構造群から，単鎖同士からなるヘテロな複合体ペア 120 個を選び出す．

(2) 120 個のタンパク質ペアから 40 個のペアをランダムに選び出す．

(3) 2 の操作を 10 回繰り返し，10 の異なるデータセットを作成する．

ここで，120 個のデータセットをそのままベンチマークとして用いないのは，データセット内のタンパク質の性質の偏りによる影響をなるべく排除するためである．また，特に記述がない限り，実験には bound ペアを用いた．精度を評価する際は，元の複合体構造が確認されたタンパク質ペアを “正例”，それ以外を “負例” として評価を行った．ただし，負例として扱うタンパク質ペア全てが必ず相互作用しないかどうかということについては確認していない．

4.2 ベンチマークを用いた評価実験

実験はそれぞれのデータセットについて，各タンパク質ペアをレセプターとリガンドに分けて入力タンパク質とし，先に述べた手順（図 3）にしたがって網羅的な PPI 予測を行い，従来手法と提案手法の間で予測精度の比較を行った．また，我々の PPI 予測手法はドッキングソフトウェアが変わっても同様の手順で予測ができるように設計されている．そこで手法がドッキングソフトウェアに依存しないことを確認する目的で，タンパク質間ドッキング計算に MEGADOCK を用いる場合の他に，ZDOCK[9] を用いた場合の結果も確認した．ZDOCK はタンパク質間ドッキング問題によく用いられるソフトウェアで，MEGADOCK と同じく剛体ドッキングを行うソフトウェアである．MEGADOCK と比べて計算速度は遅いものの，MEGADOCK にはない物理化学的性質を考慮したスコア関数を用いており，MEGADOCK よりもドッキング精度が良いという特徴をもつ．

評価指標には F 値（recall と precision の調和平均）を用いる．相互作用しているかどうか判定するための閾値 E^* は，各データセットごとに F 値を最大にする値を用いた．

4.3 実問題に対する評価実験

Matsuzaki らは大腸菌の走化性パスウェイのデータセットに対して PPI 予測を行っている [4]．本研究でもこのデータセットに対して評価実験を行い，提案手法の性能を評価を行った．ドッキング計算には MEGADOCK を用い，ベンチマークデータセットの評価と同じように，従来手法，提案手法の評価値それぞれに関して，F 値を最大にする E^* を最大にする値を用い，その時の予測結果について比較を行った．

5. 結果と考察

5.1 MEGADOCK を用いてドッキング計算を行った場合

作成した 10 のデータセットに対して MEGADOCK を利用して網羅的に PPI 予測を行い，それぞれの F 値の値を平均した値の比較を行った結果を**図 8** に示した．図 8 から，提案手法を用いることで予測精度が上昇していることがわかる．また，図は F 値の平均値で示しているが，10 のデータセットのうち 9 の場合において精度の上昇が確認された．

図 9 にあるデータセットの場合の予測結果の変化を示した．図中のセルひとつはタンパク質ペアひとつに相当し，対角線のペアが正例である．赤く塗りつぶされているセルは相互作用すると判定されたペアを表している．図 9 からは，提案手法は従来手法と比べて対角線以外にある赤いセルの数が減っていることがわかる．特に，従来手法で同じ行（または列）に偽陽性複数あるようなとき，提案手法ではそれらを正しく “相互作用しない” と評価できている．

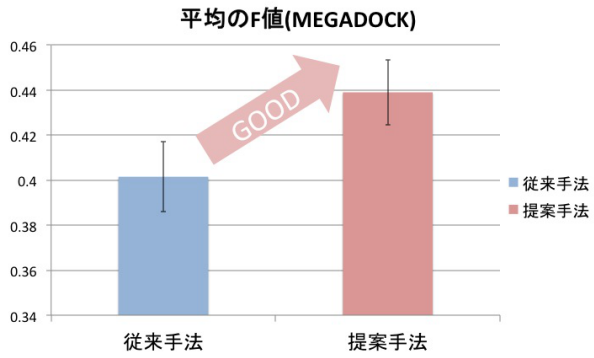

図 8 従来手法と提案手法の平均の F 値の比較（MEGADOCK）．
Fig. 8 The average f-measures for the previous method and the proposed one (with MEGADOCK).

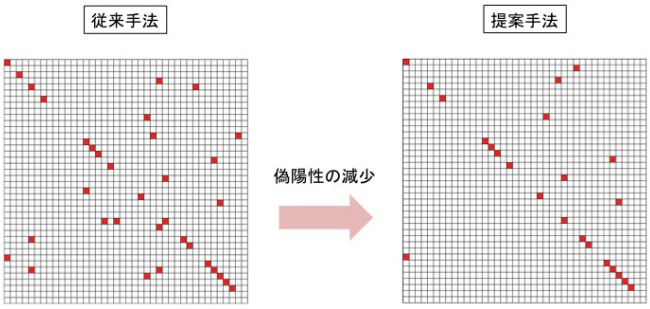

図 9 あるデータセットにおける従来手法と提案手法の予測結果．
Fig. 9 Comparison between the results of PPI prediction by using previous and proposed method.

同じ行（または列）に赤いセルが複数あることは，レセプター（またはリガンド）が複数の相手と相互作用すると判定されていることに相当しており，提案手法がそのようなペアを正しく判定できることは，提案手法が特異性のないタンパク質ペアを排除することができていることを意味する．また一方で，提案手法は新しい TP の数を増やすことにはあまり寄与しないこともわかる．

5.2 ZDOCK を用いてドッキング計算を行った場合

前述の通り，本手法はドッキングアルゴリズムそのものには依存しない設計になっている．そこで，MEGADOCK と同じベンチマークデータセットに対して ZDOCK を用いて PPI 予測を行った場合の結果を**図 10** に示す．ZDOCK を利用した場合，F 値が MEGADOCK を使用したときと比べて全体的に向上している．これは ZDOCK が MEGADOCK

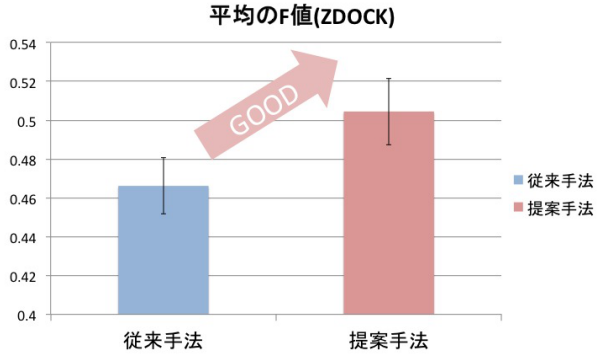

図 10 従来手法と提案手法の平均の F 値の比較（ZDOCK）
Fig. 10 The average f-measures for the previous method and the proposed one (with ZDOCK).

と比べて精密なドッキング計算をしており，得られる複合体候補構造がより正確であることが理由だと考えられ，ドッキング計算の正確性が相互作用予測精度の向上に重要であることを示唆している．しかし，ZDOCK の計算時間は MEGADOCK に比べて遅く，計算時間の短縮が重要となる網羅的な PPI 予測においては，予測精度だけでは ZDOCK の方が MEGADOCK よりも優れていると判断することはできない．また，F 値の大きさに差はあるものの，ドッキングソフトを変更した場合でも提案手法は予測精度の改善に成功してる．

5.3 大腸菌走化性パスウェイに対する予測結果

大腸菌の走化性パスウェイに対する予測結果を**図 11** に示す．従来手法は真陽性と偽陽性の個数が共に多く，提案手法は真陽性と偽陽性の個数が共に少なくなっているが，従来手法，提案手法の F 値はともに 0.462 であり，予測精度の向上は確認できなかった．これは，データセット中に含まれる “正例” のペアの割合が，提案手法が想定しているより多いためと考えられる．この割合はパスウェイ内に含まれるタンパク質の数が増加すれば通常減少するため，よりサイズの大きいデータセットで評価を行うと異なる結果が得られる可能性がある．

6. おわりに

本研究では PPI ネットワークの予測を想定し，複数のタンパク質ペアに対する予測結果を組み合わせて，PPI ネットワークにおける相互作用の特異性を表す新たな相互作用予測手法を提案し，PPI ネットワーク予測精度の向上を試みた．その結果，ベンチマークセットにおける PPI 予測精度の向上を確認した．提案手法は特に偽陽性の数を削減することに効果を発揮し，その効果はドッキングソフトウェアに依存しないという結果を得た一方で，提案手法は新たな真陽性の増加にはほとんど寄与しないこともわかった．また，実際の PPI ネットワークに本手法を適用した場合，偽陽性の個数が減少する一方で，真陽性の個数も減少するため精度の向上には至らなかった．ただし，これはネットワークに含まれるタンパク質の数が少ないことに起因している可能性があり，よりサイズの大きいデータセットでの実験が必要と考えられ，今後取り組むべき課題として挙げられる．

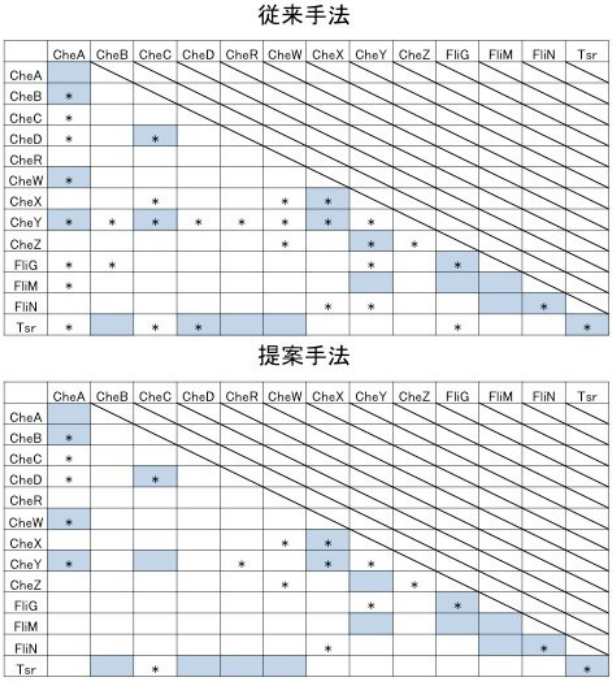

図 11 細菌走化性パスウェイに対する PPI ネットワーク予測結果．塗りつぶされているセルは正例ペアを表し，アスタリスクは “相互作用する” と予測されたペアを表す．
Fig. 11 Prediction results for the PPI network of E.coli chemotaxis pathway. Colored cells correspond to the “true” interaction. The cells marked by asterisks are predicted as “positive”.

謝辞

本研究を行う上で，実験データの提供頂いた，東京工業大学の松崎由理博士に感謝の意を表する．

参考文献

[1] http://www.genome.jp/kegg-bin/show_pathway?hsa04210
[2] Stelzl, U., Worm, U., Lalowski, M., *et al.*, A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6), 957–968, 2005.
[3] Rual, J., Venkatesan, T., Hao, T., *et al.*, Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062), 1173–1178, 2005.
[4] 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. *Journal of Bioinformatics and Com-*

putational Biology, 7(6), 991–1012, 2009.

[5] Wass, M.N., Fuentes, G., Pons, C., Pazos, F., Valencia, A., Towards the prediction of protein interaction partners using physical docking. *Molecular System Biology*, 7:469, 2011

[6] 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. *Nature Protocols*, 6, 1341–1354, 2011.

[7] Zhang, Q.C., Petrey, D., Deng, L., *et al.*, Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*, 490, 556–560, 2012.

[8] 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.

[9] Mintseris, J., Pierce, B., Wiehe, K., *et al.*, Integrating statistical pair potentials into protein complex prediction. *Proteins*, 69(3), 511–520, 2007.

[10] Ritchie, D.W., Venkatraman, V., Ultra-fast FFT protein docking on graphics processors *Bioinformatics*, 26(19), 2398–2405, 2010.

[11] Pierce, B., Weng, Z., ZRANK: Reranking Protein Docking Predictions with an Optimized Energy Function, *Proteins*, 67(4), 1078–1086, 2007.

[12] Hwang, H., Vreven, T., Janin, J., Weng, Z., Protein-protein docking benchmark version 4.0. *Proteins*, 78(15), 3111–3114, 2010.