

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

大 上 雅 史^{†1} 松 崎 由 理^{†1} 松 崎 裕 介^{†1}
佐 藤 智 之^{†2} 秋 山 泰^{†1}

タンパク質立体構造データ群から網羅的にタンパク質間相互作用 (PPI) の可能性を予測するために、我々は立体形状の相補性と物理化学的性質に基づくタンパク質ドッキングの手法を研究してきた。本研究のプロジェクトの一環として新たに開発した MEGADOCK システムは、高速なドッキング計算を行うための様々な工夫を取り入れており、中でも rPSC スコアと呼ぶスコア関数は、既存ツールの ZDOCK と比べて同等の精度を維持しながらも約 4 倍の速度向上を実現し、網羅的計算を現実のものとした。本論文では MEGADOCK システムの構成および計算モデルについて述べる。ベンチマークデータセットに適用した結果、従来手法を大きく上回る最大 F 値 0.415 を得た。さらにシステム生物学の典型的な問題の 1 つである細菌走化性シグナル伝達系のタンパク質群に MEGADOCK を応用した結果、既知の相互作用の再現をベンチマークデータと同等の精度 (F 値 0.436) で行うことに成功し、かつ生物学的に相互作用の可能性が高い組み合わせであるにも関わらず、現在までに報告されていない CheY タンパク質と CheD タンパク質の相互作用の可能性を示唆した。

**MEGADOCK: an all-to-all protein-protein interaction
prediction system using tertiary structure data
and its application to systems biology study**

MASAHITO OHUE,^{†1} YURI MATSUZAKI,^{†1}
YUSUKE MATSUZAKI,^{†1} TOSHIYUKI SATO^{†2}
and YUTAKA AKIYAMA^{†1}

An effective way to conduct exhaustive Protein-Protein Interaction (PPI) screening is one of the computational solutions for this problem. To predict all-

to-all PPI from protein structures, we have been studying the protein docking approach based on the physico-chemical properties and shape complementarity. To make these procedures that require huge number of protein dockings, we have developed high-speed protein-protein docking software “MEGADOCK” that reduces calculation time needed for docking by using several techniques that include a novel scoring function called rPSC. MEGADOCK was shown to be capable of exhaustive PPI screening by making docking calculation four times faster than conventional docking software, ZDOCK, while keeping almost the same level of accuracy in docking predictions. Here we describe an architecture and calculation model of MEGADOCK. We yielded F -measure value of 0.415 which substantially higher than previous studies when our PPI prediction system was applied to a general benchmark data. The same prediction system was applied to bacterial chemotaxis pathway, which is a typical basic problem in systems biology, and obtained the same level of accuracy (F -measure value of 0.436) with the benchmark dataset. We also found some interactions such as CheY–CheD among “False-Positives” that were worthy of further analysis.

1. はじめに

本研究では、生物学研究において近年重要視されているタンパク質間相互作用 (Protein-Protein Interaction: PPI) と呼ばれる現象を計算機で超高速に予測するため、タンパク質の形状相補性と静電相互作用のみに基づく単純化された評価モデルを提案し、フーリエ空間上での演算により計算量を大きく減じた上で、さらに大規模並列計算機上での効率的な並列計算が可能となるようにシステムの実装を行った。従来までは 1 対 1 のタンパク質間相互作用予測を行うだけでも数週間以上の計算時間を要しており、既に知られている相互作用の詳細な確認を行うのが計算機の役割と考えられていたが、本研究では、1 件ごとの予測の精度と信頼度にはやや制限があるものの、数十から数百個のタンパク質群における相互作用の可能性を網羅的に高速に予測することを初めて可能とした。

我々が開発したシステムは MEGADOCK と名付けられ、近日中に $1,000 \times 1,000$ (百万通り) の総当たり計算を実施することを目指して掲げている。本研究の成果は、たとえばガン細胞や微生物の細胞などを対象としたシステム生物学の研究において、システム内のタンパク質の制御関係を理解・予測するための新たな解析支援ツールとして広く利用される

^{†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} みずほ情報総研株式会社

Mizuho Information & Research Institute, Inc.

ことが期待される。

PPI とは、生体内のタンパク質分子が互いに結合することにより、機能の促進・抑制や、複合体形成による新たな機能獲得をする現象であり、ヒト細胞内では数万種類存在するといわれるタンパク質が相互にどのような制御関係にあるかを理解することは、病因の解明や薬剤の設計における重要な課題となっている¹⁾。

PPI 検出に関する生物学的実験手法としては、Yeast 2 Hybrid 法²⁾ や FRET 法³⁾ など数多くの手法が用いられており、ハイスループットだが結果の信頼度がやや低い手法と、精密だがコストのかかる手法が使い分けられている。我々が提案する計算に基づく相互作用予測の手法は、計算機の能力が限られていた時代においては実験手法に比べて価値の無いものであったと思われるが、数千から数万 CPU コアが比較的自由に使える近年の計算機システムを前提とすれば、むしろどのような実験よりも遙かにコストが小さく、立体構造が既知のタンパク質については、現在広く使われている配列相同性解析のごとくにまずは気軽に実施されるべき基本的なスクリーニング手法になると我々は考える。

開発した MEGADOCK は、タンパク質のペアに対してそれぞれの立体構造情報を利用してタンパク質ドッキング計算を行い、その結果に基づいて相互作用の有無を予測するシステムである(図 1)。詳しくは次章以降に述べるが、複数のタンパク質群同士の網羅的な相互作用予測を目指して、計算時間の節約のための様々な手法を提案しており、計算の高速性に特に力点を置いたシステムである。本論文では MEGADOCK の構成および相互作用の計算モデルについて述べ、高速化のための種々の工夫点について論じ、実際に相互作用の有無が実験によって確かめられているベンチマークデータセットに本手法を適用し、従来手法に比べて精度が向上したことを示す。さらにシステム生物学における典型的な例題の一つとして細菌走化性系を取り上げ、ベンチマークデータに適用したものと同等のパラメータを使って、網羅的 PPI 予測が実用的に実施できたことを示す。

2. MEGADOCK: ドッキング計算システム部

MEGADOCK の中核をなす「ドッキング計算システム部」の処理は、形状の相補性に関する項 G と静電的相互作用の項 E の計算からなる。ここで、対象とするタンパク質ペアについて片方をレセプター R 、もう片方をリガンド L と呼ぶことにする。それぞれのタンパク質を 1 辺 $1.2\text{\AA}$ のボクセル空間上に表し、タンパク質の内部か表面かなどの種別によって、各ボクセル上に異なるスコアの値を代入する。形状の相補性の項 G には、我々が提案する real Pairwise Shape Complementarity(rPSC) スコアを用いる。rPSC スコアは以下

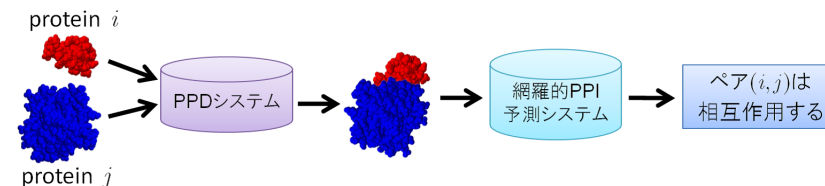


図 1 タンパク質立体構造情報を利用した、タンパク質ドッキング (PPD) による網羅的 PPI 予測
Fig. 1 All-to-all protein-protein interaction prediction using protein tertiary structure information by protein-protein docking (PPD)

のように表される⁴⁾。

$$G_R(l, m, n) = \begin{cases} \# \text{ of R atoms within } (3.6\text{\AA} + R \text{ atom } r_{vdW}) \text{ open space} \\ -27 \text{ inside of the R} \end{cases} \quad (1)$$

$$G_L(l, m, n) = \begin{cases} 0 & \text{solvent accessible surface layer of the L} \\ 1 & \text{solvent excluding surface layer of the L} \\ 2 & \text{core of the L} \\ 0 & \text{open space} \end{cases} \quad (2)$$

ただし、 r_{vdW} は原子のファンデルワールス半径を表す。それぞれの数値はいくつかのタンパク質複合体による事前実験によって最適化したものである。また、静電的相互作用による項 E については、CHARMM19⁵⁾ と FTDock の Coulomb ポテンシャルモデル⁶⁾ を組み合わせたものを利用している。以上を用いて、ドッキング予測の良さを表す評価値であるドッキングスコア S を、

$$R(l, m, n) = G_R(l, m, n) + iE_R(l, m, n) \quad (3)$$

$$L(l, m, n) = G_L(l, m, n) + iwE_L(l, m, n) \quad (4)$$

$$S(\alpha, \beta, \gamma) = \Re \left[\sum_{l=1}^N \sum_{m=1}^N \sum_{n=1}^N R(l, m, n) L(l + \alpha, m + \beta, n + \gamma) \right] \quad (5)$$

と定義する。 (α, β, γ) はリガンドの平行移動ベクトルである。MEGADOCK ではリガンドを回転・平行移動させながら全空間におけるスコアの値を畳み込み和として計算する。リガンド回転角の刻み幅は 15° とし、ある回転角に対してボクセルサイズ N に対して $N \times N \times N$

通りの平行移動を行う．その中から最も良いドッキングスコアを持つリガンドの平行移動ベクトルを，その角度におけるドッキング結果として返す．回転角のサンプリングは文献 7) に基づいており，計 3,600 通りの回転パターンで計算を行う．よって，1 つの複合体について計算されるドッキング結合部位は，ボクセルサイズが N のとき， $3,600 \times N^3$ 通りとなる．ただし角度ごとの候補構造のサンプリング数については，同様のドッキングソフトウェアである ZDOCK⁸⁾⁻¹¹⁾ では，評価実験の結果から 1 つのみ出力すれば良い（即ち出力数は 3,600 個）という結論を得ているが，新たな PPI の予測に適用する場合，構造サンプリング数を増加させることで精度を向上させられる可能性があるため，MEGADOCK では t 個の候補構造を出力可能とした．

なお，計算時間は単純に畳み込み和をとると $O(N^6)$ だが，離散フーリエ変換 (DFT) と逆離散フーリエ変換 (IFT) を用いて，

$$S(\alpha, \beta, \gamma) = \text{IFT}[\text{DFT}[\mathbf{R}(l, m, n)]^* \text{DFT}[\mathbf{L}(l, m, n)]] \quad (6)$$

とし，高速フーリエ変換 (FFT) を用いることで $O(N^3 \log N)$ に削減することが可能となる． z^* は z の複素共役を表す．rPSC の利点はここで生かされており，ZDOCK などが用いている複素数による表現モデルに比べて離散関数の虚数部に空きができる分，他の物理化学的相互作用の導入が可能となり，かつ FFT の計算回数の増加を回避し，計算の高速化を図っている．その結果，ZDOCK では FFT を 1 サンプリング角度ごとに 3 回以上行ってドッキングスコアを計算していたものを，MEGADOCK では FFT の回数を 1 回に抑えられており，結果として ZDOCK の約 4 倍の計算速度向上を実現している．

3. MEGADOCK：タンパク質間相互作用 (PPI) 予測システム部

本論文で提案する MEGADOCK における PPI 予測システム部は，ZRANK によるリランキング，クラスタリング，PPI 判定に分かれている．以下では各項目について説明する．

3.1 ZRANK リランキング

ZDOCK や MEGADOCK が探索を行うドッキング姿勢の組み合わせは $3,600 \times N^3$ 通り存在し，中にはドッキングスコアが悪くても実際には結合エネルギー値が高く現実的ではない複合体予測も多数存在する．そのような予測構造を正しく取り除くための方法としては，理想的には， $3,600 \times N^3$ 通りの全構造に対して厳密な結合エネルギーを計算することが行われるのが望ましいが，計算時間の観点から候補構造を絞り込んだ後に，やや厳密な計算を行うアプローチが有望である．

ZDOCK に対しては Pierce らが開発したリランキングシステム ZRANK¹²⁾ によって解

決を図っている．ZRANK は ZDOCK によって予測された複合体構造群に対して，ファン・デル・ワールスエネルギー，静電的相互作用エネルギー，脱溶媒和エネルギー^{*1}を，限られた範囲の原子間で計算した擬似的な相互作用エネルギーに基づいて，高速に評価をするものである．ZRANK により計算されたエネルギー値に基づいて複合体をリランキングすることで，より精度の高い順位付けに変更することができ，ある程度上位に正解に近い構造が集まるようになると考えられる．特に，後述するクラスタリングを行う際には，上位に正解構造を集めることで，クラスタリングにかかる予測複合体数の削減による計算時間の削減と，ノイズ（結合エネルギーの高い非現実的な複合体構造）が除去されることによる精度の向上が期待できる．

3.2 クラスタリング

Matsuzaki ら^{14),15)} は膨大なドッキング構造群に対して，クラスタリングによって類似性の高い構造を統合し，解析対象とする複合体予測構造を絞り込み，ドッキングおよび相互作用予測の精度を向上させる後処理システムを開発した．本手法では，文献 15) で提案されているクラスタリング手法を用いる．

3.3 PPI 評価値の計算と PPI 判定

クラスタリング結果に基づいて以下の方法でタンパク質ペアに関する評価値を決定する．なお，この手法は Matsuzaki ら¹⁶⁾ が ZDOCK を用いて相互作用予測を行った際に用いた手法を参考にし，若干の変更を加えたものである．

- (1) 各クラスタ C_i 中のデータのうち，ZRANK による評価順位が最も良いものを代表データとする．代表データについて，リランキングされた上位 1,000 個のデータのドッキングスコアに対する z 値を s_i とする．
- (2) 各クラスタ C_i のメンバ数 $|C_i|$ について調べ， N 個のクラスタ中での各メンバ数を母集団として計算した z 値を m_i とする．
- (3) m_i が閾値 m^* より大きいクラスタの集合を C' とする． C' に含まれるクラスタの代表値 s_i のうち，最大値を評価値 E とする．

$$C' = \{C_i \mid m_i \geq m^*\} \quad (7)$$

$$E = \begin{cases} \max s_i \ (i \in C') & \text{if } C' \neq \emptyset \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

*1 ACE(Atomic Contact Energy) スコア¹³⁾ を用いている．詳しくは文献 12) を参照されたい．

表 1 Protein-protein docking benchmark 2.0 から選出した 44 複合体によるサブセット
Table 1 The selected 44 complex structures from the Protein-protein docking benchmark 2.0

Complex PDB ID										
1ACB	1AK4	1ATN	1AVX	1AY7	1B6C	1BUH	1BVN	1CGI	1D6R	1DFJ
1E6E	1E96	1EAW	1EWY	1F34	1FC2	1FQ1	1FQJ	1GCQ	1GHQ	1GRN
1H1V	1HE1	1HE8	1I2M	1IBR	1KAC	1KTZ	1KXP	1KXQ	1M10	1MAH
1PPE	1QA9	1SBB	1TMQ	1UDI	1WQ1	2BTF	2PCC	2SIC	2SNI	7CEI

以上のようにして全タンパク質ペアについて評価値 E を決定したあと、 E が閾値 E^* 以上となるタンパク質ペアを「相互作用する」と判定する。

4. ベンチマークデータセットによる評価実験

タンパク質ドッキング予測に広く用いられるベンチマークである protein-protein docking benchmark 2.0¹⁷⁾ に含まれる 84 複合体のうち、タンパク質複合体がモノマー同士で形成されている 44 個の複合体によるサブセットの全要素に対して 1 対 1 のタンパク質ドッキング予測を行った。対象としたタンパク質の PDB^{18),19)} ID リストを表 1 に示す。

4.1 ドッキング予測の性能比較

4.1.1 方法

MEGADOCK 2.1 で 1 対 1 のドッキングを行い、予測複合体の結合リガンドタンパク質と結晶構造の結合リガンドタンパク質との構造差 (L-RMSD) を計算する。L-RMSD が 5 Å 以下であるドッキング構造を正解構造と定義し、ドッキング予測順位が 1 位の構造が正解構造であった場合、そのペアのドッキングが正解したと定義する。この正解の複合体数を ZDOCK 3.0¹¹⁾ と比較し評価する。なお全ての実験において、東京工業大学学術国際センターのスーパーコンピュータ “TSUBAME 1.2” を使用した。計算時間の計測値は 1CPU コアのみを用いた場合のものを示す。

4.1.2 結果と考察

MEGADOCK 2.1 と ZDOCK 3.0 で 44 複合体中正解となった個数と計算時間を表 2 に示す。ただし、MEGADOCK の $t = 1 \sim 20$ では ZRANK の計算時間も含まれる。表 2 より、MEGADOCK 2.1 は (ZRANK を併用しない場合でも) ZDOCK 3.0 と同等の正解数を出しつつ計算時間を約 4 倍高速化することに成功していることが分かる。ZRANK を併用した場合は、 t の値の増加にしたがって計算時間が増加するが、正解数も増える。例えば $t = 15$ では ZDOCK 3.0 の約半分の計算時間で良好な結果を得ている。しかしながら、本

表 2 ドッキング予測の正解数と計算時間

Table 2 The number of near native structures and calculation time for docking predictions

t	MEGADOCK								ZRANK なし	ZDOCK
	1	2	3	4	5	10	15	20		
正解数	22	23	24	24	24	24	25	25	18	19
計算時間 [分]	61	65	69	72	77	97	115	137	57	230

研究の対象とする網羅的 PPI 予測への適用を考えると、計算時間はできるだけ小さい方がよい。よって、計算時間との兼ね合いも考えて $t = 3$ が最も良いと判断した。

4.2 ベンチマークデータセットを用いた網羅的 PPI 予測実験

4.2.1 方法

表 1 の 44 個のサブセットに対して $44 \times 44 = 1,936$ 通りの網羅的 PPI 予測を行った。MEGADOCK の各角度ごとの複合体予測出力数 t を $t = 1, 2, 3, 5, 10, 20$ とし、全予測複合体数を $2,000 \times t$ 個として 44×44 通りの組合せに対してドッキング計算を行った結果を用意し、PPI 予測フローにしたがって相互作用予測を全 1,936 通りについて行った。PDB に複合体が収録されている組み合わせのペアを正例、そうでない組み合わせのペアを負例としたときの、True Positive (TP)、False Positive (FP)、True Negative (TN)、False Negative (FN) を求め、

$$F \text{ 値} = \frac{2 \cdot TP}{(TP + FP) + (TP + FN)} \quad (9)$$

によって F 値を計算し、相互作用予測性能の評価に用いる。

4.2.2 結果と考察

網羅的 PPI 予測を行った結果、 $t = 3$ のときに ZRANK を併用した場合に最大 F 値 0.415 を得た。ただしこのときの PPI 判定パラメータは $E^* = 5.5$ である。これは我々が以前に記録した閾値による相互作用判別法の F 値 0.150⁴⁾ や、機械学習で生成した判別器による F 値 0.251²⁰⁾ を大きく上回っており、網羅的な PPI 予測の精度の向上に成功していると言える。また、全てのデータ ($t = 1, 2, 3, 5, 10, 20$) に共通して、ZRANK によるリランキングを実施した場合の F 値が向上した。これより、エネルギー計算によるリランキングが PPI 予測に効果を示すことが分かる。

関連研究として、protein-protein docking benchmark 2.0 を対象に網羅的 PPI 予測を行ったものに、Yoshikawa らの AEP²¹⁾ や RFZ/LFZ²²⁾ とよばれる手法、Matsuzaki らによる手法¹⁶⁾ がある。それぞれ使用したドッキングソフトウェアや対象としているデータセッ

トの大きさが異なるが、Matsuzaki らは文献 16) 上で AEP よりも精度が良くなったことを報告しており、我々が扱ったものと同じ大きさである 44×44 のデータセットでは F 値で 0.43 を得ている。Matsuzaki らの手法で用いられたドッキングソフトウェアは ZDOCK であり、我々の手法の結果と比較すると、ドッキングの計算時間を短縮しつつ同等の精度が得られていることから、我々の手法が優れていることが分かる。また Yoshikawa らの RFZ/LFZ は、あらかじめタンパク質の機能情報を用いて複数のサブセットに分割してから網羅的 PPI 予測を行う方法であり、ドッキングには ZDOCK を用いている。RFZ/LFZ での精度は最も良いサブセット (12×12) で F 値 0.471 であり、我々の手法よりも若干良い精度を得ているが、Yoshikawa らのデータセットは 144 通りの組み合わせに対して我々の扱ったデータセットは 1,936 通りの組み合わせが存在し、10 倍以上の大きさの問題であるにも関わらず近い精度を得ていることから、我々の手法が優れていると考える。なお、あらかじめタンパク質の機能情報を用いて問題を分割する方法は、我々の想定するあらゆるタンパク質群に適用可能な初期スクリーニング手法としては適切ではないと考えているが、我々の手法にタンパク質の機能情報を同様に追加することで、ある程度予測精度を改善することは可能であると考えられる。

$t = 3$ の ROC 曲線を図 2 に示す。 m^* は 2.5 節に示してある通りクラスタのメンバ数の z 値に対する閾値である。図 2 より、 $m^* = 0.0$ (平均値) を閾値とするときが最も精度が良く、 m^* の値を増加させると精度が悪化することがわかる。また $m^* = 0.0$ ではランダムな相互作用予測 (対角線) よりも有意に精度が優れていると言える。

5. システム生物学への応用 ~ 細菌走化性系を例にして

ここまで述べた PPI 予測システムの実用的な性能を示す例として、システム生物学における典型的な問題であり、Matsuzaki らが文献 16) で取り上げている、細菌走化性系のシグナル伝達パスウェイの予測を試みた。

生物が自らの生存に関わる環境変化に適切に対応することは、生存のために重要な機能である。外界からの刺激に応答して運動が起こる性質を走性と呼ぶ。例えば大腸菌は飢餓状態になるとべん毛の遺伝子を発現し、栄養となる物質の多い方へ移動する化学走性 (走化性) を示す。この性質を実現しているのは、刺激に応じて系の内部状態を変える細胞内のシグナル伝達系である。外界の状態を感知して内部状態を変えつつ対応していくという行動は、学習という概念にもつながる興味深い現象である。また、走化性系によって制御されるべん毛モーターはきわめてエネルギー効率の良い分子モーターとしても注目され、機械分野

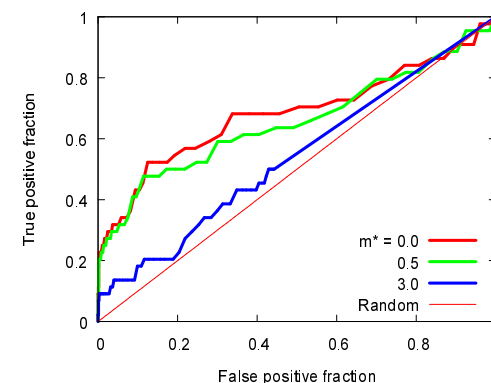


図 2 本システムのベンチマークデータセットへの適用における精度 ($t = 3$, ZRANK あり)。 m^* と E^* を変化させて行った実験における ROC 曲線を示す。縦軸が True positive fraction ($\frac{TP}{TP+FN}$) を、横軸が False positive fraction ($\frac{FP}{FP+TN}$) を示す。予測がランダムに行われた場合は斜線のようになる。

Fig. 2 Evaluation of the MEGADOCK system ($t = 3$, ZRANK used) when applied to the benchmark data. The ROC curves for varying the threshold m^* and E^* values are shown. X-axis is for the false positive fraction ($\frac{FP}{FP+TN}$) and y-axis is for the true positive fraction ($\frac{TP}{TP+FN}$). Random prediction is indicated by the diagonal.

への応用として走化性系のシミュレーションを行った研究も知られている²³⁾。

細菌の走化性を実現する分子の相互作用関係については、そのほとんどが明らかになっている^{24),25)} (図 3)。一方、系全体の示す行動の動的かつ定量的な解析においてはいくつかの未解明の問題が残されている。刺激に対する感受性や、一定時間同じ刺激レベルが続くと元の行動に戻る順応性などを細菌が実現する仕組みを説明するために、分子イメージング技術や動的モデルを用いた研究が行われている^{26)–28)}。そこで本研究では、この系の既知の PPI を「正解」と定義して MEGADOCK の評価を行うと共に、この系の制御に関わる未知の相互作用の検出を試みた。

5.1 方 法

細菌走化性系のタンパク質について、構造データを多く得られる 3 種 (*E. coli*, *S. typhimurium*, *T. maritima*) をターゲットとした。まず、KEGG²⁹⁾ 走化性系パスウェイデータ (KEGG pathway ID: *eco02030*, *stm020230*, *tma02030*) を参照し、パスウェイに存在するタンパク質データにリンクされた構造データを取得した。具体的には、パスウェイ中のタンパク質と同等の UniProt データより、LinkDB³⁰⁾ を用いて PDB の ID を得た。こ

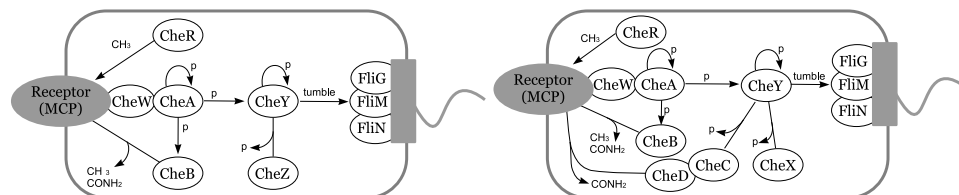


図 3 *E. coli* (左), *T. maritima* (右) の走化性シグナル伝達系。これらの細菌の運動はべん毛モーターの回転調節により制御される。べん毛モーターは CheY によって回転を制御されている。

Fig.3 Chemotaxis pathway for *E. coli* (left) and *T. maritima* (right). The motion of these bacteria are controlled by the rotation direction of their flagellar motor. CheY is responsible for the rotation direction.

これらの PDB ファイルのうち, (i) X-ray diffraction により結晶構造が決定されていること, (ii) 3.25Å より高解像度であること, (iii) アミノ酸残基数 30 以上であること (文献 31) と同様の条件) を実験対象の候補として選択した。PDB ファイルに複数のタンパク質の構造データがある場合には, PDB ファイルをタンパク質ごとに分割した。そのうち (i) Mutant のデータと (ii) 人工的に合成されたペプチド鎖のデータを対象から除外した。今回は細胞質での PPI 予測を目的としたので, ペリプラズムに位置する化学受容体のリガンド結合ドメインは除外した。ただし, 以上の条件をすべて満たす構造データが得られないタンパク質は, 変異の入ったタンパク質構造を例外的に使用した。

このように収集した 13 種のタンパク質に対応する 89 の構造データを用いて, $89 \times 89 = 7,921$ 通りのドッキングと相互作用予測を行った。走化性系データセットでは, 1 つのタンパク質について複数の構造データが存在する場合がある。ここでは, タンパク質の種類が重複する組み合わせについても全て評価値を計算し, 同等のタンパク質対の評価値のうち 1 つでも陽性と判定されるものがあれば相互作用可能と判定した。

5.2 結果と考察

図 4 に, 走化性系データへの適用における PPI 検出性能を示す。表 3, 図 5 には, クラスタリングを行わず, E^* を 5.5 として, 本システムを細菌走化性系のデータセットに適用して得た相互作用を示す。これらのパラメータは, ベンチマークデータへの適用で F 値が最大になった値である。実際に相互作用が確認されている組み合わせを灰色のセルで示した。得られた F 値は 0.436 であり, ベンチマークデータでの網羅的 PPI 予測の精度と同等の精度で予測を行うことができているといえる。また, Matsuzaki らの ZDOCK による相互作用予測では F 値 0.45 を得たと報告されており¹⁶⁾, 我々の手法が同等の精度で予測可

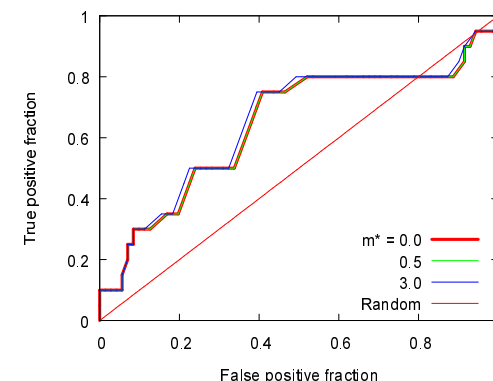


図 4 本システムの走化性系への適用における精度。 m^* と E^* を変化させて行った実験における ROC 曲線を示す。
Fig.4 Evaluation of the MEGADOCK system when applied to the chemotaxis pathway data. The ROC curves for varying the threshold m^* and E^* values are shown.

能であることを示している。

我々のシステムが検出した未知相互作用の 1 つに, CheY-CheD の組み合わせがある。CheC は, CheY の脱リン酸化酵素であるが, CheD の存在によりその活性が高まることが知られている^{32),33)}。CheY-CheC, CheC-CheD の相互作用は実験的に確認されている^{33),34)}。CheY-CheD が直接相互作用するという報告はまだない。CheY と CheC の相互作用を CheD が助長するならば, CheD が CheC と CheY の結合においてなんらかの機能を持つ可能性がある。

図 6 に, 予測された CheY-CheD 複合体の構造と, CheY-CheD の予測複合体に CheC をドッキングした結果を示す。ここに示した構造では, CheC の活性部位と CheY のリン酸化サイトは近くに位置していなかった。最終的に相互作用の有無を確認するには, 実験による解析が必要となる。CheY-CheC-CheD の三者の複合体構造を実験で解析し, 予測された複合体構造と比較することができれば興味深い。

表 3 予測されたタンパク質間相互作用. $E^* = 5.5$ とした際の結果を示す. 本システムが予測した相互作用を * 印で示す. セルが灰色の部分は既知の相互作用である.

Table 3 Results of the PPI predictions using the proposed system. The gray colored cells correspond to the known interactions.

	A	B	C	D	R	W	X	Y	Z	FliG	FliM	FliN	Tsr
CheA		-	-	-	-	-	-	-	-	-	-	-	-
CheB	*		-	-	-	-	-	-	-	-	-	-	-
CheC	*			-	-	-	-	-	-	-	-	-	-
CheD	*		*		-	-	-	-	-	-	-	-	-
CheR						-	-	-	-	-	-	-	-
CheW	*	*			*		-	-	-	-	-	-	-
CheX			*			*	*	-	-	-	-	-	-
CheY	*	*	*	*	*	*	*	*	-	-	-	-	-
CheZ						*	*	*	*	-	-	-	-
FliG	*	*						*		*	-	-	-
FliM	*							*		*	-	-	-
FliN							*	*			*	-	-
Tsr	*		*	*						*			*

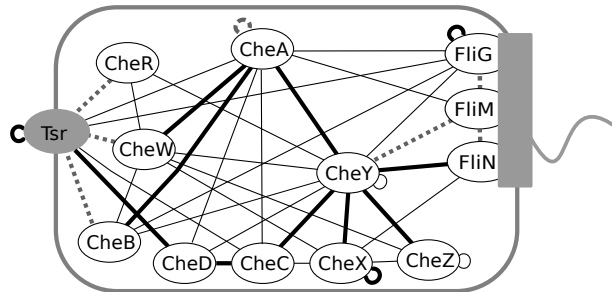


図 5 本システムにより予測されたタンパク質間相互作用. $E^* = 5.5$ とした際の結果を示す. 黒い太線は正しく予測された相互作用 (True Positives) を, 灰色の点線は既知の相互作用のうち検出できなかったもの (False Negatives) を, 細い線は予測されたがこれまでに報告されていない相互作用 (False Positives) を示す.

Fig. 5 Results of the PPI predictions from the proposed system with $E^* = 5.5$. The bold lines (True Positives), gray dashed lines (False Negatives) and thin lines (False Positives) representing the predicted or known PPIs show the relevance of the predictions.

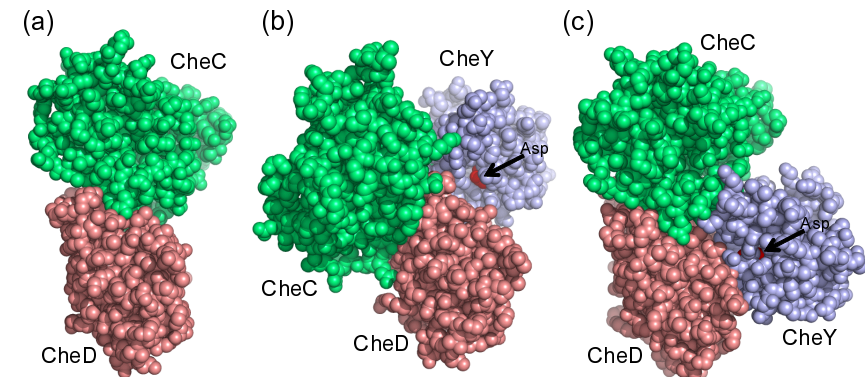


図 6 (a) 既知の CheC-CheD 複合体の結晶構造 (PDB ID: 2F9Z, chain a, c). (b) 予測された CheY (PDB ID: 1A0O, chain c)-CheD (PDB ID: 2F9Z, chain c) の複合体をドッキングのレセプターとして, CheC (PDB ID: 1XKR, chain a) をドッキングさせた時の予測. CheY のリン酸化部位を赤色で示す. (c) CheC-CheD 複合体の結晶構造 (2F9Z) をドッキングのレセプターとして, CheY (PDB ID: 1F4V, chain c) をドッキングさせた結果. CheY のリン酸化部位を赤色で示す.

Fig. 6 (a) Known structure of the CheC-CheD complex (PDB ID: 2F9Z, chains a, c). (b) Docking of CheY (PDB ID: 1A0O, chain c)-CheD (PDB ID: 2F9Z, chain c) hypothetical complex and CheC (PDB ID: 1XKR, chain a). The phosphorylation site of CheY is colored red. (c) Docking of a known structure of the CheC-CheD complex (PDB ID: 2F9Z, chains a, c) and CheY (PDB ID: 1F4V, chain c). The phosphorylation site of CheY is colored red. This hypothetical complex is also constructed using the representative data among all combinations of the CheC-CheD complexes and CheY structures.

6. おわりに

本研究では、タンパク質立体構造情報からのドッキング予測を利用した網羅的 PPI 予測に対して、MEGADOCK 2.1 システムの提案を行い、ベンチマークデータセットに適用してその精度を確認した。結果は最大 F 値が 0.415 となり、従来手法よりも大幅に高い精度での相互作用予測が可能になったことを示した。さらにシステム生物学への応用として、細菌走化性シグナル伝達系のタンパク質群に MEGADOCK 2.1 システムを適用した。現在実験的に知られている相互作用の再現と未知相互作用の可能性の検出を目的とした実験を行い、結果としてベンチマークデータと同等の F 値である 0.436 を得た。また、False Positive の中で生物学的実験結果から相互作用の可能性が高いと考えられるものとして CheY-CheD の組み合わせを得た。このタンパク質ペアに対するさらなる調査を、実験生物学者と協力して行っていく予定である。

また、現在我々は肺ガンと深く関わりがあるとされ医学的に重要なヒト EGFR シグナル伝達系^{35),36)}を対象とした、 500×500 規模の網羅的 PPI 予測に取り組んでいる。MEGADOCK は大規模並列計算機との相性が良く、数千個の CPU コアを持つような先端的な計算環境を用いれば、システム生物学における重要な系の解析を数日で実施することが可能である。

今後は、既知の配列情報などとの融合や、予測結果の信頼度をより数量的に示すことなどが課題である。

謝辞 本研究は、文部科学省 最先端・高性能汎用スーパーコンピュータの開発利用「次世代生命体統合シミュレーションソフトウェアの研究開発」、および科学研究費補助金（基盤研究 (B) 19300102）の支援を受けて行われたものである。

参 考 文 献

- 1) Ravasi, T., Suzuki, H., Cannistraci, C.V., et al.: An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man, *Cell*, Vol.140, pp.744-752 (2010)
- 2) Fields, S. and Song, O.: A novel genetic system to detect protein-protein interactions, *Nature*, Vol.340, pp.245-246 (1989)
- 3) Forster T.: Zwischenmolekulare Energiewanderung und Fluoreszenz, *Ann. Physik* Vol.437, pp.55-75 (1948)
- 4) 大上雅史, 松崎裕介, 松崎由理, 佐藤智之, 秋山泰: 物理化学的相互作用の導入による網羅的タンパク質間相互作用予測システムの高精度化, 情報処理学会研究報告, Vol.2009-BIO-17, No.11, pp.1-8 (2009)
- 5) Brooks, B.R., Brucoleri, R.E., Olafson, B.D., et al.: CHARMM: a program for

- macromolecular energy, minimization, and dynamics calculations, *J Comput Chem*, Vol.4, pp.187-217 (1983)
- 6) Gabb, H.A., Jackson, R.M. and Sternberg, M.J.E.: Modelling Protein Docking using Shape Complementarity, Electrostatics and Biochemical Information, *J Mol Biol*, Vol.272, pp.106-120 (1997)
- 7) Lattman, E.E.: Optimal sampling of the rotation function, *Acta Crystallographica Section B*, Vol.28, pp.1065-1068 (1972)
- 8) Chen, R. and Weng, Z.: Docking Unbound Proteins Using Shape Complementarity, Desolvation, and Electrostatics, *Proteins*, Vol.47, pp.281-294 (2002)
- 9) Chen, R. and Weng, Z.: A Novel Shape Complementarity Scoring Function for Protein-Protein Docking, *Proteins*, Vol.51, pp.397-408 (2003)
- 10) Chen, R., Li, L. and Weng, Z.: ZDOCK: An Initial-stage Protein-Docking Algorithm, *Proteins*, Vol.52, pp.80-87 (2003)
- 11) Mintseris, J., Pierce, B., Wiehe, K., Anderson, R., Chen, R. and Weng, Z.: "Integrating Statistical Pair Potentials into Protein Complex Prediction", *Proteins*, Vol.69, pp.511-520 (2007)
- 12) Pierce, B. and Weng, Z.: ZRANK: Reranking Protein Docking Predictions with an Optimized Energy Function, *Proteins*, Vol.67, No.4, pp.1078-1086 (2007)
- 13) Zhang, C., Vasmatazis, G., Cornette, J.L. and DeLisi, C.: Determination of atomic desolvation energies from the structures of crystallized proteins, *J. Mol. Biol.*, Vol.267, pp.707-726 (1997)
- 14) Matsuzaki, Y., Matsuzaki, Y., Sato, T. and Akiyama, Y.: Development of post-docking system for protein-protein interaction prediction, *1st Joint Workshop on Computational Science* (2008)
- 15) Matsuzaki, Y., Matsuzaki, Y., Sato, T. and Akiyama, Y.: Development of post-docking system for protein-protein interaction prediction, *IPSI-SIG Technical Report*, Vol.2008-BIO-13, No.5, pp.17-20 (2008)
- 16) Matsuzaki, Y., Matsuzaki, Y., Sato, T. and 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 Computational Biology*, Vol.7, No.6, pp.991-1012 (2009)
- 17) Mintseris, J., Wiehe, K., Pierce, B., Anderson, R., Chen, R., Janin, J. and Weng, Z.: Protein-Protein Docking Benchmark 2.0: an update, *Proteins*, Vol.60, No.2, pp.214-216 (2005)
- 18) Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., et al.: The Protein Data Bank: a computer-based archival file for macromolecular structures, *J Mol Biol*, Vol.112, pp.535-542 (1977)
- 19) RCSB: Protein Data Bank, available from (<http://www.rcsb.org/pdb/>) (accessed

- 2010-04-21).
- 20) Ohue, M., Matsuzaki, Y., Matsuzaki, Y. and Akiyama, Y.: Improvement of all-to-all protein-protein interaction prediction system MEGADOCK, *The 20th International Conference on Genome Informatics (GIW2009)*, P-033 (2009)
- 21) Yoshikawa, T., Tsukamoto, K., Hourai, Y. and Fukui, K.: Improving the accuracy of an affinity prediction method by using statistics on shape complementarity between proteins, *J Chem Inf Model*, Vol.49, pp.693-703 (2009)
- 22) Yoshikawa, T., Seno, S., Takenaka, Y. and Matsuda, H.: Improved Prediction Method for Protein Interactions Using Both Structural and Functional Characteristics of Proteins, *IPSJ Transactions on Bioinformatics*, Vol.3, pp.10-23 (2010)
- 23) Tsuji, T., Suzuki, M., Takiguchi, N. and Ohtake, H.: Biomimetic control based on a model of chemotaxis in *Escherichia coli*, *Artif. Life*, Vol.16, pp.155-177 (2010)
- 24) Wadhams, G.H. and Armitage, J.P.: Making sense of it all: bacterial chemotaxis, *Nat. Rev. Mol. Cell Biol.*, Vol.5, pp.1024-1037 (2004)
- 25) Baker, M.D., Wolanin, P.M. and Stock, J.B.: Systems biology of bacterial chemotaxis, *Curr. Opin. Microbiol.*, Vol.9, pp.187-192 (2006)
- 26) Kentner, D. and Sourjik, V.: Dynamic map of protein interactions in the *Escherichia coli* chemotaxis pathway, *Mol. Syst. Biol.*, Vol.5, pp.238 (2009)
- 27) van Albada, S.B. and Ten Wolde, P.R.: Differential affinity and catalytic activity of CheZ in *E. coli* chemotaxis, *PLoS Comput. Biol.*, Vol.5, e:1000378 (2009)
- 28) Matsuzaki, Y., Kikuchi, S. and Tomita, M.: Robust effects of Tsr-CheBp and CheA-CheYp affinity in bacterial chemotaxis, *Artif. Intell. Med.*, Vol.41, pp.145-150 (2007)
- 29) Kanehisa, M. and Goto, S.: KEGG: kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.*, Vol.28, pp.27-30 (2000)
- 30) Fujibuchi, W., Goto, S., Migimatsu, H., Uchiyama, I., Ogiwara, A., Akiyama, Y. and Kanehisa, M.: DBGET/LinkDB: an integrated database retrieval system, *Pac. Symp. Biocomput.*, pp.683-694 (1998)
- 31) Hwang, H., Pierce, B., Mintseris, J., Janin, J. and Weng, Z.: Protein-protein docking benchmark version 3.0, *Proteins*, Vol.73, pp.705-709 (2008),
- 32) Park, S.Y., Chao, X., Gonzalez-Bonet, G., Beel, B.D., Bilwes, A.M. and Crane, B.R.: Structure and function of an unusual family of protein phosphatases: the bacterial chemotaxis proteins CheC and CheX, *Mol. Cell*, Vol.16, pp.563-574 (2004)
- 33) Szurmant, H., Muff, T.J. and Ordal, G.W.: *Bacillus subtilis* CheC and FliY are members of a novel class of CheY-P-hydrolyzing proteins in the chemotactic signal transduction cascade, *Biol. Chem.*, Vol.279, pp.21787-21792 (2004)
- 34) Rosario, M.M. and Ordal, G.W.: CheC and CheD interact to regulate methylation of *Bacillus subtilis* methyl-accepting chemotaxis proteins, *Mol. Microbiol.*, Vol.21, pp.511-518 (1996)
- 35) Normanno, N., Maiello, M.R. and De Luca, A.: Epidermal growth factor receptor tyrosine kinase inhibitors(EGFR-TKIs): simple drugs with a complex mechanism of action?, *Journal of Cellular Physiology*, Vol.194 pp.13-19 (2002)
- 36) Selvaggi, G., Novello, S., Torri, V., et al.: Epidermal growth factor receptor over-expression correlates with a poor prognosis in completely resected non-small-cell lung cancer, *Annals of Oncology*, Vol.15, pp.28-32 (2004)