情報処理学会論文誌　数理モデル化と応用　Vol. 3　No. 3　91–106 (Oct. 2010)

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]

タンパク質間相互作用（Protein-Protein Interaction，PPI）に関するネットワークの解明は，細胞システムの理解や構造ベース創薬に重要な課題であり，網羅的 PPI 予測手法の確立が求められている．タンパク質立体構造データ群から網羅的に相互作用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，なかでも 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．その結果，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合せであるにもかかわらず，現在までに報告されていないものとして，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]

The elucidation of the protein-protein interaction (PPI) network is an important problem in the understanding of the cellular system and structure-based drug design. An effective way to conduct exhaustive 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 realize 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

†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.

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うえで，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，すで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 × 1,000（百万通り）の総当たり計算を実施することを目標として掲げている．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な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はる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どうしの網羅的な相互作用予測（**図 2**）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お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る．

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 1 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．

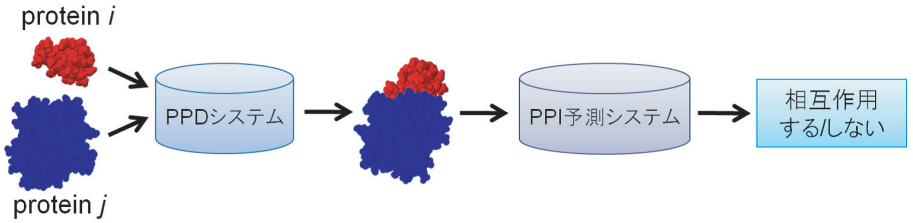

図 1　タンパク質立体構造情報を利用した，タンパク質ドッキング（protein-protein docking，PPD）による網羅的 PPI 予測
Fig. 1　All-to-all protein-protein interaction (PPI) prediction using protein tertiary structure information by protein-protein docking (PPD).

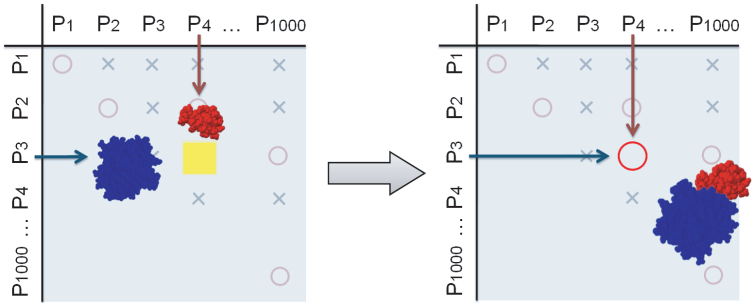

図 2　網羅的 PPI 予測の概略
Fig. 2　Scheme of all-to-all protein-protein interaction prediction.

2.　MEGADOCK：ドッキング計算システム部

図 1 に示した MEGADOCK システムのドッキング計算部について以下に詳細を述べる．

2.1　タンパク質ドッキングの概要

本論文が指すタンパク質ドッキングとは，タンパク質がほぼ剛体と見なせて複合体形成の際に構造が変化をともなわないという仮定のもと，表面形状の相補性を主として計算する数理モデルに基づいた手法である．タンパク質ドッキングという用語は，広くは分子動力学法

を用いた分子シミュレーションによる方法なども含むものであるが，分子動力学法のように原子の挙動を精密にモデル化して時系列的に変化を調べるシミュレーション手法では，1つのタンパク質ペアに関する計算だけでも数週間以上の時間を要し，我々の目的であるタンパク質群どうしの網羅的な計算には向かない．タンパク質のような巨大な分子間のドッキング計算を大量に行うには，より単純化されたモデルを選択する必要がある．

単純化されたドッキングモデルの代表例としては，タンパク質を3次元ボクセルに分割してそれぞれを離散関数としてスコア化し，それらの相関関数によって評価値を決定するボクセルモデルがあり，MolFit[4]，FTDock[5]，ZDOCK[6)–9)] などのタンパク質ドッキングソフトウェアがこのモデルを利用している．スコアの値にはそれぞれのソフトウェアごとの特徴が現れているが，たとえば Boston 大学の ZDOCK ではドッキング予測を少しでも精密にするために，形状相補性の計算には複素数で表される Pairwise Shape Complementarity（PSC）スコア[7)] を用い，さらに複数の物理化学的相互作用を考慮することを行ってきた．結果として高い精度でのドッキング予測が可能になったが，我々が目標とする網羅的 PPI 予測に利用するには，1回のドッキングの計算時間が増大しすぎている．そこで我々は網羅的 PPI 予測用ドッキングシステムとして，独自の MEGADOCK システムを開発することとした[10)–12)]．

2.2　ドッキング計算法

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 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辺が 1.2Å 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，我々が本論文で新規に提案する real Pairwise Shape Complementarity（rPSC）スコアを用いる．rPSC スコアは以下のように表される[11)]．

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

表 1　Protein-protein docking benchmark 2.0 から選出した 23 個の複合体
Table 1　The selected 23 complex structures from the Protein-protein docking benchmark 2.0.

Complex PDB ID									
1ACB	1AK4	1AVX	1AY7	1B6C	1CGI	1D6R	1E96	1EAW	1EWY
1GCQ	1GHQ	1GRN	1HE1	1KAC	1KTZ	1PPE	1SBB	1UDI	2PCC
2SIC	2SNI	7CEI							

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

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，タンパク質ドッキング予測に広く用いられているベンチマークデータセットである protein-protein docking benchmark 2.0[13)] に含まれる，23個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．事前実験に用いたタンパク質複合体の PDB[14),15)] ID リストを**表 1** に示す．

rPSC スコアは実数のみでの表現ではあるが，表面形状の相補性をできるだけ精密に表現したものである．ZDOCK の PSC スコアほど精度は良くないが，後述するように計算時間の面で有利なモデルとなっている．rPSC スコアの与え方を2次元の概略図として表したものを**図 3** に示す．

また，静電的相互作用による項 E については次のように計算する．ボクセル $i(l,m,n)$ に対する電界 ϕ_i を，

$$\phi_i = \sum_j \frac{q_j}{\varepsilon(r_{ij}) r_{ij}}, \qquad \varepsilon(r) = \begin{cases} 4 & (r \le 6\text{Å}) \\ 38r - 224 & (6\text{Å} < r < 8\text{Å}) \\ 80 & (r \ge 8\text{Å}) \end{cases} \tag{3}$$

と定義する．ただし q_j はボクセル j の電荷，r_{ij} は i と j のユークリッド距離[⋆1]，$\varepsilon(r)$ は誘電率をモデル化した関数である．アミノ酸残基ごとに CHARMM19[16)] に基づいて原子に電荷を与え，ボクセルに分割してボクセル電荷 $q(l,m,n)$ を決定し，静電的相互作用の項 $E_R(l,m,n), E_L(l,m,n)$ を，

⋆1 2Å 以下の場合は 2Å と定める．

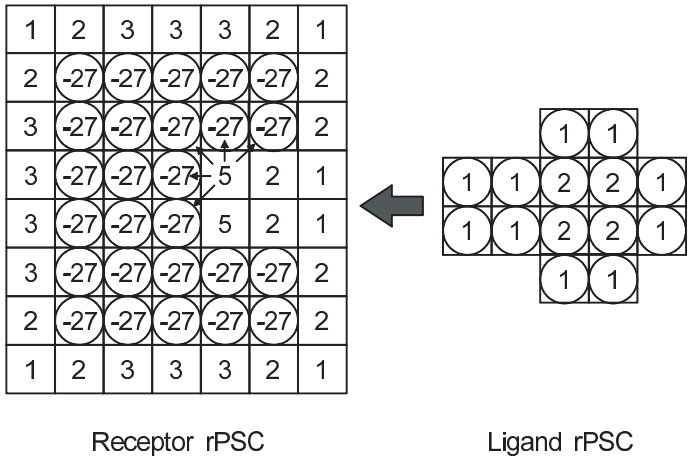

図 3　rPSC スコアの付与の概略図．本来は 3 次元ボクセルであるが，ここでは簡単のため 2 次元上で表している．四角と丸が組み合わさっているボクセルはタンパク質の実体を含んでおり，四角のみのボクセルは空間である．なお，スコア値として 0 が与えられるボクセルは省略している

Fig. 3　rPSC scoring model. The model consists of 3D voxels. Here the model is shown by two dimensions for simplicity. The voxels with a square and a circle interlace correspond to the area occupied by protein, and voxels with only squares are non-occupied space. Voxels that have score 0 are not shown here.

$$E_R(l,m,n) = \begin{cases} \phi_{i(l,m,n)} & \text{(entire voxel excluding core)} \\ 0 & \text{(core of the R)} \end{cases} \tag{4}$$

$$E_L(l,m,n) = q(l,m,n) \tag{5}$$

と決める．

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 $\boldsymbol{S}$ を

$$\boldsymbol{R}(l,m,n) = G_R(l,m,n) + iE_R(l,m,n) \tag{6}$$

$$\boldsymbol{L}(l,m,n) = G_L(l,m,n) + iwE_L(l,m,n) \tag{7}$$

$$\boldsymbol{S}(\alpha,\beta,\gamma) = \Re\left[\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N}\boldsymbol{R}(l,m,n)\boldsymbol{L}(l+\alpha,m+\beta,n+\gamma)\right] \tag{8}$$

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数 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 17) 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，ボクセル数が N のとき，$3{,}600 \times N^3$ 通りとなる．

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，

$$\boldsymbol{S}(\alpha,\beta,\gamma) = \Re\left[\mathrm{IFT}\left[\mathrm{DFT}[\boldsymbol{R}(l,m,n)]^{*}\mathrm{DFT}\left[\boldsymbol{L}(l,m,n)\right]\right]\right] \tag{9}$$

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．

2.3　複合体構造サンプリング数の増加

ZDOCK では，各サンプリング角度ごとに複合体の構造を 1 つのみ報告していた．これはドッキングすることが知られているペア間で正しい姿勢を求める問題においては，角度ごとの複合体構造サンプリング数を増やしてもあまり効果がないという評価実験の結果に基づくものであるが，新たな PPI 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．我々の 1 つの作業仮説として，「相互作用を有するタンパク質ペアはドッキング予測における結合部位が空間的に密に分布する」というものがある，たとえばドッキング計算結果として得られた複数の結合予測位置がある 1 カ所に集中している場合と，複数箇所に散在している場合とでは，1 カ所に集中しているペアの方が相互作用している可能性が高いであろうという仮説である．すなわち，角度ごとの構造数を複数個に増やすことによって，より結合予測位置の分布が明確になり，相互作用の予測がしやすくなると考えられる．そこで，MEGADOCK のドッキングシステム部の出力構造数については，各角度ごとに 1 個ではなく t 個の構造を出力することとした．今後，t は角度ごとの出力構造数を表すものとする．

2.4　網羅的 PPI 予測のための大規模並列化

MEGADOCK を用いて，多数のレセプタタンパク質とリガンドタンパク質の間での網羅

的 PPI 予測を行う場合，それぞれのペアごとの計算は基本的には独立であるため，同時に並列計算することが可能である．ただしここで，リガンドタンパク質については，回転による変換を施すたびに FFT 計算が必要であるが，レセプタタンパク質は回転させないという非対称性に注意する必要がある．

MEGADOCK は MPI ライブラリにより並列化されており，複数のレセプタとリガンドが入力されたとき，ユーザの指定する方式により，複数のプロセッサ間でレセプタとリガンドを分配した並列計算を行う（現在までのところ 256〜512 コア程度の並列計算を実施している）．このとき，m 個のレセプタと n 個のリガンドを受け取ったプロセッサは，n 個のリガンドの 1 つずつを順に取り出し，指定された角度刻みごとに FFT 化を行い，そのたびに最内ループとして m 個のレセプタとの比較を行う．これは回転角ごとの FFT を無駄に繰り返さないためである．m 個のレセプタの FFT は 1 度だけ計算してメモリ上に保持されている．

I/O 能力が比較的高いシステムにおいては，網羅的計算におけるレセプタまたはリガンドの FFT をあらかじめ全システム内で 1 度だけ計算し，様々な相手のタンパク質に対してディスクからの読み出しで実施する方が高速な場合がある（東工大 TSUBAME 1.2 システムの例では最大 3 倍程度高速である）．このため MEGADOCK では，FFT 結果をライブラリから読み出す機能も実装している．レセプタとリガンドがともに FFT ライブラリから読み出されるケースでは，畳み込み和の計算結果から逆 FFT を得る際の 1 回のみフーリエ変換が必要となり，変換の計算量は 1/3 となる．

ただしこのとき，ライブラリ化する FFT の底をどのようにとるかにトレードオフが存在する．MEGADOCK では，タンパク質を囲む立方体のサイズを必要最小限に抑えるため，2，3，5，7，11 の 5 種類の素数を FFT の底として自動的に組み合わせる機能を実現している．たとえば 2 のべき乗しか許さない実装では，サイズを 1 ボクセル分大きくする目的だけで無駄に 8 倍の体積となっていたものが，きめ細かい節約が可能になる．しかしライブラリ化の場合，畳み込みの相手のタンパク質と底を揃える必要があり，底の種類を豊富にしすぎるとライブラリ容量の増加につながり，かえって低速化を招く．我々の検討結果では，2，3，5 の 3 種類の素数の組合せを底として許してライブラリ化するのが，全体性能の面で優れていた．

I/O 能力に比べて計算能力が高いシステム，たとえば FFT 計算を GPU 上で実現する GPU 版の MEGADOCK においては，FFT ライブラリの読み込み機能は用いずに，FFT の底を自由に選ぶオプションを選択するほうが全体性能が高くなる．

現在の MEGADOCK の実装は，MPI ライブラリにより大きな単一のジョブとして並列実行する形式となっているが，クラウド環境などでの実現においては，むしろタンパク質ペアごとの細かい多数のジョブとして管理すべきであろう．これは将来の課題である．

3. MEGADOCK：タンパク質間相互作用（PPI）予測システム部

本研究で新たに提案する MEGADOCK における PPI 予測システム部について述べる．PPI 予測システム部の構成を**図 4** に示す．以下では各項目について説明する．

3.1 エネルギースコアに基づくリランキング

ドッキングソフトウェアが探索を行うドッキング姿勢の組合せは $3{,}600 \times N^3$ 通り存在し，これらから順位付けを行って一部を抽出するという操作がドッキング計算である．ドッキング計算に用いられるモデルは計算の高速性を重視しているためにさほど精密ではなく，実際にはこのドッキングスコアが良くても結合エネルギー値が高く現実的ではない複合体予測が多数存在する．そのような予測構造を正しく取り除くためには様々な方法が考えられる．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うのが望ましいが，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 においては Pierce らが開発したリランキングシステム ZRANK [18] 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・

(1) 網羅的ドッキング予測（MEGADOCK ドッキングシステム部）
　複合体出力数は $2000 \times t$ 個（t は回転角ごとの出力構造数）
(2) エネルギースコアに基づくリランキング
　$2000 \times t$ 個を ZRANK のエネルギースコアによってリランキングし，
　上位 1000 個を取り出す
(3) 構造類似度に基づくクラスタリング
　1000 個の構造をクラスタリングにかける
(4) PPI 評価値の計算
　メンバ数が閾値 m^* 以上のクラスタの中で，最も良い順位を持つ
　予測構造のドッキングスコア（z 値）を評価値 E とする
(5) PPI 判定
　評価値 E がある閾値 E^* 以上となるタンパク質ペアを「相互作用する」と判定

図 4　提案する網羅的 PPI 予測手法
Fig. 4　Proposed all-to-all protein-protein interaction prediction method.

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スコアに基づいて，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スコア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け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　構造類似度に基づくクラスタリング

Matsuzaki ら[20),21)] 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 21) 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[22)] 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 の中のデータのうち，ZRANK による評価順位が最も良いものを代表データとする．

(2)　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を母集団とした代表データのドッキングスコアの z 値を s_i とする．

(3)　各クラスタのメンバ数を母集団としたクラスタ C_i のメンバ数の z 値を m_i とする．

(4)　m_i が閾値 m^* より大きいクラスタの集合を C' とする．C' に含まれる代表値 s_i の中で最大のものを評価値 E とする．

$$C' = \{C_i \mid m_i \geq m^*\} \tag{10}$$

$$E = \begin{cases} \max s_i \ (i \in C') & (C' \neq \emptyset) \\ 0 & (C' = \emptyset) \end{cases} \tag{11}$$

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．

ここまでで述べた MEGADOCK の最新版を，MEGADOCK 2.1 と呼ぶ．

4.　ベンチマークデータセットによる評価実験

protein-protein docking benchmark 2.0 に含まれる複合体のうち，タンパク質複合体がモノマどうしで形成されている 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 ID リストを**表 2** に示す．

ここでは MEGADOCK のドッキング性能，特にドッキング出力構造を角度あたり t 個に増やした効果と，MEGADOCK に対する ZRANK によるリランキングの適用の効果を検証することを目的とする．

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 root mean square deviation（L-RMSD）を計算した．原子数 N_a のタンパク質 i と j のすべての原子に，ある一定の規則で順序付けし，k 番目に対応付けられた原子の位置ベクトルをそれぞれ $\boldsymbol{r}_k^i$，$\boldsymbol{r}_k^j$（$k = 1, 2, \ldots, N_a$）とすると，RMSD は次式で表される．

$$\mathrm{RMSD} = \sqrt{\frac{1}{N_a}\sum_{k=1}^{N_a}\left|\boldsymbol{r}_k^i - \boldsymbol{r}_k^j\right|^2} \ [\text{Å}] \tag{12}$$

また，L-RMSD が 5Å 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．

1 対 1 のドッキング性能については，結果を ZDOCK 3.0[9)] と比較し評価する．なおす

表 2　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 2　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

⋆1 ACE（Atomic Contact Energy）スコア[19)] を用いている．詳しくは文献 18) を参照されたい．

表 3　MEGADOCK 2.1 と ZDOCK 3.0 のドッキング性能比較
Table 3　Docking performance comparison of MEGADOCK 2.1 and ZDOCK 3.0.

複合体	MEGADOCK 2.1				ZDOCK 3.0			
PDB ID	Min L-RMSD	Hits	Best Rank	Best L-RMSD	Min L-RMSD	Hits	Best Rank	Best L-RMSD
1ACB	1.49	4	9	1.49	1.68	9	12	4.52
1AK4	13.70	0	–	–	1.57	2	56	1.57
1ATN	1.38	2	202	3.86	1.59	2	1	1.59
1AVX	2.56	3	7	2.56	2.47	4	2	2.47
1AY7	1.81	4	5	1.81	2.05	6	16	2.05
1B6C	1.92	2	3	1.92	1.76	1	1	1.76
1BUH	1.47	2	70	1.47	2.02	7	5	2.02
1BVN	1.50	8	2	2.93	1.91	14	1	1.91
1CGI	1.02	10	1	1.02	1.33	11	1	1.33
1D6R	1.85	7	3	1.85	1.51	1	1,512	1.51
1DFJ	2.21	2	19	2.21	2.77	2	2	4.56
1E6E	1.34	5	2	1.34	1.15	6	1	1.15
1E96	12.64	0	–	–	4.73	1	1,790	4.73
1EAW	1.46	6	1	1.46	2.02	3	3	2.02
1EWY	1.18	1	607	1.18	1.71	6	5	1.71
1F34	1.74	2	1	1.74	1.79	2	1	1.79
1FC2	1.71	2	41	1.71	2.09	2	11	2.09
1FQ1	1.58	1	48	1.58	1.97	1	15	1.97
1FQJ	1.91	1	140	1.91	1.13	1	330	1.13
1GCQ	1.35	7	1	1.35	1.37	4	3	1.37
1GHQ	13.87	0	–	–	20.95	0	–	–
1GRN	1.42	4	1	1.42	1.74	4	1	1.74
1H1V	12.15	0	–	–	5.31	0	–	–
1HE1	1.44	4	1	1.44	1.77	4	1	1.77
1HE8	7.86	0	–	–	4.28	1	393	4.28
1I2M	1.69	1	1	1.69	1.73	1	1	1.73
1IBR	2.09	1	1	2.09	2.13	1	1	2.13
1KAC	1.76	1	41	1.76	1.27	5	9	3.42
1KTZ	1.59	2	221	1.59	1.62	7	63	1.62
1KXP	2.51	1	1	2.51	2.14	1	1	2.14
1KXQ	1.34	4	10	1.36	1.36	5	1	1.36
1M10	2.93	1	1	2.93	3.40	2	1	3.40
1MAH	1.37	8	1	1.37	1.52	7	1	1.52
1PPE	1.43	22	1	1.43	1.46	21	2	2.39
1QA9	1.03	1	40	1.03	1.46	4	23	1.46
1SBB	20.42	0	–	–	6.08	0	–	–
1TMQ	1.81	2	1	1.81	2.07	5	1	2.07
1UDI	1.31	10	1	1.31	1.18	10	1	1.18
1WQ1	1.04	4	1	1.04	1.81	3	1	1.81
2BTF	1.33	5	1	2.89	1.24	5	6	1.24
2PCC	5.03	0	–	–	1.77	2	32	1.77
2SIC	1.85	3	2	1.85	1.98	5	1	1.98
2SNI	1.61	8	1	1.61	1.10	7	1	1.10
7CEI	1.36	6	1	1.36	1.26	5	5	4.78

べての実験において，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

表 3 に MEGADOCK 2.1（$t = 1$）と ZDOCK 3.0 のドッキング結果を示す．出力構造数はどちらも 2,000 個とした．なお，表 3 の値には以下のものを用いている．

- **Min L-RMSD**：予測構造 2,000 個の中で L-RMSD が最も小さかったものの L-RMSD 値 [Å]
- **Hits**：予測構造 2,000 個の中で L-RMSD が 5Å 以下のもの（正解構造）の個数

表 4　Best Rank の範囲とそれに含まれる複合体の個数
Table 4　The number of successful complex predictions included in each range of Best Ranks.

Best Rank	MEGADOCK 2.1	ZDOCK 3.0
1（正解数）	18	19
2〜10	9	10
11〜100	6	8

表 5　ZRANK による MEGADOCK のドッキング予測の正解数
Table 5　The number of successful complexes predicted by MEGADOCK and ZRANK.

角度あたり候補数 t	1	2	3	4	5	10	15	20	ZRANK なし
正解数	22	23	24	24	24	24	25	25	18

- **Best Rank**：正解構造の中で，ドッキングシステムのつけた順位が最も良かったものの順位値
- **Best L-RMSD**：Best Rank に記載した順位の構造の L-RMSD 値 [Å]

表 3 より，MEGADOCK 2.1 のドッキング性能は ZDOCK 3.0 に若干劣るものの，ほぼ同等な精度のドッキング予測結果が得られているといえる．特にドッキング結果で重要と考えられる正解数（表 3 中の灰色のセルの数）が，MEGADOCK 2.1 では 18 個，ZDOCK 3.0 では 19 個となっており，ZDOCK に近い予測精度を得られていることが分かる．また，**表 4** に表 3 の Best Rank の範囲別に含まれる複合体の個数を集計したものを示す．表 4 において，Best Rank が 1 の行の値は正解数を表しており，Best Rank が 2〜10 の行は（1 位に正解構造がなかったが）2 位〜10 位の中で正解構造が見つかった複合体の数を，11〜100 の行は同様に 11 位〜100 位の中で正解構造が見つかった複合体の数を表している．1 位だけでなく他の順位を比較しても，MEGADOCK 2.1 は ZDOCK 3.0 の性能に大きく劣ってはいないことが分かる．なお，1 つのペアのドッキング予測にかかる平均計算時間は ZDOCK 3.0 が 230 分なのに対し MEGADOCK 2.1 は 57 分であり，ZDOCK に比べて約 4 倍高速に，かつ同等の精度でドッキング計算が可能であるといえる．

ここまでの結果に対し，MEGADOCK の角度あたりの出力構造数 t を増加させて ZRANK を組み合わせることで結果がどのように変わるかを調べた．t を変化させて ZRANK を使用したときの，44 複合体中で正解となった個数を**表 5** に示す．またこのときの ZRANK の平均計算時間を**表 6** に示す．表 5 より，ZRANK をかけた場合は t の値の増加に従って 1 位が正解である複合体の数が増えていることが分かる．しかし表 6 に示すように，t の値を増

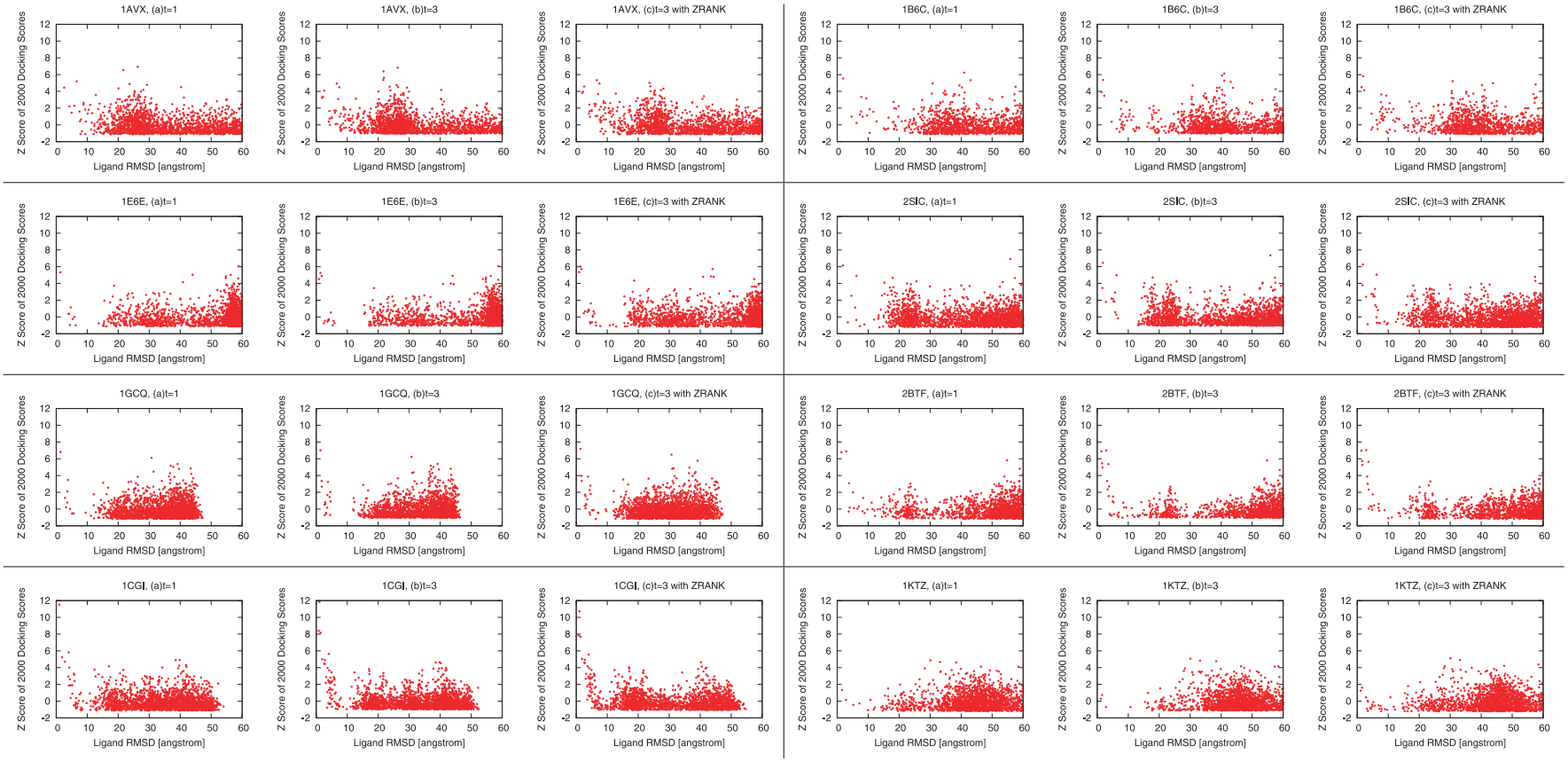

図 5　Ligand RMSD とドッキングスコア（の z 値）の 2 次元プロット
Fig. 5　Ligand RMSD versus docking score (z score) plots.

表 6　ZRANK によるリランキングに要した平均計算時間 [分]
Table 6　Calculation time for ZRANK re-ranking [min].

角度あたり候補数 t	1	2	3	4	5	10	15	20
計算時間 [分]	4.0	8.1	12.1	15.4	20.1	40.2	58.2	80.3

加させると ZRANK による計算時間も線型に増えていくので，正解数との兼ね合いで，ここでは $t = 3$ が適切であると判断した[⋆1]．

なお，ZRANK をかけない場合は，予測数が増えても 1 位の予測構造が変わることはなく，正解数は 18 個であった．$t = 3$ にすることで新たに正解となった複合体は，1AVX，1B6C，

⋆1 後の網羅的 PPI 予測における性能評価実験で，$t = 3$ で最も良い結果が得られることを述べる．

1BVN，1DFJ，1E6E，1KAC，1KXQ，2SIC の 8 個であり，逆に $t = 3$ にすることで不正解となった複合体は 1GCQ と 2BTF の 2 例存在した．ドッキング予測結果のスコアと L-RMSD との関係を調べるために，これらの複合体を含めた 8 個の複合体（1AVX，1B6C，1E6E，2SIC，1GCQ，2BTF，1CGI，1KTZ）に関するドッキングスコアと L-RMSD の分布を**図 5** に示す．図 5 は，(a) $t = 1$ のもの，(b) $t = 3$ のもの，(c) $t = 3$ のときに ZRANK を使用したもの，のそれぞれについて，上位 2,000 個のドッキングスコアを母集団とする z 値と，そのスコアを持つ予測複合体の L-RMSD を上位 2,000 個プロットしたものである．

グラフの点が左上に集中していると，L-RMSD 値が小さい，すなわち真の解に近いときに予測のスコアが高いということであり，正しい予測が可能であることになる．全体を通して (a) $t = 1$ よりも (b) $t = 3$，(b) $t = 3$ よりも (c) $t = 3$ と ZRANK を併用したものの方が，程度の差はあるものの，左側（L-RMSD が 10Å 以下）にプロットされた点が増加する傾向があることが分かる．また 1B6C や 2SIC などに見られるように，高いドッキングスコアにもかかわらず L-RMSD の大きいノイズと見なされる予測構造が，ZRANK によってうまく除外されている．1GCQ と 2BTF については ZRANK によって逆にノイズを生成してしまった結果となったが，Best Rank の値の変化はいずれも 1 から 2 とわずかなものであり，大きな精度の悪化を引き起こしているわけではない．

ドッキング予測結果の構造例を**図 6** に示す．左は PDB の 1CGI，右は 1KTZ であり，surface 表示のタンパク質がレセプタで，リボン表示のタンパク質がリガンドである．リガンドは 2 つ表示しており，緑色で表されているものが MEGADOCK のドッキング予測結果で，赤色で表されているのものが実際の正解構造となる X 線結晶構造解析による構造である．1CGI については MEGADOCK の予測が 1 位となったものの結果であり，L-RMSD は 1.02Å である．1KTZ は，最も小さい L-RMSD を得た予測が ZRANK によるリランキング後で 7 位のものであり，1.59Å であった．形状による予測難易度の大小はあるものの，実際の構造に近い複合体構造の予測が可能となっていることが見てとれる．

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 2 の 44 組の複合体に対してレセプタとリガンドの全対全の網羅的な PPI 予測を行い，精度を検証する．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 $44 \times 44 = 1{,}936$ 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 予測フローに従って相互作用予測を全 1,936 通りについて行った．もともと結合していた 44 組のレセプタとリガンドの組合せを正例，そうでない組合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，

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

によって F 値を計算し，相互作用予測性能の評価に用いた．

PPI 予測システム部の平均計算時間は，1 つのタンパク質ペアの 1,000 個のドッキング予測構造のクラスタリングに約 8 分，ZRANK によるリランキングは表 6 に示したとおりである．

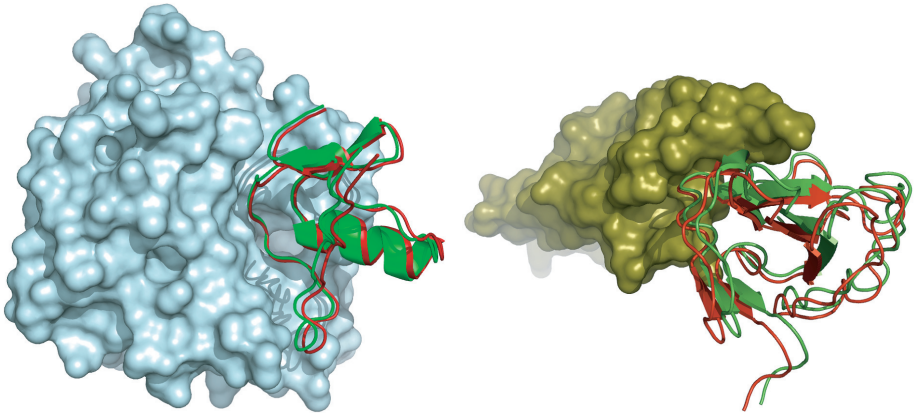

図 6　ドッキング予測によって得られた複合体構造．左は 1CGI，右は 1KTZ であり，surface 表示のタンパク質がレセプタで，リボン表示のタンパク質がリガンドである．緑は MEGADOCK のドッキング予測結果で，赤は X 線結晶構造解析によるものである

Fig. 6　Complex structure predicted by docking (left: 1CGI, right: 1KTZ). Proteins shown by surface correspond to receptors and that shown by ribbon representation correspond to ligands. Green colored ligands show the prediction by MEGADOCK and Red colord ligands are X-ray structure.

4.2.2　結果と考察

$t = 1$，2，3，5，10，20 のそれぞれの相互作用予測結果を**表 7** に示す．表 7 より，ZRANK をかけた場合で，$t = 3$ のときに最大 F 値 $= 0.415$ をとることが分かる．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 $= 0.150$ [11] や，機械学習で生成した判別器による F 値 $= 0.251$ [12] 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

表 7　44 複合体の網羅的タンパク質間相互作用予測結果
Table 7　Result of 44 × 44 protein-protein interaction prediction.

角度あたり候補数 t	1	2	3	5	10	20
ZRANK なしの予測の F 値	0.300	0.299	0.286	0.277	0.273	0.281
ZRANK ありの予測の F 値	–	0.391	0.415	0.340	0.366	0.318

しているといえる．また，すべてのデータ（$t = 1$，2，3，5，10，20）に対して，ZRANK によるリランキングを実施した部分の F 値が向上しており，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，Matsuzaki らによる手法[22]，Yoshikawa らの AEP[23] や RFZ/LFZ[24] と呼ばれる手法がある．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 22) 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，計算速度の面で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，あらかじめデータセット内のタンパク質を機能情報によってサブセットに分割し，そのサブセット内で相互作用予測を行う手法である．我々のシステムはそのような情報を事前に調べることなく，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の利用を想定しているため，直接的な比較は難しい．相互作用の予測精度で比較を行うとすると，RFZ/LFZ で得られた最も良い精度は F 値 0.471 であり，そのときのサブセットは $12 \times 12 = 144$ 通りのタンパク質ペアからなるものである．我々の手法よりも良い F 値が得られているが，我々のデータセットは 1,936 通りであり，Yoshikawa らが最も良い精度を得た 144 通りのサブセットの 10 倍以上の大きさの問題であるため，単純な比較は困難である．なお，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．

クラスタのメンバ数の閾値によって True Positive fraction と False Positive fraction がどのように動くかを確認するため，表 7 の中で最も良好な結果である $t = 3$ の ROC 曲線を**図 7** に示す．m^* は 2.5 節に示してあるとおりクラスタのメンバ数の z 値に対する閾値である．図 7 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分かる．$m^* = 0.0$ や $m^* = 0.5$ 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いえる．

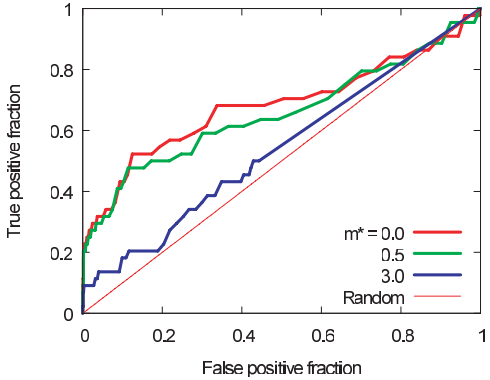

図 7　本システムのベンチマークデータセットへの適用における精度（$t = 3$，ZRANK あり）．m^* と E^* を変化させて行った実験における ROC 曲線を示す．縦軸が True positive fraction（$\frac{TP}{TP+FN}$）を，横軸が False positive fraction（$\frac{FP}{FP+TN}$）を示す．予測がランダムに行われた場合は対角線のようになる
Fig. 7　Evaluation of the docking post-processing system ($t = 3$, ZRANK used). The ROC curves 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.

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 22) 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．たと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，栄養となる物質のより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はきわめてエネルギー効率の良い分子モータとしても注目され，機械分野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[25]．

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

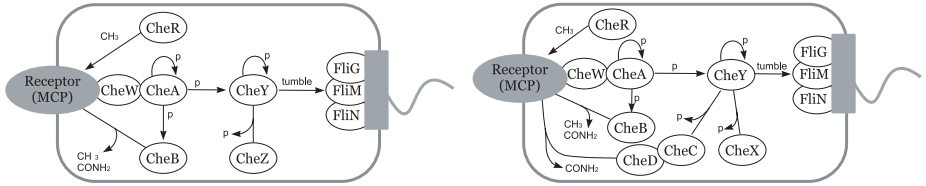

図 8　*E. coli*（左），*T. maritima*（右）の走化性シグナル伝達系．これらの細菌の運動はべん毛モータの回転調節により制御される．べん毛モータはリン酸化状態の CheY によって回転を制御されている．栄養物質など好ましい環境シグナルを受容体（Methyl-accepting Chemotaxis Proteins，MCP）が感知すると，CheA の自己リン酸化が抑制される．すると，CheA からリン酸基を受け取っていた CheY のリン酸化レベルが下がり，CheY はべん毛モータに結合しにくくなる．その結果，細菌はより長く直進運動を続けるようになり，好ましい環境に近づいていく．続いて，CheR，CheB などによる MCP のメチル化状態の制御により刺激への順応がおこる．MCP には Tar，Tsr，Trg，Tap など，検知する刺激の種類により数種類ある

Fig. 8　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. The phosphorylation state of CheY is responsible for the rotation direction. When the receptors (Methyl-accepting Chemotaxis Proteins, MCP) sense favorable signals such as those indicating nutrition molecules in the environment, CheA autophosphorylation is inhibited. Then the phosphorylation level of CheY will be reduced because of the repression of phosphotransfer from CheA. That low phosphorylation level of CheY reduces its affinity to the flagellar motor, which causes more frequent counterclockwise rotation and longer periods of smooth swimming of the cell. In addition, the stimulated receptors also undergo a gradual change in the methylation level controlled by CheR and CheB. That causes adaptation to the signal. The MCP family comprises Tar, Tsr, Trg, Tap and Aer, each of which senses distinct signals.

いる[26),27)]（**図 8**，**表 8**）．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．刺激に対する敏感性や，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[28)–30)]．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 の評価を行うとともに，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．今回は走化性系のより多くのタンパク質種を取り扱えるよう，3 生物種のデータをすべて用いて，ホモログどうしは区別せずに同一のタンパク質種として扱った．これらの種の間でホモロガスな走化性系タンパク質は，構造データを用いた他の大規模 PPI 解析例における基準（アミノ酸配列の類似性が 40%以上）[31)] で冗長と見なしている範囲に含まれる．

表 8　走化性系のタンパク質．† は，今回の対象生物種のうち *T. maritima* のみに，‡ は *E. coli* と *S. typhimurium* のみにみられる

Table 8　Proteins that constitute the chemotaxis system.

タンパク質	役割
MCP	刺激物質の化学受容体（methyl-accepting chemotaxis proteins，MCP）． 本論文では Tsr（Ser 受容体）を扱う．
CheA	自己リン酸化酵素（ヒスチジンキナーゼ）． 自己リン酸化し，CheY と CheB にリン酸基を供与する．
CheB	MCP 脱メチル化，脱アミド化酵素．リン酸化されると脱メチル化の活性が上昇する．
CheC†	CheY の脱リン酸化酵素．CheD と結合することで脱リン酸化の活性が上昇する．
CheD†	MCP 脱アミド化酵素．
CheR	MCP メチル化酵素．
CheW	MCP と CheA の足場タンパク質．
CheX†	CheY の脱リン酸化酵素．
CheY	リン酸化されるとべん毛モータと相互作用し，時計回りの回転（菌体は方向転換）を促す．
CheZ‡	CheY の脱リン酸化酵素．
FliM	べん毛モータタンパク質．
FliN	べん毛モータタンパク質．
FliG	べん毛モータタンパク質．

構造データの収集は以下のように行った．まず，KEGG[32)] 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，LinkDB[33)] を用いて PDB の ID を得た．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 34) と同様の条件）を実験対象の候補として選択した（**表 9**）．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．また，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，ある 1 つのタンパク質について以上の条件をすべて満たす構造データが得られない場合には，変異の入ったタンパク質構造を例外的に使用した（表 9 の注を参照）．

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，MEGADOCK 2.1 による $89 \times 89 = 7,921$ 通りのドッキングと相互作用予測を行った．

表 9　実験に使用した走化性系のタンパク質構造データ．† 例外として使用したデータ．134 番目の残基に変異がある（Glu → Lys）

Table 9　Chemotaxis dataset derived from PDB.

PDB ID	Chain	Organism	Molecule	Domain
1FFG	B,D	*E. coli*	CheA	P2
1FFS	B,D	*E. coli*	CheA	P2
1FFW	B,D	*E. coli*	CheA	P2
1A0O	A,C,E,G	*E. coli*	CheY	
1BDJ	A	*E. coli*	CheY	
1CHN	A	*E. coli*	CheY	
1F4V	A,B,C	*E. coli*	CheY	
1FFG	A,C	*E. coli*	CheY	
1FFS	A,C	*E. coli*	CheY	
1FFW	A,C	*E. coli*	CheY	
1FQW	A,B	*E. coli*	CheY	
1HEY	A	*E. coli*	CheY	
1JBE	A	*E. coli*	CheY	
1KMI	Y	*E. coli*	CheY	
1ZDM	A,B	*E. coli*	CheY	
2B1J	A,B	*E. coli*	CheY	
3CHY	A	*E. coli*	CheY	
1KMI†	Z	*E. coli*	CheZ	
1QU7	B	*E. coli*	MCP (Tsr)	Cytoplasmic domain
1I5N	A,B,C,D	*S. typhimurium*	CheA	P1
1A2O	A,B	*S. typhimurium*	CheB	
1CHD	A	*S. typhimurium*	CheB	C-terminal catalytic domain
1AF7	A	*S. typhimurium*	CheR	
1BC5	A	*S. typhimurium*	CheR	
2CHE	A	*S. typhimurium*	CheY	
2CHF	A	*S. typhimurium*	CheY	
2FKA	A	*S. typhimurium*	CheY	
2FLK	A	*S. typhimurium*	CheY	
2FLW	A	*S. typhimurium*	CheY	
2FMF	A	*S. typhimurium*	CheY	
2FMH	A	*S. typhimurium*	CheY	
2FMI	A	*S. typhimurium*	CheY	
2FMK	A	*S. typhimurium*	CheY	
2PL9	A,B,C	*S. typhimurium*	CheY	
2PMC	A,B,C,D	*S. typhimurium*	CheY	
1TQG	A	*T. maritima*	CheA	P1
1U0S	A	*T. maritima*	CheA	P2
2CH4	A,B	*T. maritima*	CheA	P4, P5 (Residues 355–671)
1XKR	A	*T. maritima*	CheC	
2F9Z	A,B	*T. maritima*	CheC	
2F9Z	C,D	*T. maritima*	CheD	
2CH4	W,Y	*T. maritima*	CheW	
1SQU	A,B	*T. maritima*	CheX	
1XKO	A,B	*T. maritima*	CheX	
1TMY	A	*T. maritima*	CheY	
1U0S	Y	*T. maritima*	CheY	
2TMY	A	*T. maritima*	CheY	
3TMY	A,B	*T. maritima*	CheY	
4TMY	A,B	*T. maritima*	CheY	
1LKV	X	*T. maritima*	FliG	C-terminal domain (Residues 104–335)
1QC7	A,B	*T. maritima*	FliG	C-terminal domain
2HP7	A	*T. maritima*	FliM	CheC-like domain
1O6A	A,B	*T. maritima*	FliN	C-terminal domain (Residues 59–154)
1YAB	A,B	*T. maritima*	FliN	Residues 68–154
2CH7	A,B	*T. maritima*	MCP	Cytoplasmic domain

ドッキングの予測構造数のパラメータ t は，ベンチマークデータで良い結果を得た $t = 3$ を用いた．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合せについてもすべて評価値を計

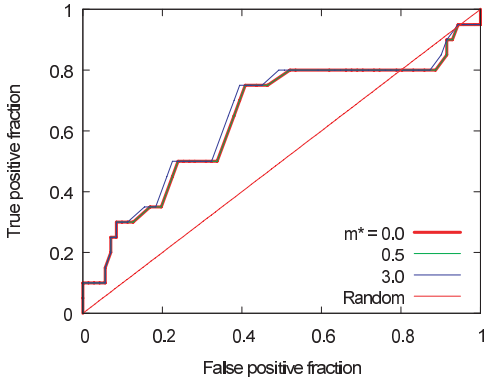

図 9　本システムの走化性系への適用における精度．m^* と E^* を変化させて行った実験における ROC 曲線を示す．縦軸が True positive fraction（$\frac{\mathrm{TP}}{\mathrm{TP+FN}}$）を，横軸が False positive fraction （$\frac{\mathrm{FP}}{\mathrm{FP+TN}}$）を示す．予測がランダムに行われた場合は対角線のようになる

Fig. 9　Evaluation of the docking post-processing system. The ROC curves for varying the threshold m^* and E^* values are shown. X-axis is for the false positive fraction ($\frac{\mathrm{FP}}{\mathrm{FP+TN}}$) and y-axis is for the true positive fraction ($\frac{\mathrm{TP}}{\mathrm{TP+FN}}$). Random prediction is indicated by the diagonal.

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 9 に，走化性系データへの適用における PPI 検出性能を示す．**図 10**，**表 10** 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をそのまま用いた．実際に相互作用が確認されている組合せを灰色のセルで示した．得られた F 値は 0.464 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 0.49 を得たと報告されており[22]，我々の手法がほぼ同等の精度で予測可能であることを示している．

提案システムの目的は PPI の一次スクリーニングであり，この過程で絞り込んだ PPI 候補はさらに詳細なエネルギー計算や実験によって解析されることが想定される．今回は偽陽性と判定した未確認の PPI のうち，存在が確認されれば生物学的意義があると思われる組合せの例として，CheY–CheD の相互作用（$E = 6.05$）をあげる．CheC は CheY の脱

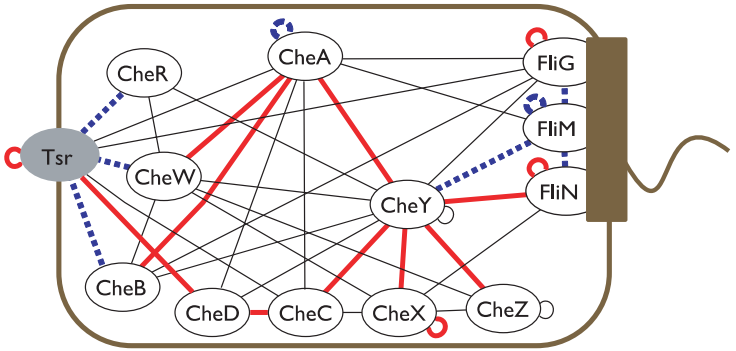

図 10　本システムにより予測されたタンパク質間相互作用．クラスタリングを行わず，$E^* = 5.5$ とした際の結果を示す．赤い太線は正しく予測された相互作用（True Positives）を，青い点線は既知の相互作用のうち検出できなかったもの（False Negatives）を，細い線は予測されたがこれまでには報告されていない相互作用（False Positives）を示す

Fig. 10　Results of the PPI predictions from the proposed system with $E^* = 5.5$, yielded without clustering. The red bold lines (true positives), blue dashed lines (false negatives) and thin lines (false positives) representing the predicted or known PPIs show the relevance of the predictions.

表 10　予測されたタンパク質間相互作用．クラスタリングを行わず，$E^* = 5.5$ 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 10　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	∗		∗	∗						∗			∗

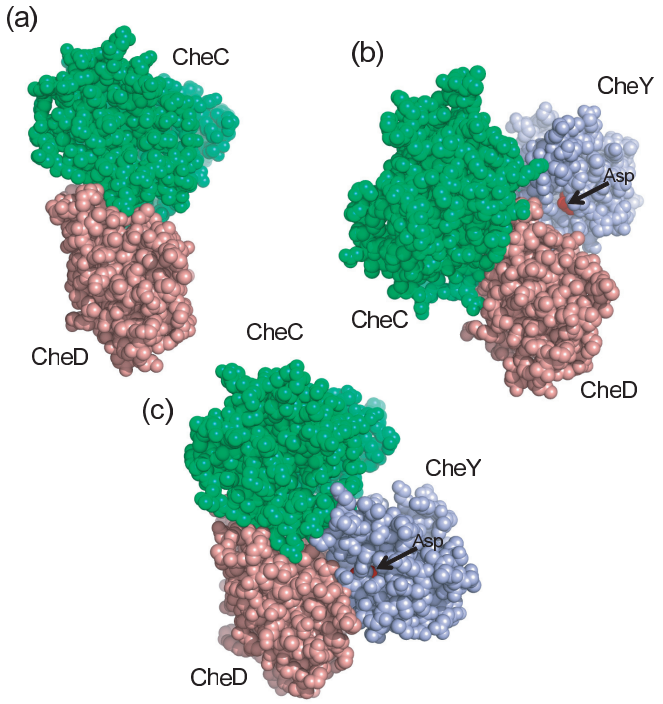

図 11　(a) 既知の CheC–CheD 複合体の結晶構造（PDB ID: 2F9Z, chain a, c, *T. maritima*）．(b) 予測された CheY（PDB ID: 1A0O, chain c, *E. coli*）–CheD（PDB ID: 2F9Z, chain c）の複合体をドッキングのレセプタとして，CheC（PDB ID: 1XKR, chain a, *T. maritima*）をドッキングさせたときの予測．CheY のリン酸化部位を赤色で示す．レセプタとして用いたのは，CheY–CheD の予測結果のうち最も高いドッキングスコアを示した複合体構造である．(c) CheC–CheD 複合体の結晶構造（2F9Z）をドッキングのレセプタとして，CheY（PDB ID: 1F4V, chain c, *E. coli*）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. 11　(a) Known structure of the CheC–CheD complex (PDB ID: 2F9Z, chains a, c, *T. maritima*). (b) Docking of CheY (PDB ID: 1A0O, chain c, *E. coli*) – CheD (PDB ID: 2F9Z, chain c) hypothetical complex and CheC (PDB ID: 1XKR, chain a, *T. maritima*). The phosphorylation site of CheY is colored red. The docking prediction with the highest E value among all the combinations of the hypothetical complex and CheC structure data is shown. (c) Docking of a known structure of the CheC–CheD complex (PDB ID: 2F9Z, chains a, c) and CheY (PDB ID: 1F4V, chain c, *E. coli*). 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.

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[35),36)]．CheY–CheC，CheC–CheD の相互作用は実験的に確認されているが[36),37)]，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．CheD は受容体（Methyl-accepting Chemotaxis Proteins，MCP）のメチル化状態を変化させる酵素であり，CheC は CheD が MCP と相互作用する部位を覆う位置に結合する．Chao らは，CheC と CheD が存在し，CheY のリン酸化レベルが高い場合には，MCP とリン酸化された CheY（CheYp）が CheD の作用を取り合い，CheYp レベルが直接 MCP のメチル化レベルを制御するというメカニズムを考察している[38)]．この提案が正しければ，CheY と CheD の間に直接の相互作用関係があり，CheY–CheC–CheD の相互作用が調節されている可能性も考えられる．CheD 存在下で CheC の脱リン酸化活性が上がるという事実は，CheD が CheY と結合して，CheC と正しい位置で結合するように導くというモデルによっても説明できる．このような動的なモデルの検討には，分子動力学法によるシミュレーションなどの解析が必要である．**図 11** (b) に，予測された CheY–CheD 複合体の構造に CheC をドッキングした結果を示す．この 3 タンパク質複合体のドッキングで得た PPI 評価値は $E = 5.44$ であった．比較のため，図 11 (c) に既知の CheC–CheD 複合体の結晶構造と CheY のドッキング結果を示す．このドッキングで得た PPI 評価値は $E = 5.44$ であり，仮の CheY–CheD 複合体をもとにした CheC とのドッキングにおける PPI 評価値と同等であった．なお，どちらの構造でも CheY のリン酸化部位は CheC の活性部位の近くには位置していなかった．図 11 にはそれぞれのドッキング結果について PPI 評価値が最も高かった複合体構造を示したが，予測した新規 PPI の妥当性を構造レベルで議論するためには，高いドッキングスコアを持つ複数の予測複合体構造を対象とし，より詳細な結合エネルギーの評価方法を適用するなどの解析を行う必要があると思われる．

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 0.464 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 の組合せを得た．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[39),40)] 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) Katchalski-Katzir, E., Shariv, I., Eisenstein, M., et al.: Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques, *Proc. National Academy of Sciences of the United States of America*, Vol.89, No.6, pp.2195–2199 (1992).
5) Gabb, H.A., Jackson, R.M. and Sternberg, M.J.E.: Modelling Protein Docking using Shape Complimentarity, Electrostatics and Biochemical Information, *J. Mol. Biol.*, Vol.272, pp.106–120 (1997).
6) Chen, R. and Weng, Z.: Docking Unbound Proteins Using Shape Complementarity, Desolvation, and Electrostatics, *Proteins*, Vol.47, pp.281–294 (2002).
7) Chen, R. and Weng, Z.: A Novel Shape Complementarirty Scoring Function for Protein-Protein Docking, *Proteins*, Vol.51, pp.397–408 (2003).
8) Chen, R., Li, L. and Weng, Z.: ZDOCK: An Initial-stage Protein-Docking Algorithm, *Proteins*, Vol.52, pp.80–87 (2003).
9) Mintseris, J., Pierce, B., Wiehe, K., Anderson, R., Chen, R. and Weng, Z.: Integrat-

ing Statistical Pair Potentials into Protein Complex Prediction, *Proteins*, Vol.69, pp.511–520 (2007).
10) Akiyama, Y., Sato, T., Matsuzaki, Y. and Matsuzaki, Y.: MEGADOCK — A rapid screening system for all-to-all protein docking analysis with pre-calculated Fourier library of protein structures, *Proc. 2008 Annual Conference of the Japanese Society for Bioinformatics*, P-032 (2008).
11) 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).
12) 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).
13) 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).
14) Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H. and Shindyalov, I.N.: The Protein Data Bank, *Nucleic Acids Research*, Vol.28, No.1, pp.235–242 (2000).
15) RCSB: Protein Data Bank. available from http://www.rcsb.org/pdb/ (accessed 2010-04-21)
16) Brooks, B.R., Bruccoleri, 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).
17) Lattman, E.E.: Optimal sampling of the rotation function, *Acta Crystallographica Section B*, Vol.28, pp.1065–1068 (1972).
18) Pierce, B. and Weng, Z.: ZRANK: Reranking Protein Docking Predictions with an Optimized Energy Function, *Proteins*, Vol.67, No.4, pp.1078–1086 (2007).
19) Zhang, C., Vasmatzis, 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).
20) Matsuzaki, Y., Matsuzaki, Y., Sato, T. and Akiyama, Y.: Development of post-docking system for protein-protein interation prediction, *1st Joint Workshop on Computational Science* (2008).
21) 松崎裕介，松崎由理，佐藤智之，秋山　泰：タンパク質間相互作用予測のためのドッキング後処理システムの開発，情報処理学会研究報告，Vol.2008-BIO-013, No.5, pp.17–20 (2008).
22) 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).
23) 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).
24) Yoshikawa, T., Seno, S., Takenaka, Y. and Matsuda, H.: Improved Prediction Method for Protein Interactions Using Both Structural and Functional Characteristics of Proteins, *IPSJ Trans. Bioinformatics*, Vol.3, pp.10–23 (2010).
25) 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).
26) 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).
27) Baker, M.D., Wolanin, P.M. and Stock J.B.: Systems biology of bacterial chemotaxis, *Curr. Opin. Microbiol.*, Vol.9, pp.187–192 (2006).
28) Kentner, D. and Sourjik, V.: Dynamic map of protein interactions in the *Escherichia coli* chemotaxis pathway, *Mol. Syst. Biol.*, Vol.5, p.238 (2009).
29) 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).
30) 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).
31) Hue, M., Riffle, M., Vert, J.P. and Noble, W.S.: Large-scale prediction of protein-protein interactions from structures, *BMC Bioinformatics*, Vol.11, No.1, pp.144–152 (2010).
32) Kanehisa, M. and Goto, S.: KEGG: kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.*, Vol.28, pp.27–30 (2000).
33) 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).
34) 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).
35) 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).
36) 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).
37) 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).
38) Chao, X., Muff, T.J., Park, S.Y., Zhang, S., Pollard, A.M., Ordal, G.W., Bilwes, A.M. and Crane, B.R.: A receptor-modifying deamidase in complex with a signaling phosphatase reveals reciprocal regulation, *Cell*, Vol.124, pp.561–571 (2006).
39) 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).
40) 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).

(平成 22 年 4 月 22 日受付)
(平成 22 年 6 月 16 日再受付)
(平成 22 年 6 月 30 日採録)

大上　雅史（学生会員）
昭和 62 年生．平成 21 年東京工業大学工学部情報工学科卒業．現在，東京工業大学大学院情報理工学研究科計算工学専攻修士課程在学中．バイオインフォマティクスの研究に従事．平成 22 年情報処理学会バイオ情報学研究会学生奨励賞受賞．日本バイオインフォマティクス学会，日本データベース学会各学生会員．

松崎　由理（正会員）
昭和 50 年生．平成 18 年慶應義塾大学大学院政策・メディア研究科博士課程修了．博士（学術）．同大学研究員を経て，現在，東京工業大学産学官連携研究員．タンパク質間相互作用，細胞内シグナル伝達系の研究に従事．

松崎　裕介（正会員）
昭和 60 年生．平成 22 年東京工業大学大学院情報理工学研究科計算工学専攻修士課程修了．同年株式会社 NTT データ入社．

佐藤　智之
昭和 41 年生．平成 3 年東京都立大学大学院理学研究科物理学専攻修士課程修了．同年株式会社富士総合研究所（現，みずほ情報総研株式会社）入社．主としてバイオインフォマティクス，計算化学に関連したプログラムの開発・並列処理，調査・解析に従事，現在に至る．サイエンスソリューション部シニアコンサルタント．

秋山　　泰（正会員）
昭和 36 年生．平成 2 年慶應義塾大学大学院理工学研究科博士課程修了．工学博士．同年電子技術総合研究所研究官，平成 4 年京都大学化学研究所助教授，平成 8 年新情報処理開発機構研究室長，平成 13 年産業技術総合研究所生命情報科学研究センター長．平成 19 年東京工業大学大学院情報理工学研究科教授．生命情報科学，並列処理応用に従事．日本バイオインフォマティクス学会，人工知能学会，日本薬物動態学会各会員．