DEIM Forum 2010 E4-4

エネルギー計算に基づくリランキングとクラスタリングを用いた網羅的タンパク質間相互作用予測手法の提案

大上　雅史† 　松崎　裕介† 　松崎　由理† 　佐藤　智之†† 　秋山　泰†

† 東京工業大学 大学院情報理工学研究科計算工学専攻　〒 152–8550 東京都目黒区大岡山 2–12–1–W8–76
†† みずほ情報総研株式会社　〒 101–8443 東京都千代田区神田錦町 2–3
E-mail: †{ohue, matsuzaki, y_matsuzaki}@bi.cs.titech.ac.jp, akiyama@cs.titech.ac.jp
††satou_t@bi.cs.titech.ac.jp

あらまし　タンパク質間相互作用 (Protein-Protein Interaction : PPI) 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群から網羅的に相互作用を予測するために，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．本研究ではドッキング予測構造の持つエネルギー値によるリランキングを行い，ドッキング予測精度を向上させ，その技法を 44×44 のタンパク質間相互作用の総当たり解析に応用して性能を評価した．その結果，従来よりも高精度な網羅的相互作用予測が可能になったことを示す．

キーワード　タンパク質間相互作用，タンパク質ドッキング，構造生物学，ハイパフォーマンス・コンピューティング

Proposal of an all-to-all protein-protein interaction prediction technique using clustering and re-ranking based on energy calculations

Masahito OHUE†, Yusuke MATSUZAKI†, Yuri MATSUZAKI†,

Toshiyuki SATO††, and Yutaka AKIYAMA†

† Graduate School of Information Science and Engineering, Tokyo Institute of Technology
Ookayama 2–12–1–W8–76, Meguro-ku, Tokyo, 152–8550 Japan
†† Mizuho Information & Research Institute, Inc.
Kanda-Nishikicho 2–3, Chiyuda-ku, Tokyo, 101–8443 Japan
E-mail: †{ohue, matsuzaki, y_matsuzaki}@bi.cs.titech.ac.jp, akiyama@cs.titech.ac.jp
††satou_t@bi.cs.titech.ac.jp

Abstract　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. To predict all-to-all PPI from protein structures, we have been studying the protein docking approach based on the physical/chemical properties and shape complementarity. In this study, we employ a re-ranking method at the docking structure ranking stage based on energy values of decoys and have succeeded to improve the accuracy of docking prediction. In addition, we applied the proposed method to predict PPIs in an all-to-all fashion. As a result of application to 44×44 all-to-all PPI analysis, we obtained an improved performance than previous methods.

Key words　protein-protein interaction, protein docking, structure biology, high-performance computing

1. はじめに

タンパク質間相互作用 (Protein-Protein Interaction: PPI) は生命現象において中心的な役割を果たしていることが近年解明されつつあり，この相互作用ネットワークの解明は，細胞内のシグナル伝達経路の特定や，それをターゲットとした創薬に対する重要課題となっている．相互作用予測は，タンパク質のアミノ酸配列モチーフを用いた手法 [1]〜[3] や，既知のドメイン間相互作用情報などのデータベースに基づく手法 [4], [5] などがよく用いられていたが，近年タンパク質の 3 次元構造が次々

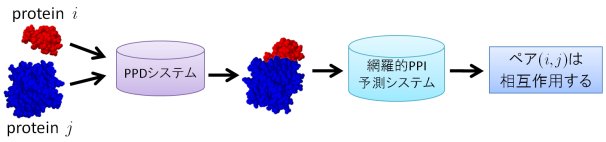

図 1　タンパク質ドッキングによる網羅的 PPI 予測

と決定されており，またタンパク質立体構造が複合体形成に深く関わっていることが分かってきたため，今後は豊富に得られつつある 3 次元情報を利用した手法が主流になると考えられている．3 次元情報を利用した相互作用予測手法は既にいくつか存在するが，自由エネルギー変化の推定など厳密に物理化学的な量を見積もろうとするアプローチでは，タンパク質 1 対 1 の相互作用予測にも数日から数週間かかるものがほとんどであった．また，相互作用ネットワーク解明のためには複数のタンパク質群の相互作用予測シミュレーションを網羅的に行う必要があり，予測計算回数は数万回から数百万回にのぼるため，現実の時間内でネットワーク予測を行うことは不可能であると見られてきた．しかし比較的計算時間の小さい形状相補性に基づくタンパク質ドッキング (Protein-Protein Docking: PPD) を利用することで，網羅的 PPI 予測を現実的な時間で行うことが可能となった．ここで問題となるのは，PPD システムが本来複合体を形成しないタンパク質ペアも，また相互作用をするタンパク質ペアも同等にドッキング計算を行い，システムが尤もらしいとする複合体構造を出力する点である．すなわち，出力されたドッキング構造から相互作用をするかしないかの判断を行う必要があるということであり，その判定を行う PPI 予測システムの確立が求められる（図 1）．

2. 手　　法

2.1 タンパク質ドッキング

本稿が指すタンパク質ドッキング (Protein-Protein Docking: PPD) とは，タンパク質を剛体とみなし，複合体形成の際にタンパク質構造が変化しないという仮定の元，タンパク質を 3 次元ボクセル空間上で表わして表面形状の相補性を主として計算するシミュレーション手法である．これまでに MolFit [6] や FTDock [7]，ZDOCK [8]〜[10] などの PPD ソフトウェアが開発され，現在まで用いられてきたが，いずれも網羅的 PPI 予測を目的としたものではないこともあり，我々は網羅的 PPI 予測用ドッキングシステムとして独自に “MEGADOCK” [11]〜[13] というシステムを開発してきた．以下ではその現行バージョンである MEGADOCK Ver.2.1 [12] について示す．

2.2 MEGADOCK Ver.2.1

MEGADOCK Ver.2.1 のドッキングは形状相補性と静電的相互作用の計算からなる．MEGADOCK で用いている PPD の良さを表すスコア $\boldsymbol{S}$ は以下の式で与えられる．

$$\boldsymbol{R}(l,m,n) = G_R(l,m,n) + iE_R(l,m,n) \tag{1}$$

$$\boldsymbol{L}(l,m,n) = G_L(l,m,n) + iwE_L(l,m,n) \tag{2}$$

$$\boldsymbol{S}(\alpha,\beta,\gamma) = \Re\left[\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N}\boldsymbol{R}(l,m,n)\boldsymbol{L}(l+\alpha,m+\beta,n+\gamma)\right] \tag{3}$$

$\boldsymbol{R}$ はレセプタータンパク質を，$\boldsymbol{L}$ はリガンドタンパク質を $N \times N \times N$ の 3 次元ボクセルに分割したときの離散関数で，(l,m,n) はそのボクセルの座標である．それぞれの離散関数は形状相補性による項 (G) と静電的相互作用による項 (E) で表わされ，スコア $\boldsymbol{S}$ は (α,β,γ) をリガンドの平行移動ベクトルとした相関関数（の実部）として計算される．w は静電的相互作用の重みを表すパラメータである．形状相補性による項 (G) は，以下で表わされる real Shape Complementarity(rPSC) スコア [12] を用いている．

$$G_R(l,m,n) = \begin{cases} \#\text{ of atoms within } (D + r_{vdW}) & \text{open space} \\ 3\rho & \text{solvent excluding surface of the R} \\ 9\rho & \text{core of the R} \end{cases} \tag{4}$$

$$G_L(l,m,n) = \begin{cases} 0 & \text{solvent accessible surface layer of the L} \\ 1 & \text{solvent excluding surface layer of the L} \\ \delta & \text{core of the L} \\ 0 & \text{open space} \end{cases} \tag{5}$$

ただし，ρ, δ はマッチや衝突のペナルティの大きさを決めるパラメータであり，D は距離のカットオフパラメータである．Ver.2.1 では $\rho = -3, \delta = 2, w = 2000, D = 3.6$Å を用いている．R と L はレセプタータンパク質，リガンドタンパク質を示す．また，静電的相互作用による項は以下のように計算される．ボクセル $i(l,m,n)$ に対する電界 ϕ_i を，

$$\phi_i = \sum_j \frac{q_j}{\varepsilon(r_{ij})r_{ij}} \tag{6}$$

$$\varepsilon(r) = \begin{cases} 4 & (r \leqq 6\text{Å}) \\ 38r - 224 & (6\text{Å} < r < 8\text{Å}) \\ 80 & (r \geqq 8\text{Å}) \end{cases} \tag{7}$$

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(注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 [14] 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，$E_R(l,m,n), E_L(l,m,n)$ を，

$$E_R(l,m,n) = \begin{cases} \phi_{i(l,m,n)} & \text{entire voxel excluding core} \\ 0 & \text{core of the R} \end{cases} \tag{8}$$

$$E_L(l,m,n) = q(l,m,n) \tag{9}$$

とする．

MEGADOCK ではこのドッキングスコア $\boldsymbol{S}$ をリガンドを回転・移動させながら計算していく．1 つのリガンド回転角に対して複合体の平行移動 α, β, γ を，ボクセルサイズ N に対して N^3 通り探索し，それを 15° 刻みで 3600 通り行う．よって，1 つの複合体について計算対象となるドッキング結合部位とリガンド回転パターンの組み合わせは，$3600 \times N^3$ 通りとなる．この中から R. Chen らが開発した ZDOCK [10] では，各回転角ごとに最良のドッキングスコアを与えるもの 1 つを暫定解として，計 2000 個の予測をドッキングスコア順に順位付けして出力するが，MEGADOCK では後のリランキングやクラスタリングと組み合わせることを考えて，各角度ごとに t 個報告可能とした．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)

(注1)：2Å 以下の場合は 2Å と定める．

を用いて，

$$\boldsymbol{S}(\alpha,\beta,\gamma) = \mathrm{IFT}\,[\mathrm{DFT}[\boldsymbol{R}(l,m,n)]^{*}\mathrm{DFT}\,[\boldsymbol{L}(l,m,n)]] \qquad (10)$$

とし，高速フーリエ変換を用いることで $O(N^3 \log N)$ となる．z^* は z の複素共役を表す．

2.3 リランキングシステム

ドッキングシステムは予測複合体を 1 つではなく，$3600 \times N^3$ 通りの解に順位付けをしてその上位となるものを複数個（例えば 2000 個）出力する．それらの中には，順位は良いが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には様々な方法が考えられるが，タンパク質複合体の結合エネルギーを 1 つ 1 つ計算して，その結果に基づいて再度順位付けを行っていく方法が最善である．しかし，厳密な結合エネルギーの計算はタンパク質のような巨大分子では 1 個につき数時間〜数日のオーダで時間がかかるため現実的ではない(注2)．ZDOCK も同様の問題を抱えているが，ZDOCK に対しては B.Pierce らが開発したリランキングシステム ZRANK [15] によって解決を図っている．ZRANK は予測された 2000 個の複合体構造に対してファン・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(注3)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，高速にリランキングを行うものであり，その評価値は

$$E_{zrank} = w_1 E_{vdW} + w_2 E_{Elec} + w_3 E_{DS} \qquad (11)$$

と表される．ZRANK により計算された評価値 E_{zrank} 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（正解でない複合体構造）が除去されることによる精度の向上が期待できる．

2.4 クラスタリング

Yusuke Matsuzaki ら [17], [18] は ZDOCK や MEGADOCK の出力する膨大な予測複合体に対して，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 Post Docking System(PDS) を開発している．各予測構造間の距離には Yusuke Matsuzaki らの提案した構造間非類似度 D を用いる．非類似度は，

$$D = D_L + D_R \qquad (12)$$

と定義される．2 つの予測構造間でレセプターをフィッティングさせ，そのときのリガンドの中心座標を $\boldsymbol{x}_{L1}, \boldsymbol{x}_{L2}$ とするとき，リガンドに関する予測構造間距離 D_L は 3 次元ユークリッド距離 $|\boldsymbol{x}_{L1} - \boldsymbol{x}_{L2}|$ で計算される．

またリガンドを回転させずにフィッティングし，その中心座標 $\boldsymbol{x}_{L1}$ を中心に (ψ,θ,φ)[rad] 回転を行った座標 $\boldsymbol{x}_{R1}$ を

$$\boldsymbol{x}_{R1} = \boldsymbol{R}_z(\psi)\boldsymbol{R}_x(\theta)\boldsymbol{R}_z(\varphi)(\boldsymbol{x}_{R1} - \boldsymbol{x}_{L1}) + \boldsymbol{x}_{L1} \qquad (13)$$

と計算し，同様に $\boldsymbol{x}_{R2}$ を計算する．ただし，$\boldsymbol{R}_i(j)(i \in \{x,y,z\}, j \in \{\psi,\theta,\varphi\})$ は軸 i の周りに角度 j[rad] だけ回転させる回転行列である．これらを用いて，D_R はユークリッド距離 $|\boldsymbol{x}_{R1} - \boldsymbol{x}_{R2}|$ で計算される．

以上の操作によって得られた D_L, D_R を用いて構造間非類似度 D を計算する．この D を全構造間で求め，その値の小さな構造をまとめながらクラスタリングを行っていく．

2.5 タンパク質間相互作用予測

PPI 予測を行うフローを以下に示す．

（1） MEGADOCK により all-to-all ドッキングを行う．出力する複合体数は各角度ごとに t 個とし，予測数は $2000 \times t$ 個とする．

（2） ZRANK によりリランキングを行う．$2000 \times t$ 個の予測をすべて ZRANK にかけ，リランキングされた予測構造のうち上位 2000 個を取り出す．

（3） 抽出された上位 2000 個の予測に対し，PDS によるクラスタリングを行う．クラスタリングにはいくつかの手法があるが，タンパク質相互作用予測には群平均法が適していることが示唆されている [18]．

（4） 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．

（a） 各クラスタ C_i の中のデータのうち，MEGADOCK によるスコアが最大のものを代表データとする．代表データについて，2000 の全候補データの MEGADOCK スコアに対する z 値（標準化係数）を s_i とする．

（b） 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．

（c） 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．

$$C' = \{C_i \mid m_i \leqq m^*\} \qquad (14)$$

$$E = \begin{cases} \max s_i \ (i \in C') & \text{if } C' \neq \emptyset \\ 0 & otherwise \end{cases} \qquad (15)$$

（5） 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．

(注2)：もちろん，全てのドッキングサイトにおける全結合構造 ($3600N^3$ 通り) についてエネルギー計算を行うのが望ましいが，当然ながら現実的ではない．しかし 2000 個などに候補を絞ることにより，先の理由で精密な計算は無理でも，少し細かいモデルによる計算であれば扱うことができる．

(注3)：ACE(Atomic Contact Energy) スコア [16] を用いている．詳しくは文献 [15] を参照されたい．

表 1　Protein-protein docking benchmark 2.0 から選出した 44 個のモノマータンパク質複合体によるサブセット

Complex PDB ID
Rigid-body(34)
1AK4, 1AVX, 1AY7, 1B6C, 1BUH, 1BVN, 1CGI, 1D6R, 1DFJ, 1E6E, 1E96, 1EAW, 1EWY, 1F34, 1FC2, 1FQJ, 1GCQ, 1GHQ, 1HE1, 1KAC, 1KTZ, 1KXP, 1KXQ, 1MAH, 1PPE, 1QA9, 1SBB, 1TMQ, 1UDI, 2BTF, 2PCC, 2SIC, 2SNI, 7CEI
Medium Difficulty(6)
1ACB, 1GRN, 1HE8, 1I2M, 1M10, 1WQ1
Difficult(4)
1ATN, 1FQ1, 1H1V, 1IBR

表 2　ZRANK による MEGADOCK2.1 のドッキング予測の正解数

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	19

表 3　ZRANK によるリランキングに要した平均計算時間 [分]

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

3. 実　　験

3.1 タンパク質ドッキング予測

3.1.1 方　　法

PPD 予測に広く用いられているベンチマークである Protein-protein docking benchmark 2.0 [19] 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 [20] ID リストを表 1 に示す．表 1 における 3 つの分類は，ベンチマークセットの作者が示しているドッキング予測の難易度の目安である．ここでは MEGADOCK の出力複合体数と，ドッキング予測の正解数との関係を調べることと，MEGADOCK に対する ZRANK の効果を検証することを目的とする．なお，評価対象とするタンパク質をモノマーに絞ることで，2 量体やそれ以上のドッキングの際の複雑な挙動の影響を考慮する必要がなくなり，ドッキングと相互作用との関係が明確になると考え，本研究ではモノマータンパク質のみを扱った．

3.1.2 結果と考察

MEGADOCK2.1 によって t の値を変えながら 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 root mean square deviation(L-RMSD) を計算した．ドッキング予測が 1 位と出力されたもので，L-RMSD が 5Å 以下となった予測複合体を正解とみなし，44 複合体中正解となった個数を表 2 に示す．またこのときの ZRANK の平均計算時間を表 3 に示す．使用した計算機は東京工業大学学術国際センター (GSIC) のスーパーコンピュータ “TSUBAME”である．表 2 より，ZRANK をかけた場合は t の値にしたがって 1 位が正解となる複合体の数が増えていることが分かる．しかし表 3 に示すように，t の値を増加させると 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 $t = 3$ が最も良いと判断した．なお，ZRANK をかけない場合は，予測数は増えるが 1 位の予測構造が変わることはなく，正解数は 19 個であった．

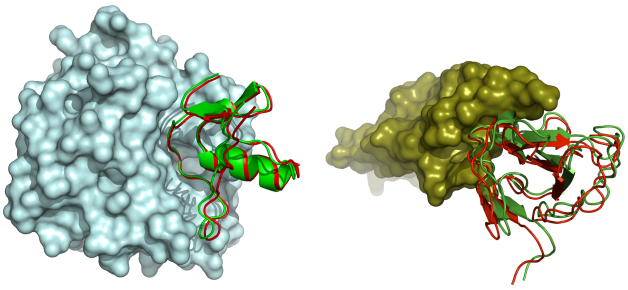

図 2　ドッキング予測によって得られた複合体構造．左は 1CGI，右は 1KTZ であり，surface 表示のタンパク質がレセプターで，リボン表示のタンパク質がリガンドである．緑は MEGADOCK のドッキング予測結果で，赤は X 線結晶構造解析によるものである．

ドッキング予測結果の構造例を図 2 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ーで，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．

3.2 網羅的タンパク質間相互作用予測

3.2.1 方　　法

ドッキング予測実験に用いたものと同一のサブセットに対して網羅的 PPI 予測を行った．ドッキング予測データは，MEGADOCK 2.1 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 $2000 \times t$ 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ものを使用する．PPI 予測フローにしたがって相互作用予測を全 1936 通り（44^2 通り）について行い，Protein Data Bank [20] 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，

$$F\text{-measure} = \frac{2 \cdot \mathrm{TP}}{(\mathrm{TP} + \mathrm{FP}) + (\mathrm{TP} + \mathrm{FN})} \tag{16}$$

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．計算時間は平均して，1 つのタンパク質ペアのドッキングに 57 分，1 つのタンパク質ペアの 2000 個のドッキング予測構造のクラスタリングに 32 分である．また，ZRANK によるリラン

表 4　44 複合体の網羅的タンパク質間相互作用予測結果

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

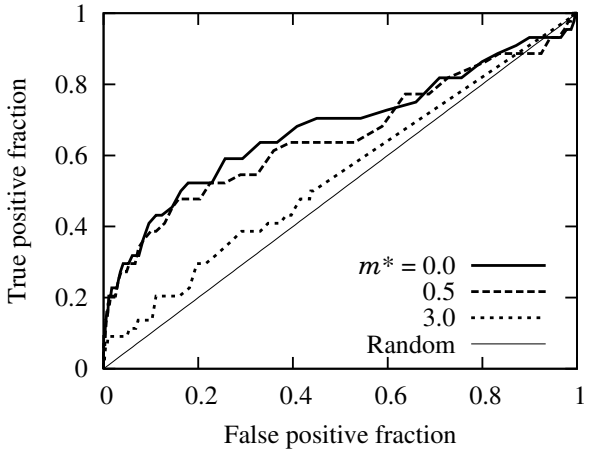

図 3　ROC 曲線（$t = 3$，ZRANK あり）

キングの計算時間は表 3 に示した通りである．

3.2.2　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 4 に示す．表 4 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 MEGADOCK 2.1 を用いて相互作用予測を行った従来手法である η-method の F 値 $= 0.150$ [12] や，同じく従来手法である Boosting を用いた学習手法による F 値 $= 0.251$ [13] を上回っており，PPI 予測精度の向上に成功していると言える．また，すべてのデータに対して ZRANK によるリランキングをかけたものの 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．

クラスタのメンバ数の閾値によって True Positive fraction と Flase Positive fraction がどのように動くかを確認するため，表 4 の中で最も良好な結果である $t = 3$ のときに ZRANK を使用したものの結果の ROC 曲線を図 3 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 3 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．また図 3 に示したように $m^* = 0.0$ 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．

3.2.3　Docking Score vs. ZRANK Score Plots

表 4 の最大 F 値をとるときの TP, FN, FP, TN からそれぞれ 2 個ずつ複合体を抽出し，ドッキングスコアの z 値と ZRANK スコアをプロットしたものを図 4 に示す．ドッキングスコアは正の方が，ZRANK スコアは負の方が良い値であり，グラフが右下に集中すると良い傾向であることがいえるが，TP サンプルは大きく右下に予測点を持つ傾向があることが分かる．それ以外の FN, FP, TN サンプルに関してはあまり大差ない結果となっており，これだけでは差異を見出すことはできないが，TP サンプルとは明らかに違う挙動を示しており，TP サンプルの予測がドッキングスコアと ZRANK スコアによって正しく行える可能性を示唆するものであるといえる．

4.　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，エネルギー計算によるリランキング (ZRANK) とドッキング予測複合体群のクラスタリング (PDS) を組み合わせた新たな予測手法の提案を行った．今後はシステム生物学の主要なモデル系の 1 つである細菌走化性系 [21], [22] に適用する予定である．この系はすでに Yuri Matsuzaki ら [23] が ZDOCK によって網羅的相互作用予測を試みているが，MEGADOCK を用いた相互作用予測によって再評価を行う．また，肺ガンと深く関わりがあるとされ，医学的にも重要視されているヒト EGFR シグナル伝達系 [24], [25] への応用を行い，新たな相互作用の発見を目指す．

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．

文　　献

[1] H.X. Zhou, Y. Shan: “Prediction of protein interaction sites from sequence profile and residue neighbor list”, *Proteins*, 44(3): 336-343, 2001.

[2] Y. Ofran, B. Rost: “Predicted protein-protein interaction sites from local sequence information”, *Federation of European Biochemical Societies Letters*, 544(1-3): 236-239, 2003.

[3] A. Koike, T. Takagi: “Prediction of protein-protein interaction sites using support vector machines”, *Protein Engineering, Design & Selection*, 17(2): 165-173, 2004.

[4] M. Deng, S. Mehta, F. Sun: “ Inferring domain-domain interactions from protein-protein interactions”, *Genome Research*, 12: 1540-1548, 2002.

[5] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N.J. Krogan, S. Chung, A. Emili, M. Snyder, J.F. Greenblatt, M. Gerstein, “A bayesian networks approach for predicting protein-protein interactions from genomic data”, *Science*, 302(5644): 449-453, 2004.

[6] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A.A. Friesem, C. Aflalo, I.A. Vakser: “Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques”, *Proceedings of the National Academy of Sciences of the United States of America*, 89(6): 2195-2199, 1992.

[7] H.A. Gabb, R.M. Jackson, M.J.E. Sternberg: “Modelling Protein Docking using Shape Complimentarity, Electrostatics and Biochemical Information”, *J Mol Biol*, 272: 106-120, 1997.

[8] R. Chen, Z. Weng: “Docking Unbound Proteins Using Shape Complementarity, Desolvation, and Electrostatics”, *Proteins*, 47: 281-294, 2002.

[9] R. Chen, L. Li, Z. Weng: “ZDOCK: An Initial-stage Protein-Docking Algorithm”, *Proteins*, 52: 80-87, 2003.

[10] R. Chen, Z. Weng: “A Novel Shape Complementarirty Scoring Function for Protein-Protein Docking”, *Proteins*, 51: 397-408, 2003.

[11] Y. Akiyama, T. Sato, Y. Matsuzaki, Y. Matsuzaki: “MEGADOCK - A rapid screening system for all-to-all protein docking analysis with pre-calculated Fourier library of protein structures”, *Proceedings of the 2008 Annual Conference of the Japanese Society for Bioinformatics*: P032,

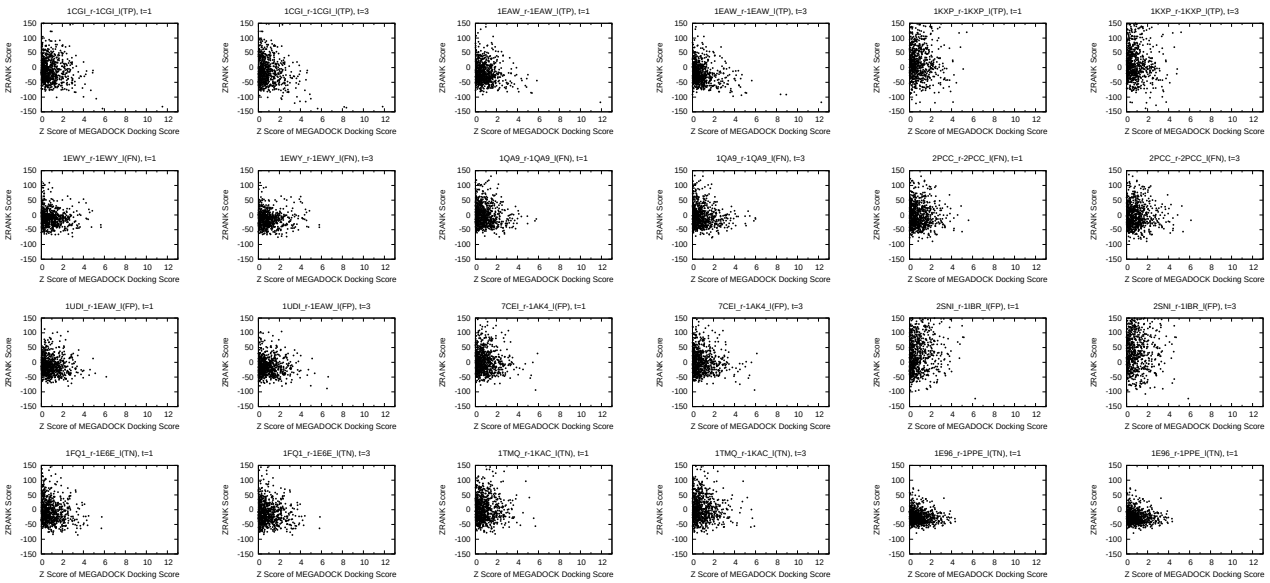

図 4　TP,FN,FP,TN の各サンプルにおけるドッキングスコア (z 値) と ZRANK スコアの比較

2008.

[12] M. Ohue, Y. Matsuzaki, Y. Matsuzaki, T. Sato, Y. Akiyama: “Improvement of all-to-all protein-protein interaction prediction system by introducing physicochemical interaction”, *IPSJ-SIG Technical Report*, 2009-BIO-17(11): 1-8, 2009.

[13] M. Ohue, Y. Matsuzaki, Y. Matsuzaki, Y. Akiyama: ”Improvement of all-to-all protein-protein interaction prediction system MEGADOCK”, *The 20th International Conference on Genome Informatics(GIW2009)*: P-033, 2009.

[14] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M. Karplus: “CHARMM: a program for macromolecular energy, minimization, and dynamics calculations”, *J Comput Chem*, 4: 187-217, 1983.

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

[16] Zhang C., Vasmatzis G., Cornette J. L., DeLisi C.: “Determination of atomic desolvation energies from the structures of crystallized proteins”, *J. Mol. Biol.*, 267: 707-726, 1997.

[17] Y. Matsuzaki, Y. Matsuzaki, T. Sato, Y. Akiyama: “Development of post-docking system for protein-protein interation prediction”, *1st Joint Workshop on Computational Science*, Saitama, Japan, 2008.

[18] Y. Matsuzaki, Y. Matsuzaki, T. Sato, Y. Akiyama: “Development of post-docking system for protein-protein interation prediction”, *IPSJ-SIG Technical Report*, 2008-BIO-13(5): 17-20, 2008.

[19] J. Mintseris, K. Wiehe, B. Pierce, R. Anderson, R. Chen, J. Janin, Z. Weng: “Protein-Protein Docking Benchmark 2.0: an update”, *Proteins*, 60(2): 214-216, 2005.

[20] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer Jr, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi, *J. Mol. Biol.*, 112: 535, 1977. `http://www.rcsb.org/pdb/`.

[21] G.H. Wadhams, J.P. Armitage: “Making sense of it all: bacterial chemotaxis”, *Nat. Rev. Mol. Cell Biol.*, 5: 1024-1037, 2004.

[22] M.D. Baker, P.M. Wolanin, J.B. Stock: “Systems biology of bacterial chemotaxis”, *Curr. Opin. Microbiol.*, 9: 187-192, 2006.

[23] Y. Matsuzaki, Y. Matsuzaki, T. Sato, Y. Akiyama: “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*, 7(6): 991-1012, 2009.

[24] N. Normanno, M.R. Maiello, A. De Luca: “Epidermal growth factor receptor tyrosine kinase inhibitors(EGFR-TKIs): simple drugs with a complex mechanism of action?”, *Journal of Cellular Physiology*, 194: 13-19, 2002.

[25] G. Selvaggi, S. Novello, V. Torri, E. Leonardo, P. De Giuli, P. Borasio, C. Mossetti, F. Ardissone, P. Lausi, G.V. Scagliotti: “Epidermal growth factor receptor overexpression correlates with a poor prognosis in completely resected non-small-cell lung cancer”, *Annals of Oncology*, 15: 28-32, 2004.