Vol.2009-BIO-17 No.11
2009/5/26

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

大上雅史[†1] 松崎裕介[†1] 松崎由理[†1]
佐藤智之[†2] 秋山泰[†1]

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 解析法の確立が求められている．我々は網羅的タンパク質ドッキングシステム “MEGADOCK” を開発してきたが，その予測精度は充分ではなく，改善が望まれていた．本稿では，実数のみで表わされる新たな形状相補性モデルである rPSC(real Pairwise Shape Complementarity) を提案し，さらに静電的相互作用を組み合わせることで，計算時間の大幅な増加を招くことなく精度の向上に成功したことを示す．

Improvement of all-to-all protein-protein interaction prediction system by introduciong physicochemical interaction

Masahito Ohue,[†1] Yusuke Matsuzaki,[†1] Yuri 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. The establishment of the all-to-all PPI analytical method is also a highly demanded task. We developed an all-to-all protein docking system “MEGADOCK” for this purpose. In this study, we propose a new shape complementarity model rPSC (real Pairwise Shape Complementarity) to improve prediction of MEGADOCK. We also added electrostatic interaction to the imaginary term of the scoring function. We successfully improved the precision without causing large increase of calculation time.

†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.

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．これまで計算機による相互作用予測手法がいくつも開発されてきたが，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．

ところが，タンパク質形状の相補性に主に注目したタンパク質ドッキングの発達によりその様相は変化しつつある．形状相補性に重きを置く方法は必ずしも精密な予測手法ではないが，1 対 1 の予測にかかる時間を数時間オーダにまで落とすことが可能であるため，多数のタンパク質群同士の網羅的な PPI 予測を現実的な時間で行えるようになった．だが，実際の PPI は形状相補性以外にも物理化学的相互作用による様々な影響が関係しており，それらを考慮しなければ PPI 予測精度は向上しない．

我々は網羅的 PPI 予測を行うために，計算時間を優先したタンパク質間ドッキングシステムである “MEGADOCK”[1] を開発してきた．その初版である MEGADOCK Ver.1.0 は，Katchalski-Katzir らの提案した形状相補性計算[2] のみを行うことで，網羅的 PPI 予測を現実的な時間で可能なものとした．しかしその予測精度は充分なものではなく，さらなる改善が求められていた．

2. タンパク質ドッキングに関する既存研究

2.1 概　　略

タンパク質ドッキングの手法としては分子動力学法 (molecular dynamics: MD) が広く用いられている．しかし MD は小さな分子のナノ秒単位での挙動を調べるだけでも何時間もシミュレーションする必要があり，タンパク質のような高分子の挙動のシミュレーションを行うためには何日も費やさねばならない．

一方，形状相補性を主とするタンパク質ドッキングは，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で，表面形状の相補性に基づい

て計算を行っていくものである．そのため MD などの精密な計算手法に比べれば，この手法は計算時間の面では非常に高速であるが，予測精度は劣ると言わざるをえない．しかし PPI は本来生物学的実験によって導き出されてきたものであり，その実験コストを考えればタンパク質ドッキングによって相互作用部位や相互作用タンパクペアをある程度見積もることができることは非常に有用である．

2.2 MolFit

MolFit[2] は E. Katchalski-Katzir らが開発したタンパク質同士の形状相補性計算に高速フーリエ変換 (FFT) を用いた最初のドッキングプログラムである．

レセプタータンパク質 **a** とリガンドタンパク質 **b** を $N \times N \times N$ の 3 次元ボクセルに分割し，ボクセルを (l, m, n) の座標で表す．次に以下の式によりスコアをそれぞれのボクセル $\overline{a}_{l,m,n}, \overline{b}_{l,m,n}$ に与える．

$$\overline{a}_{l,m,n} = \begin{cases} 1 & \text{on the surface of the R}^{\star 1} \\ \rho & \text{inside of the R} \\ 0 & \text{outside of the R} \end{cases} \tag{1}$$

$$\overline{b}_{l,m,n} = \begin{cases} 1 & \text{on the surface of the L} \\ \delta & \text{inside of the L} \\ 0 & \text{outside of the L} \end{cases} \tag{2}$$

この式 (1), (2) を以後 K-K スコアと呼ぶ．これにより複合体 **c** の形状相補性のスコアは，スコア関数 $\overline{a}, \overline{b}$ の相関関数として，以下のように計算される．

$$\overline{c}_{\alpha,\beta,\gamma} = \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} \overline{a}_{l,m,n} \times \overline{b}_{l+\alpha,m+\beta,n+\gamma} \tag{3}$$

ここで α, β, γ はレセプタータンパク質 **a** に対するリガンドタンパク質 **b** の相対的な移動によって表される空間の格子ステップ数を表す．また，ρ，δ はパラメータである$^{\star 2}$．

さて，式 (3) は $O(N^3)$ の掛け算の計算に，(α, β, γ) で可能となる平行移動パターン N^3 を乗じた $O(N^6)$ の計算が必要となる．しかしこの式は畳み込み和であるため，離散フーリエ変換 (DFT) を用いた計算が可能であり，高速フーリエ変換 (FFT) によって計算量は $O(N^3 \log N)$ まで落とすことができる．

2.3 FTDock

FTDock[3] は MolFit に用いられている K-K スコアに静電的相互作用の項を取り入れたドッキングプログラムである．

ボクセル $i(l, m, n)$ に対する電界 ϕ_i を

$$\phi_i = \sum_j \frac{q_j}{\varepsilon(r_{ij}) r_{ij}} \tag{4}$$

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 Euclid 距離，$\varepsilon(r)$ は誘電率をモデル化したもので，

$$\varepsilon(r) = \begin{cases} 4 & (r \leq 6\text{Å}) \\ 38r - 224 & (6\text{Å} < r < 8\text{Å}) \\ 80 & (r \geq 8\text{Å}) \end{cases} \tag{5}$$

として与えられる関数である．アミノ酸残基ごとにあらかじめ決められた電荷値の表に基づいて各原子に電荷を与え，タンパク質をボクセルに分割してボクセル電荷 $q_{l,m,n}$ を決定する．これらを用いてレセプタータンパク質 **a** とリガンドタンパク質 **b** の静電的相互作用スコア $E^a_{l,m,n}, E^b_{l,m,n}$ を，

$$E^a_{l,m,n} = \begin{cases} \phi_{i(l,m,n)} & \text{entire voxel excluding core} \\ 0 & \text{core of the R} \end{cases} \tag{6}$$

$$E^b_{l,m,n} = q_{l,m,n} \tag{7}$$

と定義する．以上より複合体 **c** の静電的相互作用スコアは，

$$E^c_{\alpha,\beta,\gamma} = \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} E^a_{l,m,n} \times E^b_{l+\alpha,m+\beta,n+\gamma} \tag{8}$$

となる．FTDock では，式 (8) で得られた静電的相互作用スコアを式 (3) の形状相補性スコア (K-K スコア) とともに考慮している．

2.4 ZDOCK

ZDOCK[4)–7)] は Boston 大学の Zhiping Weng らによって開発されたドッキングソフト

⋆1 R は Recepter Protein，L は Ligand Protein を表す．
⋆2 E. Katchalski-Katzir らは $\rho = -15$，$\delta = 1$ としている．

ウェアであり，現在も改良が続けられている．ZDOCK にはいくつかのバージョンがあるが，ここでは現在公開されているものの中で広く利用されている ZDOCK Ver.2.3 について述べる．ZDOCK Ver.2.3 では相互作用部位特定のためのスコア関数として，以下に示す形状相補性に基づくスコア (SC)，脱溶媒和自由エネルギーに基づくスコア (DS)，静電的相互作用に基づくスコア (ELEC) を組み合わせたものを利用している．

2.4.1 Shape Complementarity: SC

形状相補性に基づくスコアであり，3 つのスコアの中で最も重要視しているスコアである．基本的な考え方は MolFit と同様であるが，ZDOCK では少々異なったスコア関数を提案している．

レセプター，リガンドの各ボクセル (l,m,n) に対し，

$$\Re[\overline{a}_{l,m,n}] = \begin{cases} \#\text{ of R atoms within } (D + \text{R atom } r) & \text{open space} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

$$\Im[\overline{a}_{l,m,n}] = \begin{cases} 3 & \text{solvent excluding surface of protein} \\ 9 & \text{protein core} \\ 0 & \text{open space} \end{cases} \tag{10}$$

$$\Re[\overline{b}_{l,m,n}] = \begin{cases} 1 & \text{if this voxel is the nearest voxel of a L atom} \\ 0 & \text{open space} \end{cases} \tag{11}$$

$$\Im[\overline{b}_{l,m,n}] = \Im[\overline{a}_{l,m,n}] \tag{12}$$

によってスコアを与える．$\Re, \Im$ はそれぞれ複素数の実数部，虚数部をとる関数であり，D はカットオフパラメータである．これにより，複合体スコアは

$$\text{SC Score}_{\alpha,\beta,\gamma} = \Re\left[\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} \overline{a}_{l,m,n} \times \overline{b}_{l+\alpha,m+\beta,n+\gamma}\right] \tag{13}$$

と定義される．このスコア関数を PSC(Pairwise Shape Complementarity) という．レセプターに窪みがあり，リガンドの形状がそれと合致すると大きなプラスのスコアが与えられるように設計されており，K-K スコアに比べてより形状相補性に重点を置いた計算ができるようになっている．

2.4.2 Desolvation Free Energy: DS

タンパク質原子と溶媒との 2 箇所の接触を，タンパク質原子同士と水同士に置き換えたときの自由エネルギーの変化としてスコアを与える Atomic Contact Energy(ACE) スコア[8] を利用している．

2.4.3 Electrostatics: ELEC

FTDock において用いられている手法と同様であるが，電荷に CHARMM19(Chemistry at HARvard Macromolecular Mechanics)[9] を用いる点で異なっている．

2.4.4 ZDOCK Ver.3.0 について

現在 ZDOCK の最新バージョンは 3.0 である．ZDOCK Ver.3.0 では DS の代わりに IFACE（interface atomic contact energy)[7] を用いており，原子対エネルギー計算のために 6 つの FFT を追加しているため，ZDOCK Ver.3.0 は ZDOCK Ver.2.3 と比べて約 3 倍遅くなっているが，その精度は ZDOCK Ver.2.3 よりも高いと考えられる．

3. MEGADOCK〜形状相補性スコアの提案と物理化学的相互作用の導入

3.1 MEGADOCK システムの現状 (Ver.1.0)

MEGADOCK[1] は 1000×1000 の大規模ドッキング計算を現実のものにするという理念によって開発された網羅的ドッキングシステムであり，

- 形状相補性に基づくタンパク質ドッキング計算
- 高速なタンパク質ドッキング計算のためのフーリエ変換データベース
- 網羅的探索研究のための計算環境
- 網羅的探索研究のための計算結果解析

などからなる統合環境である．フーリエ変換の計算結果をデータベース化するなど，網羅的ドッキングを高速に行うための工夫が随所になされているが，核となるドッキング計算は K-K スコアによる形状相補性を用いたものであるため，精度の改善が求められている．

3.2 形状相補性スコアの提案

ZDOCK の PSC スコアは形状相補性をうまく表した画期的なスコアであり，K-K スコアに比べて各段に良い性能を示すことが分かっている．だがこのスコアは複素数で表現されているため，他の相互作用に関する計算を行う場合は独自に畳み込み和を計算することになり，3 次元複素 FFT を複数回行わねばならない．しかし，FFT の回数を増やすと計算時間が倍増してしまうため，極力抑える必要がある．

本稿では，ZDOCK の PSC スコアのアプローチから着想を得た，実数のみで表現された

Vol.2009-BIO-17 No.11
2009/5/26

新たな形状相補性スコアである **real Pairwise Shape Comprimentarity (rPSC) スコア**を提案する．rPSC スコアは以下の式で表される．

$$\overline{a}_{l,m,n} = \begin{cases} \#\text{ of R atoms within } (D + \text{R atom } r) & \text{open space} \\ 3\rho & \text{solvent excluding surface of the R} \\ 9\rho & \text{core of the R} \end{cases} \tag{14}$$

$$\overline{b}_{l,m,n} = \begin{cases} 0 & \text{solvent accessible surface layer of the L} \\ 1 & \text{solvent excluding surface layer of the L} \\ \delta & \text{core of the L} \\ 0 & \text{open space} \end{cases} \tag{15}$$

ρ と δ はパラメータである．この rPSC スコアを K-K スコアに代わって導入した MEGADOCK を，以後 **MEGADOCK Ver.2.0** と呼ぶ．

3.3 物理化学的相互作用の導入

rPSC スコアの提案によって，虚数項に物理化学的相互作用を入れることができる．具体的には，rPSC スコアを $G^a_{l,m,n}$，$G^b_{l,m,n}$ とし，物理化学的相互作用を $P^a_{l,m,n}$，$P^b_{l,m,n}$ とし，

$$\overline{a}_{l,m,n} = G^a_{l,m,n} + iP^a_{l,m,n} \tag{16}$$

$$\overline{b}_{l,m,n} = G^b_{l,m,n} + iwP^b_{l,m,n} \tag{17}$$

と表す．w は重みパラメータである．これらを用いて複合体 $\overline{c}_{\alpha,\beta,\gamma}$ を式 (3) によって求め，スコアを

$$\begin{aligned} \text{Score}_{\alpha,\beta,\gamma} &= \Re[\overline{c}_{\alpha,\beta,\gamma}] \\ &= \sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N}\left(G^a_{l,m,n}G^b_{l+\alpha,m+\beta,n+\gamma} - wP^a_{l,m,n}P^b_{l+\alpha,m+\beta,n+\gamma}\right) \end{aligned} \tag{18}$$

と定める．ただしこの方法では物理化学的相互作用は 1 つしか導入することができないので，本稿では静電的相互作用を適用する．理由としては，静電的相互作用は影響範囲が広く，PPI においても極めて重要な相互作用であると考えられていること，FTDock や ZDOCK のようにドッキングに適用した研究事例も多いことが挙げられる．静電的相互作用のスコアは FTDock の計算式 (4)-(8) を採用する．すなわち，式 (16),(17) の $P^a_{l,m,n}$，$P^b_{l,m,n}$ がそれぞれ式 (6),(7) の $E^a_{l,m,n}$，$E^b_{l,m,n}$ に置き換わる．ただし原子の電荷は CHARMM19[9] のものを利用する．この静電的相互作用を MEGADOCK Ver.2.0 に追加したものを，以後 **MEGADOCK Ver.2.1** と呼ぶ．

4. 1 vs. 1 protein docking

4.1 MEGADOCK の予測性能評価

4.1.1 方　　法

ZDOCK 開発チームが提供している PDB データセットである ZDOCK Benchmark 2.0[10] のタンパク質複合体を用いて，MEGADOCK による 1 対 1 のドッキング予測実験を行った．用いた複合体の PDB-ID を表 4.1.1 に示す．

表 1　タンパク質複合体リスト
Table 1　Protein Complex List from ZDOCK Benchmark 2.0

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

評価には実際の複合体結晶構造との Ligand Cα RMSD 値を用いる．また，ZDOCK Benchmark には複合体構造からそれぞれのタンパク質を切り離した *bound* 構造と，複合体生成前の結晶構造である *unbound* 構造の両方が用意されているので，本稿では両者の結果を示す．

比較に用いる評価指標に関しては，各複合体に関して

- **RMSD**：予測結果の上位 2000 個の中で Ligand RMSD が最も小さかったものの RMSD 値
- **Hits**：予測結果の上位 2000 個の中で RMSD が 10Å よりも小さかったものの個数
- **Rank**：RMSD が 10Å よりも小さかった予測結果の中で，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値

を求め，これらの評価値によって予測性能を比較する．比較対象は K-K スコアを用いている MEGADOCK Ver.1.0，rPSC スコアに改良した MEGADOCK Ver.2.0，rPSC スコアに静電的相互作用を加えた MEGADOCK Ver.2.1 である．ただし，パラメータは事前実験によって得られた $\rho = -3, \delta = 2, w = 2000, D = 3.6$Å を用いている．

4.1.2 結　　果

bound ドッキングの結果を表 2 に，*unbound* ドッキングの結果を表 3 に示す．*bound* において，Ver.2.0 は Ver.1.0 に比べて多くの複合体で RMSD が改善されており，また Hits の数も増加し，Rank も一部の複合体を除いては上位を占めるようになっている．Ver.2.1 は

表 2 MEGADOCK の各バージョンにおける *bound* ドッキング性能比較
Table 2 *bound* docking performance comparison of MEGADOCK

複合体	Ver.1.0			Ver.2.0			Ver.2.1		
PDB ID	RMSD	Hits	Rank	RMSD	Hits	Rank	RMSD	Hits	Rank
1ACB	10.29	0	-	1.49	8	3	1.49	8	10
1AK4	15.72	0	-	13.70	0	-	13.70	0	-
1AVX	11.79	0	-	2.56	12	7	2.56	15	4
1AY7	1.40	34	283	1.81	12	15	1.81	14	5
1B6C	15.01	0	-	1.92	10	3	1.92	9	4
1CGI	1.92	16	150	1.02	36	1	1.02	36	1
1D6R	8.90	4	22	1.85	14	2	1.85	20	3
1E96	23.75	0	-	12.64	0	-	12.64	0	-
1EAW	4.74	4	784	1.46	32	1	1.46	35	1
1EWY	8.94	3	1312	1.03	21	36	1.03	19	35
1GCQ	2.68	28	6	1.35	10	1	1.35	12	1
1GHQ	15.79	0	-	11.97	0	-	11.97	0	-
1GRN	1.89	6	644	1.42	10	2	1.42	11	2
1HE1	1.40	14	64	1.44	8	1	1.44	7	1
1KAC	6.55	1	1422	1.76	3	390	1.76	5	42
1KTZ	15.79	0	-	12.46	0	-	1.59	7	314
1PPE	2.00	154	20	1.43	111	1	1.43	103	1
1SBB	28.82	0	-	16.02	0	-	15.94	0	-
1UDI	1.57	55	2	1.31	22	2	1.31	23	2
2PCC	4.21	15	174	6.07	5	1361	5.03	11	447
2SIC	12.49	0	-	1.85	10	3	1.85	8	3
2SNI	2.14	16	657	1.61	17	2	1.61	17	2
7CEI	3.17	7	483	1.36	22	2	1.36	28	2

表 3 MEGADOCK の *unbound* 各バージョンにおけるドッキング性能比較
Table 3 *unbound* docking performance comparison of MEGADOCK

複合体	Ver.1.0			Ver.2.0			Ver.2.1		
PDB ID	RMSD	Hits	Rank	RMSD	Hits	Rank	RMSD	Hits	Rank
1ACB	5.99	15	332	8.62	3	374	8.62	4	917
1AK4	14.84	0	-	16.39	0	-	16.39	0	-
1AVX	9.11	3	53	4.14	10	21	4.14	8	213
1AY7	3.43	22	209	3.51	6	598	3.51	12	1284
1B6C	10.68	0	-	2.43	1	1662	2.43	1	889
1CGI	2.72	59	361	3.54	10	108	3.54	11	97
1D6R	9.57	1	132	3.42	9	6	3.42	9	5
1E96	20.08	0	-	10.65	0	-	7.99	1	1218
1EAW	4.31	40	15	1.24	44	2	1.24	49	2
1EWY	3.77	25	477	3.44	8	230	4.16	21	94
1GCQ	15.33	0	-	10.98	0	-	12.27	0	-
1GHQ	19.22	0	-	14.12	0	-	7.61	1	1639
1GRN	4.84	5	874	2.75	7	290	5.10	6	102
1HE1	5.67	11	42	4.00	5	2	4.00	6	2
1KAC	3.01	5	1120	5.09	4	147	3.94	5	91
1KTZ	17.82	0	-	16.10	0	-	15.98	0	-
1PPE	2.01	88	11	1.30	105	1	1.30	105	1
1SBB	30.43	0	-	17.38	0	-	17.38	0	-
1UDI	3.03	36	1	2.91	20	1	2.91	14	10
2PCC	6.70	9	175	6.59	4	916	6.59	6	206
2SIC	13.52	0	-	2.89	5	44	2.89	4	105
2SNI	9.54	1	1003	8.81	3	696	8.81	2	718
7CEI	15.97	0	-	1.84	10	14	1.84	17	1

Ver.2.0 に比べて RMSD の値はほとんど変化がないが，若干の Hits の増加と Rank の改善がみられる．また，1KTZ のような Ver.2.0 でも正解構造を予測できなかった複合体に対して，Ver.2.1 で予測に成功するなど，Ver.1.0 から Ver.2.0，そして Ver.2.1 へと改良することで，予測精度が良くなっているといえる．

unbound では，*bound* ほどうまく予測できているとは言えないが，Ver.1.0 から Ver.2.0, Ver.2.1 へと改良することで全体的に各評価値が改善されていることがわかる．特に，Ver.1.0 では予測ができなかった複合体に対して，Ver.2.0 では 1B6C, 2SIC, 7CEI が予測に成功し，Ver.2.1 ではさらに 1E96, 1GHQ の予測に成功していることから，MEGADOCK の予測性能は向上したといえる．

4.2 既存のドッキングソフトウェアとの性能比較

4.2.1 方　　法

MEGADOCK Ver.1.0，2.0，2.1 と，ZDOCK Ver.2.3，3.0 の性能を比較する．比較は Minimum RMSD–ranking グラフを用いる．これは ranking（横軸）で示した順位までの中で最小の Ligand RMSD をとるドッキング予測構造の RMSD を，表 1 の各複合体ごとに色分けしてプロットしたものである．グラフは必ず右下がりの形となるが，より ranking の小さい段階で RMSD 値が小さくなっていれば，性能が良いと判断することができる．

4.2.2 結　　果

MEGADOCK Ver.1.0，2.0，2.1 と，ZDOCK Ver.2.3，3.0 の Minimum RMSD–ranking グラフを図 1〜図 10 に示す．また，数値比較のためにこれらのグラフの AUC(Area Under the Curve) を求めたものを表 4 に示す．

図 1〜図 3 を比べると，MEGADOCK Ver.1.0 から Ver.2.0 に改良したことで多くの複合体のグラフが左下寄りになり，予測性能が向上していることがわかる．Ver.2.1 では Ver.2.0 よりも全体的に ranking の小さい段階でグラフが降下しており，性能が改善されたことがグラフより見てとれる．さらに図 4，図 5 から MEGADOCK Ver.2.1 の予測性能は ZDOCK にかなり近づいていることがわかる．図 6〜図 8 の *unbound* ドッキングでは，バージョンの違いによる性能変化が見た目では分かりにくいが，表 4 の AUC の値より MEGADOCK Ver.2.1 は Ver.1.0，Ver.2.0 よりも性能が良く，ZDOCK の値に近いものとなっていることがわかる．

Vol.2009-BIO-17 No.11
2009/5/26

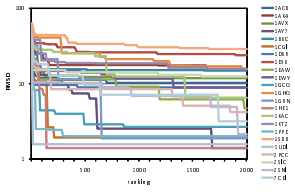

図 1 MEGADOCK Ver.1.0 の *bound* 結果
Fig. 1 *bound* docking for MEGADOCK Ver.1.0

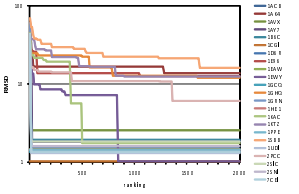

図 2 MEGADOCK Ver.2.0 の *bound* 結果
Fig. 2 *bound* docking for MEGADOCK Ver.2.0

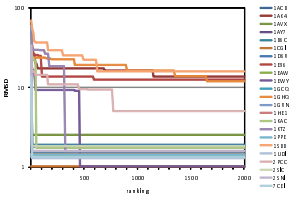

図 3 MEGADOCK Ver.2.1 の *bound* 結果
Fig. 3 *bound* docking for MEGADOCK Ver.2.1

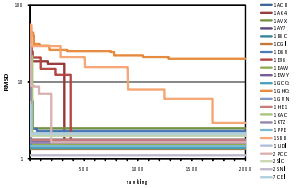

図 4 ZDOCK Ver.2.3 の *bound* 結果
Fig. 4 *bound* docking for ZDOCK Ver.2.3

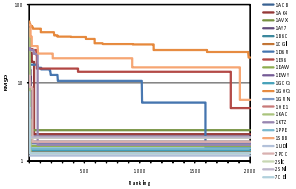

図 5 ZDOCK Ver.3.0 の *bound* 結果
Fig. 5 *bound* docking for ZDOCK Ver.3.0

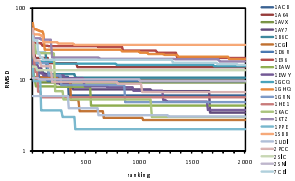

図 6 MEGADOCK Ver.1.0 の *unbound* 結果
Fig. 6 *unbound* docking for MEGADOCK Ver.1.0

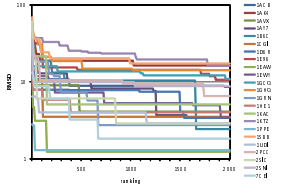

図 7 MEGADOCK Ver.2.0 の *unbound* 結果
Fig. 7 *unbound* docking for MEGADOCK Ver.2.0

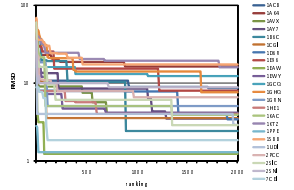

図 8 MEGADOCK Ver.2.1 の *unbound* 結果
Fig. 8 *unbound* docking for MEGADOCK Ver.2.1

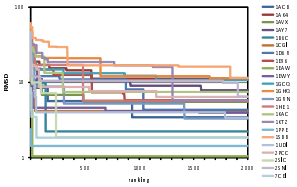

図 9 ZDOCK Ver.2.3 の *unbound* 結果
Fig. 9 *unbound* docking for ZDOCK Ver.2.3

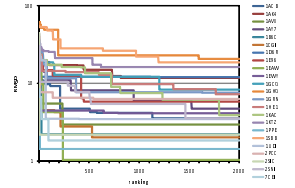

図 10 ZDOCK Ver.3.0 の *unbound* 結果
Fig. 10 *unbound* docking for ZDOCK Ver.3.0

表 4 RMSD–ranking グラフの AUC の比較
Table 4 Area under the RMSD–ranking graph curve comparison

	AUC($\times 10^{-3}$)	
	bound	*unbound*
MEGADOCK 1.0	537.6	549.6
MEGADOCK 2.0	264.0	417.2
MEGADOCK 2.1	222.3	393.3
ZDOCK 2.3	160.1	348.1
ZDOCK 3.0	211.1	380.7

4.3 1 対 1 のドッキング計算時間の比較

4.3.1 方　　法

各ドッキングシステムのドッキング計算時間を比較する．表 1 の各複合体について，1 対 1 のドッキングを行い，平均計算時間と標準偏差 σ を求め，グラフによってエラーバーとともに示す．計算機環境は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” である．

4.3.2 結　　果

平均計算時間と，$\pm\sigma$ の範囲を表すエラーバーを，図 11 に示す．図 11 から，MEGADOCK Ver.1.0 から Ver.2.0 へは計算時間の増加はほとんどないことがわかる．また，Ver.2.0 から Ver.2.1 へは静電的相互作用の一連の計算に伴う約 10 分程度の計算時間の増加が見られた．FFT の計算時間は，複合体にもよるがドッキングにかかる計算時間全体の約 80% にあたり，FFT を 2 回かける場合は計算時間が約 1.8 倍になる．MEGADOCK Ver.2.0 の計算時間の 80% はおよそ 32 分であり，FFT を 2 回かけると 32 分ほど計算時間が増加してしまうと予想されることから，MEGADOCK Ver.2.1 の 10 分の増加は深刻なものではなく，計算時間の面でも性能は良いといえる．

ZDOCK Ver.2.3 は MEGADOCK Ver.2.1 に比べ平均で 10 分ほど遅く，また複合体によるばらつきも大きいことが分かる．また，ZDOCK Ver.3.0 は ZDOCK Ver.2.3 の約 3 倍の計算時間となった．

5. 23 vs. 23 all-to-all docking

5.1 方　　法

ZDOCK Benchmark 2.0 の *bound* データセットを用いて，MEGADOCK による網羅的ドッキング実験を行った．網羅的 PPI 解析においてはタンパクペアが相互作用をしている

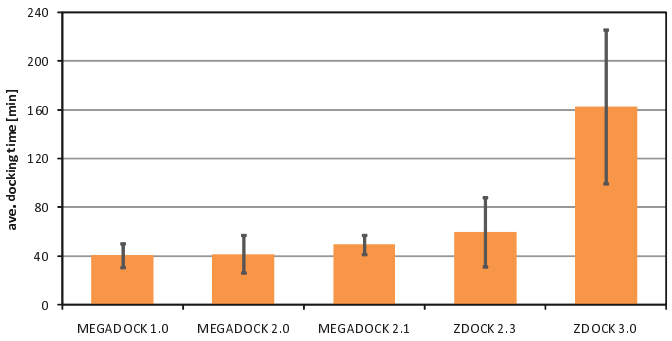

図 11 各ドッキングシステムの計算時間
Fig. 11 calculation time for docking system

かしていないかを予測することに焦点が置かれるので，ドッキングシステムの出力するスコアから相互作用の有無を判定する必要がある．本稿では，ドッキングシステムの出力する 1 位のスコアを用いて，突出して大きい値となっているタンパクペアを相互作用すると判定する．しかし，ドッキングスコアはタンパク質の大きさにある程度依存するため，タンパク質の大きさによる補正が必要となる．そこで，以下の 3 つの補正法を提案する．タンパクペア (i,j) の 1 位のドッキングスコアを S_{ij}，タンパク質 i の表面積を A_i とすると，補正スコアは

$$\eta_{ij}^{(\log)} = \frac{\log S_{ij}}{\min\{A_i, A_j\}} \tag{19}$$

$$\eta_{ij}^{(\log\log)} = \frac{\log(\log S_{ij})}{\log(\min\{A_i, A_j\})} \tag{20}$$

$$\eta_{ij}^{(2/3)} = \frac{S_{ij}^{2/3}}{\min\{A_i, A_j\}} \tag{21}$$

と表わされる．これらの補正スコアから，全 529 個の値を母集団とする z スコアを求めて，z スコアが 1.0 より大きい場合は「ペア (i,j) が相互作用する」と判定する．z スコアは，

$$z_{ij} = \frac{\eta_{ij} - \mu}{\sigma} \tag{22}$$

で求められる．μ と σ はそれぞれ全 529 個の値を母集団とする平均，標準偏差である．

23×23 個のタンパクペアのうち，PDB に登録されている 23 複合体を正例，それ以外の 506 複合体を負例とし，TP, FP, FN, TN から precision，recall，f–measure を求め，性能評価に用いる．これらの値の意味や計算式の一覧を表 5 に示す．

表 5 性能評価に用いる値一覧
Table 5 List of values to use for performance evaluation

TP(True Positive)	相互作用すると予測されたペアによる複合体が PDB に登録されている
FP(False Positive)	相互作用すると予測されたペアによる複合体が PDB に登録されていない
FN(False Negative)	相互作用しないと予測されたペアによる複合体が PDB に登録されている
TN(True Negative)	相互作用しないと予測されたペアによる複合体が PDB に登録されていない
precision(適合率)	$\text{precision} = \frac{\text{TP}}{\text{TP+FP}}$
recall(再現率)	$\text{recall} = \frac{\text{TP}}{\text{TP+FN}}$
f–measure	$f\text{−measure} = \frac{2\cdot\text{recall}\cdot\text{precision}}{\text{recall+precision}}$

5.2 結　　果

MEGADOCK Ver.1.0, 2.0, 2.1 による結果と，ZDOCK Ver.3.0 による結果を表 6 に示す．S はスコアの補正を行わずに，生のドッキングスコアのまま各評価値を求めた結果である．補正スコアの性能に関しては，f–measure から MEGADOCK Ver.1.0 と ZDOCK Ver.3.0 は S が，MEGADOCK Ver.2.0, 2.1 は $\eta_{ij}^{(2/3)}$ が最も良い結果となった．各手法の f–measure の最大値は，MEGADOCK のバージョン間で比べると Ver.1.0<Ver.2.0<Ver.2.1 といった具合に大きくなっており，性能が向上していることがわかる．また MEGADOCK Ver.2.1 の f–measure は ZDOCK Ver.3.0 にかなり近付いており，網羅的ドッキングでの性能も ZDOCK に拮抗していることがわかる．

6. おわりに

本稿では MEGADOCK システムの改良のため，従来の形状相補性スコアである K-K スコアに代わって，実数のみで表現された rPSC スコアを提案した，rPSC スコアは ZDOCK の PSC スコアが持つ表面形状の相補性を上手く表現しているところに着想を得たもので，PSC と違い実数のみで表わすことで，他の物理化学的相互作用の導入による複素 FFT の回数増加を抑えることができる利点がある．MEGADOCK Ver.2.1 では，この rPSC スコアに静電的相互作用を組み合わせることで，1 回の複素 FFT によって形状相補性計算と静電的相互作用計算を 1 度に行うことを可能とした．これによりドッキング予測性能は飛躍的に向上し，ZDOCK Ver.3.0 に迫る精度を得ることに成功した．

また目標である網羅的 PPI 予測への適用として，23×23 の網羅的タンパク質ドッキン

表 6 all-to-all ドッキングの結果
Table 6 Result of all-to-all docking prediction

	MEGADOCK Ver.1.0				MEGADOCK Ver.2.0				MEGADOCK Ver.2.1				ZDOCK Ver.3.0			
	S	$\eta^{(\log)}$	$\eta^{(\log\log)}$	$\eta^{(2/3)}$	S	$\eta^{(\log)}$	$\eta^{(\log\log)}$	$\eta^{(2/3)}$	S	$\eta^{(\log)}$	$\eta^{(\log\log)}$	$\eta^{(2/3)}$	S	$\eta^{(\log)}$	$\eta^{(\log\log)}$	$\eta^{(2/3)}$
TP	5	2	4	4	4	2	5	5	4	2	6	6	8	4	5	5
FP	69	45	72	64	77	39	61	54	76	41	59	51	74	49	85	73
FN	18	21	19	19	19	21	18	18	19	21	17	17	15	19	18	18
TN	437	461	434	442	429	467	445	452	430	465	447	455	432	457	421	433
precision	0.068	0.043	0.053	0.059	0.049	0.049	0.076	0.085	0.050	0.047	0.092	0.105	0.098	0.075	0.056	0.064
recall	0.217	0.087	0.174	0.174	0.174	0.087	0.217	0.217	0.174	0.087	0.261	0.261	0.348	0.174	0.217	0.217
f–measure	**0.103**	0.057	0.081	0.088	0.077	0.063	0.112	**0.122**	0.078	0.061	0.136	**0.150**	**0.152**	0.105	0.089	0.099

グを行った．網羅的 PPI 予測に関しては，タンパク質ドッキングによる相互作用判定や評価方法が確立していないので，本稿でもいくつかの評価方法を検討したが，充分とはいえない結果となった．しかし，MEGADOCK Ver.2.1 の rPSC と静電的相互作用の導入を行った効果は表れており，MEGADOCK Ver.1.0 に比べて精度の向上に成功しているので，網羅的 PPI 予測への実現に向けて大きな前進を遂げたといえる．

今後の課題として，網羅的ドッキングにおける相互作用判定法を確立することが挙げられる．今回は網羅的ドッキングへの試みとして 1 位のドッキングスコアのみを用いた手法を提案したが，ZDOCK の返す 2000 個の予測結果に対してクラスタリングを行うことで精度を向上させる研究事例が報告されており[11]，さらに ZDOCK を用いた相互作用ペアの予測法を実用に耐え得るレベルにまで改良し，実際の生物系に応用した研究も進んでいる[12]．これらは ZDOCK を用いたものであるが，本稿で行った MEGADOCK による予測に適用することでさらなる精度の向上が見込まれる．

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) 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*: P032，2008.
2) 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”, *Proc. Natl. Acad. Sci. USA*, 89(6): 2195-2199, 1992.
3) 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.
4) R. Chen, Z. Weng: “Docking Unbound Proteins Using Shape Complementarity, Desolvation, and Electrostatics”, *Proteins*, 47: 281-294, 2002.
5) R. Chen, L. Li, Z. Weng: “ZDOCK: An Initial-stage Protein-Docking Algorithm”, *Proteins*, 52: 80-87, 2003.
6) R. Chen, Z. Weng: “A Novel Shape Complementarirty Scoring Function for Protein-Protein Docking”, *Proteins*, 51: 397-408, 2003.
7) J. Mintseris, B. Pierce, K. Wiehe, R. Anderson, R. Chen, Z. Weng: “Integrating Statistical Pair Potentials into Protein Complex Prediction”, *Proteins*, 69(3): 511-520, 2007.
8) C. Zhang, G. Vasmatzis, J.L. Cornette, C. DeLisi: “Determination of atomic desolvation energies from the structures of crystallized proteins”, *J. Mol. Biol.*, 267: 707-726, 1997.
9) 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.
10) 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.
11) 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.
12) Y. Matsuzaki, Y. Matsuzaki, T. Sato, Y. Akiyama: “Virtual screening of protein-protein interactions: an application to known signal transduction systems”, *Proceedings of the 2008 Annual Conference of the Japanese Society for Bioinformatics*: P057，2008.