Vol.2010-BIO-20 No.4
2010/3/4

タンパク質の特性に基づく *unbound* ドッキングのための剛体予測手法の改良

松崎裕介[†1] 大上雅史[†1] 松崎由理[†1]
佐藤智之[†3] 関嶋政和[†1,†2] 秋山泰[†1]

我々が開発するタンパク質間ドッキング予測システム MEGADOCK において，そのドッキングスコアを算出する目的関数は形状の相補性と静電的相互作用を反映しているが，それぞれの項の重みは対象タンパク質によらず一定である．しかしタンパク質の形状やドッキングにおける構造変化に多様性があるため，目的関数におけるその重みをタンパク質の特性に応じて変化させることにより，予測精度の改善が期待できる．本研究では，溶媒露出面積や表面電荷，溶媒中の構造変化などの個々のタンパク質の特性に基づいて静電重みを決定し，目的関数における形状相補性の寄与を変えることで，*unbound* ドッキングの予測精度を向上させる方法を提案する．

Improvement of rigid-body prediction for *unbound* docking based on protein feature

Yusuke Matsuzaki ,[†1] Masahito Ohue ,[†1]
Yuri Matsuzaki ,[†1] Toshiyuki Sato ,[†3]
Masakazu Sekijima [†1,†2] and Yutaka Akiyama[†1]

In protein-protein docking prediction system MEGADOCK that we are developing, the target function calculating the docking scores depends on the shape complementarity and electrostatic interaction. And the weight of each term doesn't depend on the target protein and be constant. However, there are a lot of varieties in the shape of proteins and their conformational changes in the docking. Thus, the improvement of the prediction can be expected by changing this proportion in the target function based on the feature of proteins. In this study, we proposed a new method to optimize the electrostatic weight of the target function based on the feature of individual proteins such as the solvent accessible surface area, the surface charge and the structural changes, for improving the docking prediction in the *unbound* docking.

1. 序　　論

タンパク質間相互作用 (PPI: protein-protein interaction) は生命現象において中心的な役割を果たしており，その相互作用の異常が疾病の発症などに関係していることがわかってきた．従ってタンパク質が相互作用する相手，及びその複合体の構造を解析することは，医薬品の開発ターゲットを同定する過程において大きな意味を持つ．生化学的実験により複合体の構造を特定するには多くの時間と費用を要するため，近年多くの PPI 予測プログラムが開発されてきている．現在では配列モチーフを利用する方法など，様々な手法が提案されているが，非結合時の構造情報を利用して行うドッキングは，複合体構造が直接予測可能であり有用な手法である．そのドッキング予測の評価においては，既知の結合した複合体の構造を 2 つのタンパク質に分離し，それらに対してドッキング予測を適用する *bound* ドッキングと呼ばれる計算をもとに評価することが多い．しかし未知の複合体構造の特定に活用するためには，核磁気共鳴や X 線結晶解析により決定されたそれぞれの単体構造を基に複合体構造を予測する必要がある．そのため 2 つのタンパク質の構造が各々別の条件下で計測され，それらのデータ間での予測を行う *unbound* ドッキングと呼ばれる計算において，高い予測精度を実現することが求められる．

従来のドッキング予測手法では，形状相補性と静電的相互作用等の物理化学的相互作用を考慮してドッキングスコアを定義していることが多い．しかし，*bound* ドッキングと *unbound* ドッキングではそれらの間のバランスは変わりうる．例えば，ドッキング時に大きな構造変化を伴う場合，*bound* ドッキングでは相互作用面の形状が一致したとしても，*unbound* ドッキングでは構造変化のため，*bound* と同様に高い形状相補性を持つとは限らない．そのため *unbound* ドッキングでは，*bound* ドッキングと比べて形状相補性への寄与を小さくし，他の相互作用への寄与を大きくした方が良い可能性があり，その重みはタンパク質ごとで一定ではないと考えられる．すなわち，形状相補性がタンパク質の表面形状に強く依存す

†1 東京工業大学　大学院情報理工学研究科　計算工学専攻
Department of Computer Science Graduate School of Information Science and Engineering, Tokyo Institute of Technology
†2 東京工業大学　学術国際情報センター
Global Scientific Information and Computing Center, Tokyo Institute of Technology
†3 みずほ情報総研株式会社
Mizuho Information & Research Institute, Inc.

る以上，タンパク質の構造変化が大きい “柔らかい” タンパク質と，構造変化の小さい “硬い” タンパク質との間で同様の計算を行うことは適切ではないと言える．

そこで本研究では，各タンパク質の特性をもとに静電的相互作用の寄与の大きさをタンパク質ごとで変化させ，従来の剛体ドッキング予測を改良することを提案する．ドッキングソフトウェアには内部の目的関数を操作できる MEGADOCK[1] を用いる．現在，MEGADOCK においてドッキングスコアを算出する目的関数は，複数のタンパク質に対して *bound* ドッキングにてベンチマークテストを行い，その結果をもとにパラメータを最適化してある．それにより形状相補性のスコアと静電的相互作用のスコアとの関係は，対象とするタンパク質によって変わることは無く，固定されたパラメータから計算されたスコアをもとに予測結果を出力する．そのため，それらのどのタンパク質に対しても全く同様の関数によりドッキング予測を行うのではなく，その異なる性質に基づいて目的関数を変化させることでドッキング予測の精度を向上させられる可能性があると考えられる．そこで，タンパク質が持つ性質とドッキング予測時の関数との関係を抽出し，その精度をさらに向上させる方法を提案する．

2. MEGADOCK の目的関数について

網羅的 PPI 予測システムとして，当研究室で開発された MEGADOCK の目的関数について述べる．

2.1 目的関数の計算法

関数は形状相補性 (rPSC)[2] と静電的相互作用 (ELEC) についてのものである．実数部で rPSC を，虚数部で ELEC を表現することにより一度の FFT でそれらの計算が可能となっており，高速なドッキング計算を実現している．

2.1.1 real Pairwise Shape Complementarity (rPSC)

各グリッドへの rPSC スコアの与え方は以下のとおりである．レセプター **a**，リガンド **b** の各グリッド (l,m,n) に対し，rPSC スコア $G^a_{l,m,n}, G^b_{l,m,n}$ を

$$G^a_{l,m,n} = \begin{cases} \#\text{ of Recepter atoms within } (D + \text{Receptor atom radius}) & \text{open space} \\ 3\rho & \text{solvent excluding surface of protein} \\ 9\rho & \text{protein core} \end{cases} \tag{1}$$

$$G^b_{l,m,n} = \begin{cases} 0 & \text{solvent accessible surface layer of Ligand} \\ 1 & \text{solvent excluding surface layer of Ligand} \\ \delta & \text{core of Ligand} \\ 0 & \text{open space} \end{cases} \tag{2}$$

として与える．D はカットオフパラメータであり $D = 3.6(\text{Å})$ としている．また ρ, δ もパラメータであり，$(\rho, \delta) = (-3, 2)$ に定めている．

2.1.2 静電的相互作用 (ELEC)

静電的相互作用は分子間相互作用の中でも影響範囲が広く，タンパク質間相互作用においても重要な相互作用の一つであるとの考えから，物理化学的相互作用の中で最初に導入された．

グリッド $i(l,m,n)$ に対する電界 ϕ_i を

$$\phi_i = \sum_j \frac{q_j}{\varepsilon(r_{ij}) r_{ij}} \tag{3}$$

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

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

として与えられる関数である．ボクセル電荷 $q_{l,m,n}$ はアミノ酸残基ごとに決められた電荷値の表に基づいて与えられる．これらを用いてレセプター **a**，リガンド **b** の静電的相互作用スコア $E^a_{l,m,n}, E^b_{l,m,n}$ を

$$E^a_{l,m,n} = \begin{cases} \phi_{i(l,m,n)} & \text{(entire grid excluding core)} \\ 0 & \text{(core of molecule)} \end{cases} \tag{5}$$

$$E^b_{l,m,n} = q_{l,m,n} \tag{6}$$

と定義する．

2.1.3 目 的 関 数

rPSC のスコアを S_{rPSC}，静電相互作用のスコアを S_{ELEC} とすると，ドッキングスコア S は

$$S = S_{rPSC} + wS_{ELEC} \tag{7}$$

で計算される．ここで w はパラメータで，$w = 2000$ である．

Vol.2010-BIO-20 No.4
2010/3/4

表 1 データセット (Protein-Protein Docking Benchmark 2.0 より)
Table 1 Data set(from Protein-Protein Docking Benchmark 2.0).

複合体名					分類
1AK4 †	1AVX	1AY7 †	1B6C †	1BUH †	
1BVN	1CGI †	1D6R	1DFJ	1E6E	
1E96	1EAW	1EWY	1F34	1FQJ	
1GCQ	1GHQ	1HE1	1KAC	1KTZ	*Rigid-body*
1KXP	1MAH	1PPE	1QA9	1SBB	
1TMQ	1UDI	2BTF	2PCC	2SIC	
2SNI	7CEI				
1ACB †	1GRN	1I2M	1M10 †	1WQ1	*Medium Difficulty*
1ATN	1FQ1 †	1H1V			*Difficult*

† リガンド・レセプターともに分子動力学計算を実施したタンパク質.

3. 静電重み α と予測精度

2.1 で記述した通り，MEGADOCK では最終的なスコアを (7) 式により求めている．この静電的相互作用に重み α を掛けることで，形状相補性と静電的相互作用の関係を変化させる．すなわち，$\alpha > 1$ のとき静電的相互作用の寄与を重視し，$\alpha < 1$ のときは形状相補性を相対的に重視することになる．この α を本研究では以降，“静電重み” と称する．

3.1 データセット

対象とするタンパク質は Protein-Protein Docking Benchmark 2.0[3] に含まれるモノマー同士の複合体のうち 40 例を選出し，*unbound* ドッキングでの実験を行う．その複合体の名称と Protein-Protein Docking Benchmark 2.0 におけるカテゴリーを表 1 に示す．ここで *bound* ドッキングによる予測ではなく *unbound* ドッキングを対象としているのは，複合体から分離させた *bound* 構造の場合，*unbound* の場合と比べて形状相補性に依存しやすくなる懸念があるため，及び本研究の最終的な目標を *unbound* ドッキングの精度向上に定めているためである．

3.2 予測精度の評価

MEGADOCK の目的関数を改良するにあたって重要となるのは，α の値によってドッキング結果の改善が見られるか否かである．α を変化させてもドッキング予測の精度が目に見えて向上しないのであれば，余分な操作を加えて α を決定する必要は無いことになる．そこでまず本研究では，α を変化させることで実際にドッキングの予測精度が改善するのかを確認した．

ドッキング精度の良さの判断には，複合体全体の C_α 原子を対象とした RMSD(Root mean square deviation) を，予測構造と実験で得られた構造との間で計算したものを用いる．この RMSD を，デフォルトで出力される 2000 位までの予測構造について計算し，各順位までで最小の RMSD をプロットしたグラフを作成する．そして RMSD が高い順位で小さな値になるかどうかを，このグラフの AUC(Area Under the Curve) から判断する．この AUC について，$\alpha = 1$ の際の AUC に対する割合 R_{AUC} を計算し，その値から α と予測結果との関係について調べる．つまり，AUC の値は小さければ小さいほど良く，その場合 R_{AUC} の値も小さくなる．

なお，Protein-Protein docking Benchmark 2.0 から得られるタンパク質のデータは，*bound* 構造と *unbound* 構造について立体構造だけでなく配列にも差異が見られる場合がある．従って，BLAST を用いて *bound* と *unbound* の配列をアラインメントし，挿入部位と欠失部位以外の Cα 原子に限定して RMSD を計算する．すなわち置換部位や低複雑度領域の Cα も RMSD 計算に含めることとする．

結果の評価方法として AUC を選ぶ理由は 2 つある．1 つ目は，MEGADOCK のパラメータの最適化において検証した値であるため，2 つ目は特定の予測構造の精度に依存しにくく，出力された複数の予測結果に対して総合的に精度を判断できるためである．

3.3 予測順位・C_RMSD と α の関係 (AUC)

α の値を 0.01, 0.05, 0.07, 0.1, 0.2, 0.5, 1, 2, 5, 10, 15, 20, 100 と変化させたときの，順位と C_RMSD の関係を表すグラフにおける R_{AUC} を表 2 に示す．

この結果をもとに，対象としたタンパク質群を以下の 4 通りに分類する．

a) $\alpha > 1$ で AUC が減少する
b) $\alpha < 1$ で AUC が減少する
c) $\alpha \neq 1$ で AUC が増加する
d) その他

大まかにこの 4 通りに分けると，分類 (a) と (b) に属するタンパク質は以下のようになる．

a) 1AK4, 1AY7, 1BUH, 1E6E, 1KTZ, 1QA9, 2PCC, 2SNI, 1M10, 1WQ1, 1H1V
b) 2SIC

上記のタンパク質は，適切な α を選択することで結果が改善が期待できる．従って，これらのタンパク質にて適切な α が選択されるよう，α の値を定めることが重要となる．ただし (b) に分類されるタンパク質は 1 つしか Benchmark 2.0 に含まれていないため，α の決定の際には (a) の分類の精度向上を重視することとする．

表 2 各 α についての R_{AUC} の値
Table 2 The value of R_{AUC} from each α.

α	0.01	0.05	0.07	0.1	0.2	0.5	1	2	5	10	15	20	100
1AK4	0.97	0.97	0.97	0.97	0.96	0.98	1	0.99	0.93	0.9	0.92	0.87	0.79
1AVX	0.98	0.98	0.98	0.98	0.97	0.97	1	1.1	1.82	2.09	2.36	2.55	3
1AY7	1.07	1.07	1.07	1.07	1.07	1.04	1	0.84	1.04	1.2	1.09	1.07	1.35
1B6C	1.08	1.13	1.13	1.13	1.1	1.05	1	1.06	1.17	2.11	2.56	2.53	3.01
1BUH	1.09	1.09	1.09	1.09	1.09	1.06	1	0.85	0.63	0.68	0.77	0.94	0.77
1BVN	1.02	1.02	1.01	1.01	1.01	1.01	1	1	1	1.01	1.02	1.01	1.03
1CGI	1.01	1	1	1	1	1	1	1.04	1.65	2.15	2.32	2.33	2.6
1D6R	0.98	0.98	0.97	0.97	0.97	0.97	1	1.15	1.62	2.03	2.1	2.16	2.75
1DFJ	1.02	1.02	1.02	1.02	1.02	1.01	1	1	1.16	1.21	1.23	1.27	2.3
1E6E	1.05	1.05	1.05	1.05	1.05	1.04	1	0.89	0.87	0.89	0.88	0.88	1.03
1E96	1.04	1.04	1.04	1.04	1.04	1.04	1	1.07	1.14	1.12	1.07	1.07	1.23
1EAW	1.07	1.06	1.06	1.05	1.04	1.02	1	0.99	1.19	1.72	2.22	2.35	4.21
1EWY	1.09	1.08	1.08	1.07	1.06	1.04	1	0.99	0.99	1	1.07	1.22	1.56
1F34	1.03	1.03	1.03	1.03	1.03	1.01	1	1.01	1.18	1.45	1.92	2.05	2.47
1FQJ	0.98	0.98	0.98	0.98	0.98	0.99	1	0.97	0.99	1.02	1.03	1.09	1.08
1GCQ	1.01	1.01	1.01	1.01	1	1	1	1.01	1.07	1.06	1.06	1.05	0.95
1GHQ	1.02	1.02	1.02	1.01	1.01	1	1	1.01	1.09	1.19	1.23	1.24	0.84
1HE1	1	1	1	1	1	1	1	1	1	1	1	1	1
1KAC	1.07	1.07	1.07	1.07	1.06	1.03	1	0.97	0.93	0.93	1	1.09	1.42
1KTZ	1.03	1.03	1.03	1.03	1.03	1.02	1	0.87	0.52	0.49	0.49	0.5	0.56
1KXP	1.71	1.12	1.12	1.11	1.08	1.04	1	0.96	0.96	0.98	1.24	1.73	5.61
1MAH	1.06	1.05	1.05	1.05	1.04	1.02	1	0.94	1.33	2.19	2.96	2.89	3.64
1PPE	1	1	1	1	1	1	1	1	1.12	2.22	4.29	8.74	7.34
1QA9	1.04	1.05	1.06	1.07	1.1	1.15	1	0.84	0.71	0.65	0.69	0.7	0.85
1SBB	1.15	1.14	1.14	1.14	1.13	1.13	1	0.99	1.15	1.11	1.11	1.1	1.3
1TMQ	1.1	1.06	1.05	1.05	1.03	1	1	1.06	1.26	1.63	1.82	1.74	1.59
1UDI	0.97	0.97	0.97	0.97	0.96	0.94	1	0.96	0.98	1.32	1.46	1.64	2.06
2BTF	0.99	1	1	1	1	1	1	0.99	0.95	0.95	0.97	0.98	1
2PCC	1.11	1.11	1.04	1.03	1.03	1.01	1	0.91	0.84	0.87	0.9	0.94	1.35
2SIC	0.78	0.79	0.79	0.79	0.8	0.86	1	1.88	3.51	3.7	3.54	3.51	3.98
2SNI	1.05	1.05	1.05	1.05	1.04	1.02	1	1.03	1.11	1.04	1.05	0.88	0.83
7CEI	1.14	1.13	1.13	1.12	1.09	1.03	1	0.99	1	1.07	1.24	1.5	2.97
1ACB	0.97	0.97	0.97	0.97	0.97	0.99	1	0.98	0.95	1.04	1.09	1.11	1.43
1GRN	1	1	1	1	1	1	1	1.01	1.04	1.05	1.1	1.07	1.01
1I2M	1.51	1.49	1.41	1.4	1.34	1.23	1	0.96	0.97	0.98	1.04	1.03	1.35
1M10	1	1	1	1	1	1	1	1.04	1.03	0.64	0.64	0.68	0.82
1WQ1	1.18	1.18	1.18	1.18	1.17	1.1	1	0.64	0.59	0.6	0.61	0.62	1.31
1ATN	1	1	1	1	1	1	1	1	1	1	1	0.99	1
1FQ1	1	1	1	1	0.99	0.99	1	1.06	0.98	1.02	1.02	1.01	1
1H1V	0.99	0.99	0.99	0.99	0.99	0.99	1	0.92	0.9	0.88	0.88	0.82	0.88

：(a) に分類される例 ($\alpha > 1$ のとき結果が改善)

表 3 *bound* 構造と *unbound* 構造の差 (RMSD)
Table 3 Difference (RMSD) between *bound* structure and *unbound* structure.

	lig.	rec.	sum
1H1V	1.59	13.39	14.98
1FQ1	0.80	3.29	4.10
1KTZ	3.19	0.62	3.81
1I2M	2.65	1.14	3.79
2BTF	2.71	0.75	3.46
1ACB	1.76	1.49	3.25
1CGI	1.47	1.78	3.25
1ATN	2.71	0.41	3.12
1KXP	0.89	2.14	3.03
1M10	1.24	1.70	2.93
7CEI	1.18	1.72	2.90
1FQJ	0.54	1.88	2.42
1BUH	1.35	1.03	2.38
1GRN	1.74	0.63	2.37

	lig.	rec.	sum
1D6R	1.27	1.02	2.29
1B6C	0.35	1.92	2.28
1GCQ	1.11	1.10	2.21
1DFJ	0.66	1.51	2.17
1E6E	1.07	1.03	2.10
1WQ1	0.97	1.06	2.04
1AK4	0.52	1.46	1.98
1EWY	1.07	0.80	1.87
1GHQ	1.09	0.73	1.82
1F34	0.68	1.01	1.69
1HE1	0.92	0.74	1.67
1UDI	0.47	1.09	1.56
1QA9	0.84	0.67	1.52

	lig.	rec.	sum
1TMQ	0.39	1.10	1.49
1MAH	0.74	0.74	1.49
1KAC	0.50	0.96	1.46
1SBB	0.89	0.53	1.42
1E96	0.72	0.63	1.35
1AY7	0.51	0.60	1.11
1BVN	0.64	0.45	1.09
1EAW	0.58	0.51	1.09
1AVX	0.49	0.55	1.03
2SIC	0.25	0.64	0.88
1PPE	0.43	0.44	0.87
2PCC	0.34	0.49	0.84
2SNI	0.28	0.46	0.74

：(a) に分類される例 ($\alpha > 1$ のとき結果が改善)
：α の小さな増加 (5 程度まで) では結果が大幅に悪くならない例

3.4 *bound* 構造と *unbound* 構造の差異

ドッキング時のタンパク質の構造変化の度合いは，ドッキング予測計算における形状相補性の寄与の大きさと関係すると考えられる．そのため表 3 に，*bound* 構造と *unbound* 構造との間で計算した RMSD の値を示す．表中の sum の値をもとに，ドッキング時の構造変化が大きいタンパク質から順に並べている．

タンパク質の “柔らかさ” と形状相補性の寄与に関係があるとすれば，この値が大きいほど α を増大させたときの予測結果の改善が期待できる．表 3 について，リガンドとレセプターの RMSD の合計が大きいものだけに着目すると，(a) に分類されていないケースでも 1CGI を例外とすれば，5 程度までの α の増加が微小な結果改善につながる 1ACB や 1KXP などのケースや，1FQ1 のように α の変化が AUC にほとんど影響しないケースばかりである．従って，タンパク質のドッキング時の構造変化が大きいほど，形状相補性への寄与が小さい傾向があると言える．

4. 最適な静電重み α の推定

4.1 関連付けるタンパク質の性質

表 2 において結果が改善するような α の値を，ドッキング前のタンパク質の性質から推

測することで，α を決定する式を定義する．そこで次の数値との相関を調べる．

(1) 分子動力学 (Molecular Dynamics, MD) 計算による構造変化の度合い
(2) 残基数
(3) 溶媒露出表面積 (Solvent Accessible Surface Area, SASA)
(4) タンパク質の表面電荷

以上の情報を組み合わせて，予測精度を向上させる α を決定することを試みる．

4.1.1 MD 計算時の構造変化の度合い

“溶媒中での構造変化が大きいタンパク質は，ドッキング予測時の形状相補性の寄与が小さくなる”という仮説を立て，各タンパク質の “柔らかさ”との関係を調べることでその仮説を検証する．そこで MD 計算で得た構造変化の度合いからタンパク質の “柔らかさ”を推定する．MD 計算には AMBER 10[4] の sander を利用した．ただし，計算時間の問題からこの関係を調べるのはシミュレーションが完了している 8 例のみに止める．

4.1.2 タンパク質の大きさ

形状相補性に関する MEGADOCK の関数である rPSC は，タンパク質表面の凹凸が一致しているほど大きな値を示しやすいが，その数値はタンパク質の大きさに依存する傾向もある．そこで形状相補性と静電的相互作用との関係においても，少なからずタンパク質の大きさが影響すると考えられる．ここで用いる溶媒露出面積 (SASA) とは，溶媒分子を表す球を各原子の球面 (一般的にはファンデルワールス球面) に接して動かしたとき，球の中心の軌跡が描く曲面の面積である．

SASA の値は溶媒和自由エネルギーに大まかに比例していることが知られており[5]，かつ他のドッキングソフトウェアで脱溶媒和自由エネルギーのスコアを採用していることから，SASA の値を計算に用いる．SASA の計算には Surface Racer 5.0[6] を用いており，計算に際して Probe radius は 1.4 Å(ただし，エラーが発生する 1M10 ligand, 1ABB ligand, 1D6R receptor, 1F34 receptor の 4 例は 1.2 Å) としている．また残基数については，それ自体がおおよそタンパク質の大きさを示すが，SASA と組み合わせることでタンパク質を構成する各残基の大きさ等の情報としても活かすことができる．

4.1.3 タンパク質の表面電荷

正，負それぞれの電荷を帯びた部分の SASA と，極性を持った部分の SASA を，全体の SASA と同じく Surface Racer 5.0 より求め，α の決定に利用する．静電的相互作用の寄与の大きさは，ドッキングさせるタンパク質の表面電荷の影響を受ける可能性があるという推測からこの値を用いる．

表 4 MD 計算における 0ps 時の構造と 100ps 時の構造の差異 (RMSD)
Table 4 Difference (RMSD) between the structure at 0ps and 100ps in MD calculation by AMBER.

	ligand	receptor	sum	RMSD †
1AK4	1.45	0.91	2.36	3.28
1AY7	0.97	0.67	1.64	2.31
1B6C	1.30	0.66	1.96	2.61
1BUH	0.80	1.23	2.03	3.27
1CGI	0.94	1.00	1.94	2.94
1ACB	1.58	0.96	2.54	3.51
1M10	1.09	0.82	1.91	2.73
1FQ1	1.24	1.25	2.50	3.75

: (a) に分類される例 ($\alpha > 1$ のとき結果が改善)
† *bound* 構造と *unbound* 構造で計算した RMSD の合計

4.2 最適な α の推定

4.2.1 MD 計算時の構造変化の利用

熱平衡化後の初期構造と 100psec 後の構造との間での RMSD，その合計，及びそれらの *bound* 構造と *unbound* 構造との RMSD(リガンドとレセプターの合計) を表 4 に示す．なお，表中で色をつけた行は α の値を大きくすると精度が向上する分類 (a) のタンパク質である．100psec 後の構造としたのは，熱平衡化直後の構造からの大きな構造変化が十分に落ち着くためである．

この結果を見ると，表中の 3 列目 (sum) と 4 列目 (RMSD) に相関 (ピアソン相関係数 $C = 0.62$) が確認できる．このことから，ドッキング時に大きな構造変化を伴なう例を MD シミュレーションによって抽出できる可能性が示唆され，これは構造変化を考慮したドッキング予測では大きな意味を持つ．

一方，今回の小規模なターゲットでは，RMSD の差と α の値との間に明確な関係性は見出すことは難しい．(a) に分類されたタンパク質のうち，RMSD の差が明らかに大きいタンパク質は 1BUH だけにとどまり，分類されなかった他のタンパク質において比較的大きな値を取った例も散見される．これについて，もう少し大規模なデータセットで検証する必要がある．

4.2.2 タンパク質の大きさ・表面電荷の利用

続いて，タンパク質の表面積や，残基数，表面電荷等の情報を用いて，静電重み α を決定することを試みる．先に示した分類のうち (a) に属するタンパク質群，すなわち α を増大させたときに予測結果が改善する例について，この AUC の値を出来る限り小さくするよ

うな式を検討する．

ここでは 9 通りの定義を試みており，n 番目の静電重みの推定値 α_n^* を以下の式で定める．ここで，σ_n はタンパク質の特性から決定したパラメータ，z_n はデータセットの σ_n から算出した Z 値，Med は中央値，s はスケーリングパラメータである．

$$\alpha_n^* = \begin{cases} \beta^{(\sigma_n - \mathrm{Med}(\sigma_n))/s} & (n = 1, 2, 3) \\ \beta^{z_n} & (n = 4, \cdots, 9) \end{cases} \tag{8}$$

s, β について最適値を推定し，$s = 10$，$\beta = 3$ を得た．

次に (8) 式の $\sigma_n (n = 1, \cdots, 9)$ を定義する．S_{tot} は全体の SASA，S_{pol} は極性をもつ部分の SASA，S_{chg}^+, S_{chg}^- はそれぞれ＋と－の電荷を帯びた部分の SASA，N_{res} は残基数を表す．

$$\sigma_1 = S_{tot}/N_{res}(lig) + S_{tot}/N_{res}(rec) \tag{9}$$

上記の定義は，それぞれのタンパク質における残基の大きさを表す．

$$\sigma_2 = S_{tot}/N_{res}(lig) \tag{10}$$

$$\sigma_3 = S_{tot}/N_{res}(rec) \tag{11}$$

上記の 2 つは，σ_1 の定義と同様に，リガンドとレセプターそれぞれで分離して定義したものである．静電的相互作用のスコア計算において，レセプターに電界，リガンドに電荷を与えており，対称的な定義となっていない．従って，リガンドとレセプターそれぞれから算出される値に対して検証する．

$$\sigma_4 = \sigma_1 \tag{12}$$

σ_4 の定義は σ_1 と同様だが，(8) 式において指数を恣意的な値でなく Z 値にすることで，σ_5〜σ_9 と比較する．

$$\sigma_5 = S_{pol}/S_{tot}(lig) + S_{pol}/S_{tot}(rec) \tag{13}$$

極性を持った表面の占める割合を表しており，静電的相互作用のスコアに影響を及ぼす．

$$\sigma_6 = (|S_{chg}^+(lig) - S_{chg}^-(rec)| + |S_{chg}^+(rec) - S_{chg}^-(lig)|)/(S_{pol}(lig) + S_{pol}(rec)) \tag{14}$$

極性を持った表面のうち，正の電荷と負の電荷がどれほど対応しているかを表現する．

$$\sigma_7 = S_{chg}^+/S_{pol}(lig) + S_{chg}^+/S_{pol}(rec) \tag{15}$$

$$\sigma_8 = S_{chg}^-/S_{pol}(lig) + S_{chg}^-/S_{pol}(rec) \tag{16}$$

正，負それぞれについて，タンパク質表面での電荷の偏りを表しており，それらを加算することで 2 つのタンパク質での大まかな対応関係がわかる．すなわちそれぞれの値が 0 に近い，もしくは 1 に近い場合には，2 つのタンパク質が表面に同様の電荷を抱えることを表す．

$$\sigma_9 = S_{pol}/N_{res}(lig) + S_{pol}/N_{res}(rec) \tag{17}$$

最後に定義する σ_9 は，各残基の極性を持った部分がどれだけ表面に現れているかを表す．

5. 評 価 実 験

5.1 9 通りの推定値 α_n^* についての実験結果

4.2.2 で定義した $\alpha_n^* (n = 1, \cdots, 9)$ について，表 1 のデータセットに対して実験し，R_{AUC} を求める．全てのケースに対して計算した R_{AUC} の平均値，AUC < 0.9 を満たす例の個数，AUC > 1.1 を満たす例の個数を表 5 に示す．

この結果から，僅かながら平均が小さく，$R_{AUC} < 0.9$ の個数と $R_{AUC} > 1.1$ の個数の差が最も大きいため，先に挙げた推定値のうち α_8^* が最も良い結果を返していると言える．

次に各 α_n^* について，データセットの全てのタンパク質に対する実験結果をまとめたグラフを**図 1** に示す．このグラフは，$1 - R_{AUC}$ の値を積み重ねたグラフで，正になるケースと負になるケースをそれぞれ軸の上下に重ね，その大きさを比較する．すなわち，グラフが 0 よりも上に伸びるほど結果の改善が顕著となり，下に伸びるほど結果が悪くなるケースが増えることを意味する．従って，なるべくグラフの下への伸長を抑えながら，上への伸びが大きければ，予測を改善させたと言うことができる．

図 1 から α_n^* を評価すると，この結果においても α_8^* の結果が，他と比べて良好な結果を返している．他のケースと比べて負の方向への伸びが小さく，結果の大幅な悪化を招くこと無く，AUC の改善を果たしているためである．

5.2 α_1^* に対する閾値の導入と離散化

図 1 からは，$\alpha_1^*, \alpha_3^*, \alpha_4^*, \alpha_9^*$ ではどれも一つの例 (水色:1PPE) が大幅な悪化の要因となっていることも読み取れる．このタンパク質はリガンドの残基数が 29 と少なく，残基数を分母にすえる σ の定義において，σ の極度な増大の引き金となっている．このように α^* が極端に変化することは，予測結果の大幅な悪化につながる恐れがある．実用上このようなケー

Vol.2010-BIO-20 No.4
2010/3/4

表 5 9 通りの推定値 α_n^* に対する実験結果
Table 5 The result of experimentation to 9 kinds of estimation α_n^*.

	α_1^*	α_2^*	α_3^*	α_4^*	α_5^*	α_6^*	α_7^*	α_8^*	α_9^*
R_{AUC} の平均	1.21	0.98	1.23	1.06	0.99	1.01	0.99	0.97	1.04
$R_{AUC} < 0.9$ を満たす個数	6	3	2	5	4	4	4	6	6
$R_{AUC} > 1.1$ を満たす個数	4	0	7	4	1	4	2	1	5

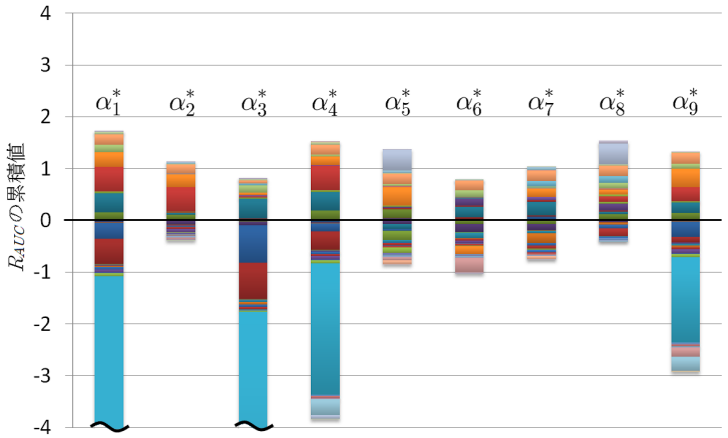

図 1 9 通りの推定値 α_n^* に対する R_{AUC} の累積値
Fig. 1 The cumulative values of R_{AUC} for 9 kinds of estimation α_n^*.

スが現れる可能性を残すことは好ましくない．従って，ある閾値で α^* の値に制限を設けること，及び数個の簡単な値に離散化して予測することを試みる．

検証の対象とするのは，先に挙げたうち α_1^* である．この値を次のように制限し，同様のデータセットに対して評価実験を行う．

(1) α_1^* に上限 M と下限 m を定め，推定値 α^* の範囲を $m \leq \alpha^* \leq M$ とする．$(m, M) = (0.1,\ 10), (0.2,\ 5), (0.5,\ 2)$ の 3 通り検証し，その値を各々$\alpha_{cut1}^*, \alpha_{cut2}^*, \alpha_{cut3}^*$ とする．

(2) α_1^* の値から離散値として α_{dis}^* を決定する．その定義は以下の通り．

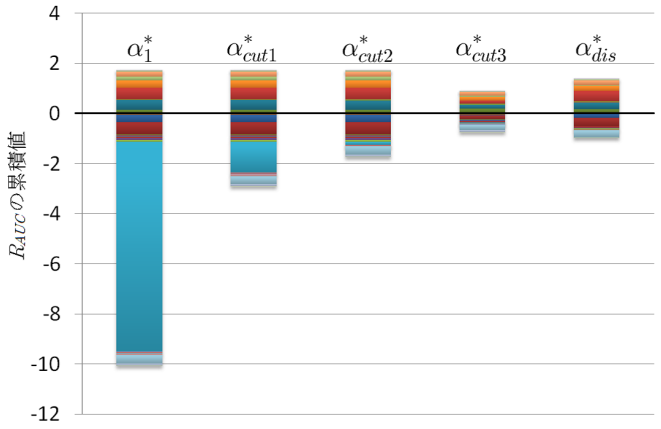

図 2 α_1^* に対する閾値の決定・離散化の実験結果
Fig. 2 The result of experiment of decision of threshold and discretization to α_1^*.

$$\alpha_{dis}^* = \begin{cases} 3 & \alpha_1^* \geq 3 \\ 2 & 2 \leq \alpha_1^* < 3 \\ 1 & 0.5 < \alpha_1^* < 2 \\ 0.5 & 0.3 < \alpha_1^* \leq 0.5 \\ 0.3 & \alpha_1^* \leq 0.3 \end{cases} \tag{18}$$

これらの定義式に基づいて，実験を行った結果を**図 2** に示す．この結果から，α_1^* に対する閾値の決定によって，特定の対象における結果の大幅な悪化が避けられたと言える．ただし 0.5〜2 という狭い範囲で定義される α_{cut3}^* では，良好な結果にまで影響を及ぼしており，離散値を取った α_{dis}^* でも良好な結果の後退を招いており，今回の実験においては α_{cut2}^* の定義が最も良い結果を返したと判断できる．

6. 考　　察

6.1 静電重み α の変化による予測精度改善の可能性

表 2 で α を増減させた際の AUC の変化を示しているが，結果が改善するような適切な α を持つケースは多く，予測精度の改善は十分に期待できる．特に 1KTZ や 1WQ1 のよう

に，$\alpha = 1$ のときの値と比べて AUC が半分近くまで改善するケースへの有効性は高いと判断できるため，この結果を反映するような α の定義式の発見は十分に価値がある．ただし，α の値を大きくしたときに結果が悪くなるケースも多く存在するため，適切に静電重み α を定義することが重要となる．

6.2 推定値 α^* の定義と予測の改善

6.2.1 α_n^* の比較について

(8)〜(17) 式で定義された 9 つの推定値 α^* を表 5，図 1 でそれぞれ比較した結果，最も良好な結果であると判断できるのは α_8^* であった．R_{AUC} の平均，R_{AUC} が小さいものの個数でも最も結果が良好である上に，結果が悪化するような α の選択を控えながら，複数のタンパク質において結果を改善させたことがグラフから読み取れる．

この α_8^* が良い結果を与えたことから，(16) 式で定義された σ_8 が表すタンパク質表面での電荷の偏りに関する情報が，*unbound* ドッキングにおいて重要な情報であることが推測される．また，極性を持った部分における正と負のバランスに偏りがあるときに，静電相互作用の影響を変化させる必要があると解釈できものの，α_7^* の結果が α_8^* と比べて劣っているため，更なる検証が必要であると考えられる．

6.2.2 閾値の設定と離散化について

また α_1^* の結果に対して閾値による上限と下限の設定や離散化を試したが，ここで α_1^* を対象として選択したのは，極端に悪化させる特別な例があったことだけが理由ではなく，システムへの静電重みの導入が比較的容易であることも理由として挙げられる．α_1^* の算出には SASA と残基数の情報が必要で，本稿では前者を Surface Racer 5.0 から獲得しているが，ドッキング計算の前に行うグリッド化において，rPSC の計算に用いるスコアをタンパク質の表面とコアの部分でそれぞれ割り当てる．従って，タンパク質表面のグリッドに割り当てられたスコアを足し合わせるだけで，SASA と相関の強い数値を求めることが可能である．実際にデータセットのレセプターについて，Surface Racer 5.0 で計算した SASA と，表面のグリッドに割り当てられたスコアの合計との間で相関係数を計算すると 0.98 という強い相関が確認できる．当然，残基数を得ることも容易であることから，α_1^* による結果を洗練することができれば，MEGADOCK への導入においても大きな前進となる．

その α_1^* への閾値の設定は，非常に有効であったと考えられる．上限 5，下限 0.2 を与えた α_{cut2}^* では，大幅な結果の悪化が起こる例 (1PPE) による影響を緩和するだけでなく，良好な結果を妨げることもなかったため閾値の設定が効果的であったと言える．

また，先の実験で最も良いと判断した α_8^* については，σ_8 の定義から極端な値を取ることは稀である，実験において極端に結果が悪化するケースが無かった等の理由により，閾値の設定が結果に及ぼす影響は小さいと考えられる．

7. 結　　論

本研究では，*unbound* ドッキングにおける予測精度向上のため，剛体予測手法における目的関数をタンパク質の特性の基づいて変化させることを提案した．形状相補性と静電的相互作用から計算される MEGADOCK の目的関数において，適切な静電重み α をタンパク質単体の性質から推測し，ターゲットごとで変化させることを試みた．その結果，ベンチマークテストにおけるドッキング結果の改善が確認されたため、適切な定義式の決定が予測精度の向上に有効であることが示された．最も予測精度の向上が顕著であった静電重みは，極性を持つ部分の SASA のうち，負に帯電している部分の割合を用いて決定した値であり，40 例中 6 例で大幅に予測結果が改善した．加えて，SASA と残基数から定めた推定値において，上限と下限を決める閾値を設定したことでも予測の改善が見られ，閾値の決定が効果的となるケースがあることも示された．これらの結果から，*unbound* ドッキング予測の精度向上に貢献できる新規アイディアを提案することができた．

謝辞　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, *Proceedings of the 2008 Annual Conference of the Japanese Society for Bioinformatics*, P032，2008.
2) M. Ohue, Y. Matsuzaki, Y. Matsuzaki, T. Sato, Y. Akiyama, *IPSJ-SIG Technical Report*, 2009-BIO-17(11), 2008.
3) Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, Janin J, Weng Z, Proteins, 60, 214-216, 2005.
4) Case, D A ; Darden, T A ; Cheatham, T E *et al.* San Francisco : University of California, 2008.
5) J Wang, W Wang, S Huo, M Lee, P A. Kollman, *J.Phys. Chem.* B, 105, 5055, 2001.
6) O V Tsodikov, M T Record, Y V Sergeev, *J Comput Chem*, 23(6), 600-609, 2002.