Vol.2009-BIO-18 No.1
2009/9/17

分子動力学法に基づくタンパク質構造サンプリングとドッキング予測の改良

松崎 裕介[†1] 松崎 由理[†1]
関嶋 政和[†1,†2] 秋山 泰[†1]

タンパク質を剛体とみなすタンパク質間ドッキング予測では，NMR や X 線結晶解析により決定された構造と溶媒中構造の違いが大きい場合やドッキングに構造変化を伴う場合，複合体の構造を正確に予測することが困難になる．そこで分子動力学計算をもとにタンパク質構造のアンサンブルをとり，クラスター分析を適用してドッキング用の候補構造群を抽出する．これによって得られた構造に対して従来のドッキング計算を実行することで，構造変化を考慮したドッキング予測を実現した．

Protein structure sampling based on molecular dynamics and improvement of docking prediction

Yusuke Matsuzaki ,[†1] Yuri Matsuzaki ,[†1]
Masakazu Sekijima [†1,†2] and Yutaka Akiyama[†1]

When a protein structure in the solvent changes greatly from its structre decided by NMR or X-ray crystallographic analysis, or a structural change is needed for docking, it is difficult to computationally predict the structure of the complex precisely with the rigid protein-protein docking method. In order to solve this problem, we calculate the ensemble of the protein conformations with molecular dynamics simulation, and then the cluster analysis is applied, and candidates structural group for rigid-docking is extracted. As a result, a protein-protein docking prediction that considers the structural change is achieved by executing the existing rigid-docking calculation with the obtained structures.

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が，従来は配列モチーフを利用する方法か，あるいは既知の構造情報を使うとしても剛体予測が主流となっており複合体の構造を基にした *bound* docking で評価することが多い．しかし NMR や X 線結晶解析により決定された構造と溶媒中構造との間に構造の差異があるケースや，ドッキングに構造変化を伴うケースもあるため，たとえ *bound* docking で高い精度が得られても各々の結晶構造等から予測を行う *unbound* docking においては，剛体予測だけで精度向上を実現することが困難な例もある．ゆえに各々のタンパク質構造からドッキング予測を行い，未知の複合体構造の特定に活用するためには，構造変化を考慮したドッキング予測システムの開発は不可欠である．

本研究では，タンパク質の構造変化を加味するために分子動力学法 (MD: Molecular Dynamics) を用いる．分子動力学法でドッキング自体のシミュレーションを実行すればより精度の高い予測が実現できるが，それには莫大な計算時間が必要となる．そのため予測する 2 つのタンパク質に対して，本稿では 1 つずつ比較的短時間のシミュレーションを行い，そのトラジェクトリからドッキング用の候補構造を抽出する．そして得られた複数の構造に対して網羅的に従来の剛体予測を適用することで，各々の結晶構造等に対する剛体予測だけでなく，構造変化を考慮したドッキング予測を行うことが可能となる．これは単体での構造と複合体構造との間の違いが大きいケースに対して有効であると考えられる．

本研究で使用したソフトウェアは，分子動力学法の計算には AMBER 10[1] の sander，従来の剛体ドッキング予測には ZDOCK 3.0.1[2)3)] である．また，その予測精度の評価に用いるタンパク質構造データは ZDOCK benchmark 2.0[4] から選出した．

†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

Vol.2009-BIO-18 No.1
2009/9/17

2. 構造サンプリング

2.1 分子動力学法の計算条件

構造サンプリングのために実行する分子動力学計算では，PDB[5] ファイルをもとにエネルギー最小化，熱平衡化を経て 1nsec のシミュレーションを行う．実行時の sander の計算条件を表 1 に示した．計算には東京工業大学の TSUBAME Grid Cluster ⋆1を使用した．

なお，本研究で対象とした ZDOCK benchmark 2.0 のタンパク質の PDB データは，http://zlab.bu.edu/julianm/benchmark/ から取得しているが，これらでは水素原子が省略されている．そのため AMBER 10 の protonate コマンドによって水素を付与した．また，PDB データに HETATM 行を含むものもあるが，これらの行はすべて削除して実行した．

2.2 候補構造の抽出

2.1 の計算で得られるトラジェクトリは，候補構造群として捉えるとその数が膨大である．そのためレセプタータンパク質とリガンドタンパク質の双方から取得する大量の候補構造に対して，網羅的に剛体予測を実行することは現実的ではない．そこでトラジェクトリの各構造に対してクラスター分析を適用[6] し，解析可能な数まで候補を絞り込む．これにより構造群を代表する少数の構造のみを対象とした網羅的な計算という形で，従来の剛体予測と組み合わせることが可能になる．

表 1 AMBER10 の計算条件
Table 1 Calculation condition of AMBER 10.

Prior processing		
Solvation box	TIP3PBOX	
	(buffer: 10Å	closeness 1.0Å)
Minimization		
minimization cycles	2000	
Equilibration		
simulation time	40psec	
temperature	0 → 300K(20psec),	300K(20psec)
ensemble	NVT	
nonbonded cutoff	8.0Å	
Molecular Dynamics		
simulation time	1nsec	
temperature	300K	
ensemble	NPT	
nonbonded cutoff	8.0Å	

⋆1 各ノードは CPU Opteron 880(2.4GH) × 8，メモリ 32GB

2.2.1 非類似度の定義

クラスター分析の適用のためには，構造間の差異を表す非類似度を定義する必要がある．ここでは構造変化の度合いを数値化する必要があるため，非類似度には Kabsch 法[7] に基づく RMSD(Root Mean Square Deviation) を用いる．ただし動作が大きくなりやすい水素原子や側鎖の原子によって非類似度が大きく変化することは好ましくないため，RMSD の値は Cα 原子のみを対象として計算する．なお，ここで RMSD の計算に使用したプログラムは http://www.personal.leeds.ac.uk/˜bgy1mm/Bioinformatics/rmsd.html から得た．

2.2.2 トラジェクトリのクラスタリング

クラスタリングのアルゴリズムは，K-means 法に代表される非階層的クラスタリングと，逐次的再計算の方法で複数に分類される階層的クラスタリングの 2 つに大きく分けられるが，本研究では階層的クラスタリングを用いて解析を行う．なぜならば非類似度の定義に RMSD を用いており，ユークリッド距離が前提となる K-means 法は適さないためである．

1nsec の分子動力学計算から得たトラジェクトリは 5psec ごとの構造からなっており，200 の構造に対してクラスタリングを実行することで 10 個の候補構造に絞り込むことにした．構造間で非類似度を計算し，それが最小となるクラスターを結合することを繰り返すが，この際の再計算の方法を変えることで (1) 群平均法, (2) 最短距離法, (3) 最長距離法, (4) ウォード法の 4 種類のアルゴリズムを実装した[8]．ただしウォード法もユークリッド距離を用いることを前提としているため，非類似度に RMSD を用いることはできない．ここでは初期クラスター に対して RMSD δ を用いて非類似度 d を $d(G_i, G_j) = \frac{1}{2}\delta^2$ とし，$G' := G_q \cup G_r$ のときクラスターの更新式を

$$d(G', G_i) = \frac{1}{|G_q| + |G_r| + |G_i|}\Big[(|G_q| + |G_i|)d(G_q, G_i\ ) + (|G_r| + |G_i|)d(G_r, G_i) - |G_i|d(G_q, G_r)\Big] \quad (1)$$

と定義しているため，ウォード法とは厳密には異なる．しかし便宜上，本稿では上記の手法をウォード法と記す．

Vol.2009-BIO-18 No.1
2009/9/17

2.2.3 ドッキング予測への応用

2.2.2 の手法を用いることでレセプターとリガンドそれぞれから小数個の候補構造を抽出することができる．本研究の目的は構造変化を考慮したドッキング予測の実現にあるため，サンプリングした構造群に対して剛体予測を実行しドッキング予測の結果とする．剛体予測には ZDOCK 3.0.1 を用いるが，これは 1 対 1 の予測に対してデフォルトで 2000 に及ぶ複数の予測結果を出力する．この 1 対 1 予測を候補構造群に対して網羅的に実行するため得られる結果の数は膨大となるが，それぞれの予測結果について計算されたスコアも出力されるため，このスコアをもとに得られた結果を再度並び替え，最終的な予測結果とする．

3. 手法の評価

3.1 評 価 対 象

ZDOCK benchmark 2.0 より表 2 に示した 5 例を選出し，レセプターとリガンド双方に対して構造サンプリングを実行した結果について，下記の 2 種類の方法で評価を行う．これら 5 例の選出においては，ドッキング予測が比較的簡単なものとそうではないものをどちらも選出することや，タンパク質の残基数が多すぎないことなどの点を考慮して決定した．ZDOCK benchmark 2.0 では予測の難易度に応じて Rigid-body，Medium Difficulty，Difficult の 3 種類の区分がなされているが，それぞれの区分に属する対象を 1 つ以上選んだ．

3.2 評 価 方 法

3.2.1 *bound* 構造との比較

unbound 構造（単体での結晶構造）から複合体の構造を予測したいため，構造サンプリングによって *bound* 構造（複合体での結晶構造）に近い構造が得られるかどうかが重要となる．そこでサンプリングで抽出した構造と *bound* 構造との間で RMSD を計算して評価を行う．ただし ZDOCK 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，BLAST2 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．

表 2 対象タンパク質の詳細
Table 2 Details of target proteins.

complex	categories	receptor ID	R_RMSD ⋆1	ligand ID	L_RMSD ⋆1
1AY7	Rigid-body	1RGH_B	0.506	1A19_B	0.602
1CGI	Rigid-body	2CGA_B	1.470	1HPT_	1.783
1ACB	Medium Difficulty	2CGA_B	1.760	1EGL_	1.493
1M10	Medium Difficulty	1AUQ_	1.235	1MOZ_B	1.697
1FQ1	Difficult	1FPZ_F	0.803	1B39_A	3.292

⋆1 レセプター，リガンドそれぞれに対して *bound* 構造と *unbound* 構造との間で Kabsch 法で計算した RMSD.

表 3 クラスタリングで得た構造と *bound* 構造との間の RMSD
Table 3 Minimum of RMSD between structures by clustering and *bound* structure.

	最小値	最大値	群平均法	最短距離法	最長距離法	ウォード法
1AY7_R	0.81	1.27	**0.82**	**0.82**	**0.82**	**0.82**
1AY7_L	0.67	1.27	0.79	0.87	0.78	**0.67**
1CGI_R	1.64	2.15	**1.67**	**1.67**	**1.67**	1.71
1CGI_L	1.27	2.08	1.34	1.34	**1.34**	1.34
1ACB_R	1.98	2.29	**1.98**	**1.98**	2.028	2.03
1ACB_L	1.37	2.32	1.71	1.71	**1.68**	1.71
1M10_R	1.31	1.89	1.38	**1.34**	1.38	1.38
1M10_L	1.66	2.51	1.81	1.81	**1.73**	**1.73**
1FQ1_R	1.07	1.65	1.26	1.32	1.22	**1.20**
1FQ1_L	2.94	3.44	**2.94**	2.98	**2.97**	2.94

3.2.2 ドッキング予測結果の評価

ドッキング予測精度は，正しい複合体結晶構造と計算機予測した複合体構造について Kabsch 法による RMSD 計算を実行し評価する．*unbound* のデータをもとにした構造と *bound* データとの比較であるため，3.2.1 と同じく若干の配列の違いが生じるが，計算対象とする Cα 原子を 3.2.1 と同じ基準で選ぶことで対処する．

3.3 結　　果

まず 3.2.1 の方法に基づいて評価を行う．**図 1**，**図 2** に示したのは，分子動力学計算から得たトラジェクトリの 5psec ごとの構造と *bound* 構造との間で RMSD を計算した結果と，ウォード法によるクラスタリングの実行によって抽出したタイムステップをプロットしたものである．これをレセプターとリガンドそれぞれで分割して示した．また表 3 はトラジェクトリの全ての構造について計算した RMSD の最小値と最大値，及びクラスタリングのアルゴリズムごとに抽出した 10 個の構造群と *bound* 構造との RMSD をそれぞれ計算しその最小値を示したものであり，**図 3** は 1ACB のリガンドを対象として 4 種類のクラスタリングアルゴリズムで得たものをプロットしたものである．

表 3 の結果から 4 つのアルゴリズムを比較すると，トラジェクトリ中から RMSD が小さ

Vol.2009-BIO-18 No.1
2009/9/17

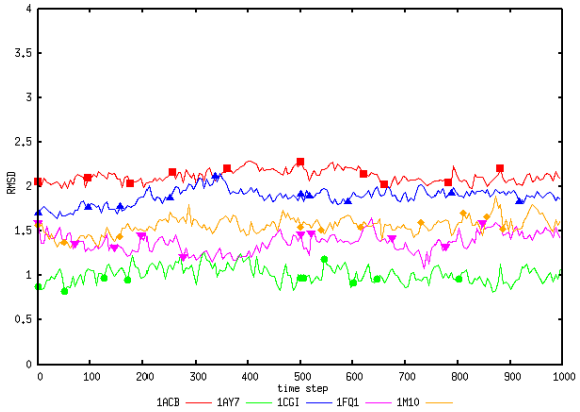

図 1 トラジェクトリと *bound* 構造との間の RMSD（レセプター）
Fig. 1 RMSD between torajectory data and bound structure of receptor.

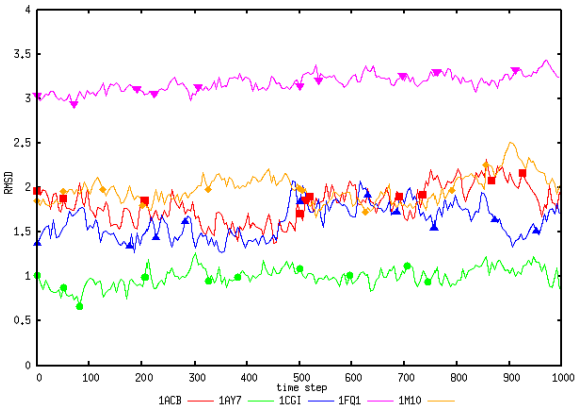

図 2 トラジェクトリと *bound* 構造との間の RMSD（リガンド）
Fig. 2 RMSD between torajectory data and bound structure of ligand.

い構造を抽出する性能で考えれば大きな差はないが，最も良い結果であると判断できるのはウォード法である．4 つのアルゴリズムの中で RMSD が最小となる例の数で見れば，10 例中 5 例で最小値を得ている最長距離法もウォード法と同じだが，ウォード法が優れていると判断する理由は 1AY7_L と 1FQ1_L にある．これらの最小値は 4 つのアルゴリズム内で最小というだけではなくトラジェクトリ全体で見ても最小となる構造を抽出しており，特に

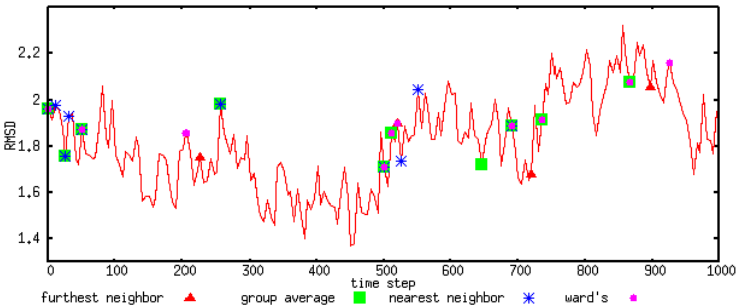

図 3 1ACB_L に対するクラスタリングアルゴリズムの比較
Fig. 3 Comparison of clustering algorithm (1ACB_L).

1AY7_L では他の 3 つのアルゴリズムより 2 割近く小さい値を得ている．また最小値を取らなかった他の 5 例の数値も，他のアルゴリズムと比べて著しく劣るものではないことから，少なくとも今回このターゲットの範囲ではウォード法が優れていると結論付けられる．

続いて図 1，図 2 の結果について見てみると，分子動力学計算前の構造よりも RMSD が小さくならない 1FQ1 のリガンドのような例もあるが，これは計算対象とした *bound* 構造も同様に X 線結晶解析などで特定された構造であるため，シミュレーション前の方が数値が小さくなりやすい可能性もある．このような *bound* 構造に近づかないケースがある一方で，表 3 の最小値と最大値の値も示す通り，1ACB, 1CGI, 1M10 らのリガンドは RMSD の値に大きな変動が見られる．その結果得られる多様性に富んだ構造の中から，クラスタリングの適用によって *bound* 構造に出来る限り近い構造群を獲得できることが理想となるが，図 3 で示した 1ACB のリガンドでは，どのアルゴリズムも RMSD が小さくなる 300〜500psec あたりの構造を抽出していない．この点でアルゴリズムに改善の余地があるものの，ウォード法や最長距離法が他の 2 つのアルゴリズムに比べて比較的離れた構造を抽出していることは，表 3 で優れた結果が出た 1 つの要因となったと考えられる．

次に 3.2.2 で示した方法により，提案する手法の評価を行う．表 4 には *unbound* 構造同士に対して従来の剛体予測でドッキング計算した場合と，分子動力学法を活用し 2.2.3 の方法によって順位付けした場合で，正しい複合体構造との RMSD を Kabsch 法で計算した結果を示す．ここでは 1 位の予測に対する値を上段に，10 位以内で最小となる値とその順位（かっこ内）を下段に記した．下段の “-” は 1 位が RMSD 最小の構造となる例である．

表 4 の結果を見ると，1AY7 の予測は剛体予測よりも比較的良好な結果が得られている．

Vol.2009-BIO-18 No.1
2009/9/17

表 4 *unbound* 構造同士の剛体予測と分子動力学法を活用した手法との性能比較
Table 4 Performance comparison of the method using molecular dynamics with rigid docking for using *unbound* structures.

	rigid docking	群平均法	最短距離法	最長距離法	ウォード法
1AY7	12.2	8.6	12.9	7.9	12.4
	8.2(2)	-	8.7(4)	-	6.5(9)
1CGI	7.2	15.7	16.2	16.1	16.0
	2.2(5)	15.4(10)	15.7(5)	15.5(3)	6.4(6)
1ACB	7.9	17.2	16.4	17.2	8.4
	4.8(4)	11.6(9)	11.5(3)	11.0(3)	8.1(9)
1M10	13.4	15.6	21.7	20.0	21.7
	-	13.9(9)	15.7(3)	13.9(6)	13.1(9)
1FQ1	18.7	17.6	15.6	17.2	21.0
	12.8(9)	17.3(9)	15.5(5)	16.9(9)	15.7(9)

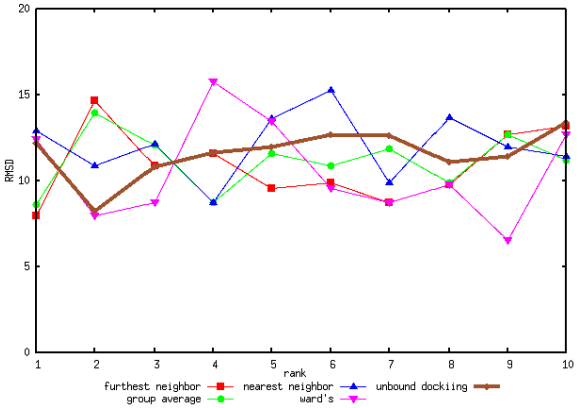

図 4 1AY7 のドッキング結果の比較
Fig. 4 Comparison of docking results (1AY7).

この 1AY7 について，**図 4** に最上位だけでなく各アルゴリズムで得た全ての予測結果の RMSD を示した． その結果から，他のクラスタリングアルゴリズムでも同様に剛体予測より低い RMSD を得ていることが分かる．従って 1AY7 はこの手法に適していると考えられ，より精度の高い構造抽出が実現できればドッキング予測の精度向上に結びつく可能性がある．また相互作用面の位置を正しく予測できるかどうかも重要であるため，**図 5**，**図 6** に剛

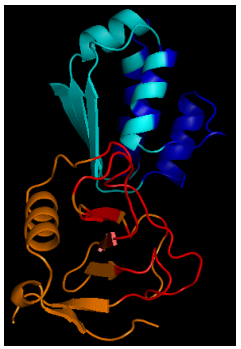

図 5 ウォード法で予測された RMSD 最小の複合体構造
Fig. 5 Complex structure with minimum RMSD predicted by Ward's method.

図 6 剛体予測で予測された RMSD 最小の複合体構造
Fig. 6 Complex structure with minimum RMSD predicted by rigid docking.

体予測とウォード法それぞれで *bound* 構造との RMSD が最小となる構造を示した．オレンジ色がレセプター，水色がリガンドのタンパク質で，それぞれの色が濃い部分が実際に複合体を形成する際の相互作用面である．これらを見比べてみると，ウォード法の予測は剛体予測よりも本来の相互作用面が 2 つのタンパク質の接触面から遠く，相互作用面の位置の予測精度が高いとは言えない．従って，剛体予測より小さな RMSD を得ることが出来たものの，正確な相互作用面の予測には課題も残る．

1AY7 の結果に良好な点がある一方で，他の 4 つの予測には問題が残る．特に 1CGI についてはウォード法以外の 3 つのアルゴリズムで大幅な RMSD の増大が見られ，この手法に適していないと言える．更に 1M10 や 1FQ1 のような剛体予測で予測が困難な例につい

ても RMSD の著しい減少はないことから，システムの改良が必要であると言わざるを得ない．この結果をもとにクラスタリングのアルゴリズムを比較すると，1CGI や 1ACB の結果からウォード法が最も良い結果を返していると言える．なおこの結果は 2.2.3 にある通り ZDOCK のスコアで順位付けをしており，特定の構造ペアの結果が多く含まれるケースもあるが，各構造ペアの 1 位の予測だけで順位付けをしても結果に改善は見られず，中には結果が悪化するケースもあった．

4. 結　　論

本研究では分子動力学法を用いてタンパク質の構造サンプリングを行い，取得した構造群に対して網羅的に剛体ドッキング予測を適用することで，タンパク質の構造変化を考慮したドッキングシステムを実現した．AMBER 10 の sander による分子動力学計算で得られたトラジェクトリからドッキングに用いる候補構造を抽出するために，RMSD を用いてクラスター分析を適用することで *bound* 構造に近い構造を抽出した．クラスター分析のアルゴリズムとして 4 種類を実装した結果，それらの性能の差は大きくないものの，本稿で用いたデータセットで最も *bound* 構造に近い構造を抽出したのはウォード法であり，トラジェクトリの中で比較的 *bound* 構造に近い構造を抽出することに成功した．サンプリングした構造群に対して ZDOCK 3.0.1 で網羅的にドッキング予測を行うと，それぞれの結晶構造からの剛体予測に比べて改善が見られたのは 1 例で，他の 4 例では予測精度を向上させることは出来なかった．その理由はシミュレーションの長さや計算条件にあると考えられる．図 1，図 2 からも読み取れるように劇的な構造の変化を捉えるには 1nsec という時間は短く，更に多様な構造を得るためには温度条件を変えるなどの対処も必要となる．

本稿で対象としたタンパク質が 5 例のみであることから今後検証するターゲットを増やすべきだが，この 5 例の結果では *unbound* docking の精度向上が果たされたとは言い難く，精度をさらに高める必要がある．そのために今回のシミュレーションで大きい構造変化が見られなかったターゲットについても，*bound* 構造が抽出できるように時間や温度を設定することが重要となる．もしトラジェクトリに *bound* 構造に近い構造が含まれるならば，クラスター分析による候補構造の抽出によってその構造を取得できる可能性が高いことが本研究で示されたため，この手法の流用による候補構造の特定が期待できる．一方ドッキング予測への応用には候補構造の抽出に比べて課題が多いが，トラジェクトリからの候補構造抽出の際に特定のタイムステップ周辺の構造が多く選出されることがある問題の解決や，候補構造群同士の網羅的なドッキングによって得られた膨大な数の結果から，より精度の高い予測結果を選ぶための改良や検証を経ることで，*unbound* dockign の精度を高められる可能性がある．また分子動力学計算だけでなく網羅的な剛体ドッキング予測にも多くの計算時間を要するが，当研究室で開発が進む網羅的ドッキングシステム MEGADOCK[9] を用いることで計算時間の短縮が期待できる．

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) Case, D A ; Darden, T A ; Cheatham, T E ; Simmerling, C L ; Wang, J ; Duke, R E ; Luo, R ; Crowley, M ; Walker, R C ; Zhang, W ; Merz, K M ; Wang, B ; Hayik, S ; Roitberg, A ; Seabra, G ; Kolossvary, I ; Wong, K F ; Paesani, F ; Vanicek, J ; Wu, X et al. San Francisco : University of California, 2008.
2) Chen R, Li L, Weng Z: “ZDOCK: an initial-stage protein-docking algorithm”, Proteins, 52:80-87, 2003.
3) Mintseris J, Pierce B, Wiehe K, Anderson R, Chen R, Weng Z. “Integrating statistical pair potentials into protein complex prediction.” Proteins, 69, 511-520, 2007.
4) Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, Janin J, Weng Z: “Protein-Protein Docking Benchmark 2.0: an update”, Proteins, 60, 214-216, 2005.
5) H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. “The Protein Data Bank” Nucleic Acids Research, 28, 235-242, 2000.
6) 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.
7) Kabsch, Wolfgang. “A solution of the best rotation to relate two sets of vectors”, Acta Crystallographica, 32:922, 1976.
8) 宮本定明: “クラスター分析入門 ファジィクラスタリングの理論と応用”, 第 2 章, 7 章, 森北出版 (株), 1999.
9) 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.