2008－BIO－13 (6)
2008／6／26

# タンパク質ドッキング予測プログラムによるシグナル伝達系のタンパク質間相互作用解析

松崎由理[†1] 松崎裕介[†1]
佐藤智之[†2] 秋山泰[†1]

我々は，立体構造データをもとにタンパク質のドッキング予測を行い，候補となるドッキング構造をクラスタリングにより非冗長化してタンパク質間相互作用を予測する，ドッキング後処理システムを開発している．ソフトウェアの利用目的として，複合体のドッキング構造そのものの予測と，多数のタンパク質の構造データに網羅的に予測プログラムを適用して相互作用する可能性のある組み合わせを発見する相互作用対検出問題の，二つの問題が考えられる．最適な後処理の方法はこの両者で異なる可能性があるが，本稿では主に後者の問題を対象とする．

本発表では，ドッキング予測ソフトウェア ZDOCK3.0.1 による結合構造の候補データをクラスター解析により評価する手法を，実際のシグナル伝達系の例として大腸菌走化性系の 9 つのタンパク質に適用した結果について述べる．実験的には未確認な相互作用も含めて，生物学的な見地からも予測結果の妥当性を検討した．また，構造データに基づく予測を実際のパスウェイ解析に用いる際に困難だった点についてまとめる．

# Analysis of protein-protein interactions of signal transduction pathways using a docking prediction program

Yuri Matsuzaki,[†1] Yusuke Matsuzaki,[†1] Toshiyuki Sato[†2] and Yutaka Akiyama [†1]

We have developed software for analyzing protein-protein interaction using tertiary structure data of target proteins by postprocessing outputs of a docking prediction software. Possible applications of our software include predicting accurate docking form of two proteins and discovering new combinations of proteins that can form complexes using amounts of protein structure data. Here we focused on the latter type of application and evaluated the proposed software by applying it to a real signal transduction pathway of bacterial chemotaxis, which has nine kinds of proteins in the pathway.

We first got docking prediction data by applying tertiary structure data of chemotactic proteins to ZDOCK 3.0.1. Then the outputs were clustered based on the structural differences of each docking prediction. We then evaluated the affinity of two proteins using the clustered data. As well as the results of evaluation of our software in protein-protein interaction prediction of bacterial chemotaxis, difficulties in applying our software to the real biological pathway are discussed.

## 1. はじめに

我々は，クラスター解析によるタンパク質ドッキング予測データの後処理システム（本稿では「ドッキング後処理システム」と表記）を提案した[1]．このシステムは，ドッキング予測ソフトウェア ZDOCK が出力する膨大な構造予測データをクラスタリングして非冗長化し，各タンパク質の組み合わせについて相互作用可能性の有無を判定する．

本発表ではドッキング後処理システムを細菌の走化性系に適用し，性能を評価した結果について述べる．細菌の走化性は古くから機構や動態の研究が進んでおり[2)3)]，タンパク質間の結合関係の多くが明らかになっているため，今回の検証材料とした．実際の生物系のパスウェイを対象にドッキングと後処理システムを利用する際，困難だった点についても述べる．

†1 東京工業大学大学院情報理工学研究科
Department of Computer Science, 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 and Research Institute

## 2. 対象と方法

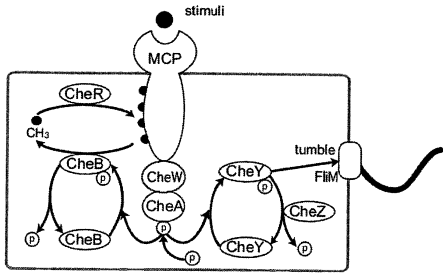
### 2.1 細菌の走化性

*E. coli* などの細菌は環境の変化に応じて遊泳パターンを変化させ，より好ましい環境に移動する．環境情報を検知して運動器官である鞭毛モーターに伝えるシグナル伝達系は，*E. coli* では主に図 1 の要素で構成されている．

細菌の走化性におけるシグナル伝達経路の詳細は種ごとに少しずつ異なるが，ヒスチジンキナーゼ活性をもつ受容体が刺激となる化学物質を検知して自己リン酸化能力を調節し，他のタンパク質にリン酸基を転移するという枠組みは共通している．今回は *E. coli* に存在する 9 種の走化性タンパク質とそのホモログを対象とした．

### 2.2 走化性に関連するタンパク質の構造データ

まず，細菌の走化性に関連するタンパク質の複合体データを PDB データベースより検索して取得した．これらの




| 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 Tar（Asp 受容体）および 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 脱メチル化酵素．リン酸化されると脱メチル化の活性が上昇する． |
| CheR | MCP メチル化酵素． |
| CheW | 足場タンパク質． |
| CheY | リン酸化されると鞭毛モーターと相互作用し，時計回りの回転を促す． |
| CheZ | CheY の脱リン酸化酵素． |

**図 1** *E. coli* の化学走性におけるシグナル伝達経路

**表 1** 評価実験に用いた複合体の PDB データ

| PDB ID | 複合体構成要素 | PDB ID | 複合体構成要素 |
|---|---|---|---|
| 1A0O | CheA, CheY | 1U0S | CheA, CheY |
| 1B3Q | CheA, CheA | 1U8T | CheY, FliM |
| 1BC5 | CheR, Tar | 1VLT | Tar, Tar |
| 1EAY | CheA, CheY | 2B1J | CheY, FliM |
| 1F4V | CheY, FliM | 2CH4 | CheA, CheW |
| 1FFG | CheA, CheY | 2D4U | Tsr, Tsr |
| 1FFS | CheA, CheY | 2PL9 | CheY, CheZ |
| 1FFW | CheA, CheY | 2PMC | CheY, CheZ |
| 1KMI | CheZ, CheY | | |

データについて，複合体の要素ごとに分離した構造ファイルを作成し，再ドッキングを行った．このようなドッキングは “Bound docking” と呼ばれる．

今回の目的は後処理システムの評価であるため，ドッキング予測自体が困難とされる，非複合体の結晶データからの新規のドッキング（“Unbound docking”）は評価における正解の基準から除外した．

具体的には，タンパク質名などのキーワード検索により取得した 24 のデータのうち．類似するものや PDB データに含まれるアミノ酸数が 10 個未満の小さすぎるものを除いた 17 個のデータ（表 1）を利用した．*E. coli* のデータ 10 件，*T. maritima* のデータ 3 件，*S. typhimurium* のデータ 4 件が含まれる．

比較のため，タンパク質ドッキング予測の精度を評価する用途で公開されている ZDOCK ベンチマークデータ[4] のうち，43 のデータを用いた相互作用対検出も行った．利用したデータの詳細については別文献[1] に記載している．以降このデータセットについては単に「ベンチマーク」と表記する．

### 2.3 ZDOCK と後処理システムの適用

ドッキング予測ソフトウェア ZDOCK（バージョン 3.0.1）は入力としてタンパク質の三次元構造データを二つ受け付け，二者のうちサイズが大きいタンパク質を “Receptor”，小さい方を “Ligand” と呼ぶ．今回は，取得した 17 のタンパク質複合体データから要素タンパク質の構造データを分離し，サイズの大きいものを “Receptor” とした．次に全 “Receptor” について，全 “Ligand” タンパク質との組み合わせ（$17 \times 17$）で，ZDOCK によるドッキング予測結果を得た．ドッキング構造候補の出力数は ZDOCK のデフォルトである 2000 とした．

次に，得られた結合状態候補の構造類似性に基づいてクラスター解析を行った．ドッキング後処理システムで，メンバー数が多く，かつスコアの高いデータを含むクラスターは相互作用に関わる構造予測データを含む可能性が高いと仮定した．このようなクラスターが存在する場合にはそのタンパク質対の親和性が高いと判定することとした．

様々なクラスタリング手法を比較した結果から，多数のタンパク質間での相互作用対検出問題においては群平均法が適していることが示唆された[1]．そこで今回は群平均法を用いたクラスタリングを行った．

### 2.4 相互作用対の検出

前項で得た $17 \times 17$ のタンパク質対についてのドッキング構造予測データそれぞれを対象に，以下の手順で相互作用可能性の評価を行った．

なお本稿では，ドッキング後処理システムの応用が可能なドッキング予測問題と相互作用検出問題について，後者に焦点を置いている．そのため，ドッキング予測を主目的にした後処理について論じた関連発表[1] とは，クラスタリング結果の利用方法が異なる点がある．

(1) 2000 の予測データを構造類似性に基づき群平均法でクラスタリングする．

(2) クラスタリング結果をもとに以下の方法でタンパク質対に関する評価値を決定する．

- 各クラスター $C_i$ 中のデータにおいて，ZDOCK 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 ZDOCK スコアに対する Z 値を $s_i$ とする．クラスタ数を $N$ とすると，$C_i (1 \leq i \leq N)$，$s_i (1 \leq i \leq N)$ である．
- 各クラスター $C_i$ のメンバ数 $|C_i|$ について調べ，$N$ 個のクラスター中での各メンバー数を母集団として計算した Z 値を $m_i$ とする[5]．
- $m_i$ が閾値 $m^*$ より大きいクラスターの集合を $C\prime$ とする．$C\prime$ に含まれるクラスターの代表値 $s_i$ のうち，最大値を評価値 $E$ とする．

$$C\prime = \{C_i | m_i \geq m^*\} \tag{1}$$

$$E = \begin{cases} max\ s_i,\ i \in C\prime & \text{if } C\prime \neq \phi \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

(3) 以上のようにして $17 \times 17$ のタンパク質対全てについて親和性の評価値 $E$ を決定した後，$E$ が閾値 $E^*$ 以上となるタンパク質対を，相互作用可能と判定する．

評価実験では，閾値 $m^*$ を 0.0〜3.0 まで 0.5 刻みで，閾値 $E^*$ を 3.0〜10.0 まで 0.1 刻みで変更して，各パラメータの組み合わせについて精度を比較した．

## 3. 結　　果

### 3.1 検出された相互作用

図 2 に，前項における閾値 $m^*$ を 0.0，閾値 $E^*$ を 5.7 とした際に検出された相互作用を示す．これらのパラメータ

(a)

| | Y | A | Tar | Y | FliM | Y | Y | Y | Y | Y | FliM | Tar | FliM | W | Tsr | Z | Z |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (1a0o)CheA | | * | | | | | | | | | | * | | * | | | |
| (1b3q)CheA | * | * | | | | * | | | * | | * | | | | * | | |
| (1bc5)CheR | | | * | | * | * | * | | * | | * | * | * | | * | | |
| (1eay)CheA | | * | | * | | * | * | * | | | | * | | | | | |
| (1f4v)CheY | | * | | * | | | | * | * | | | | * | * | * | | |
| (1ffg)CheA | | | | | | * | | | | | | | * | | | | |
| (1ffs)CheA | | | | | | | | * | | | | | | | | * | |
| (1ffw)CheA | | | | | | | | | * | | | * | | | | | |
| (1kmi)CheZ | | * | * | | * | * | * | * | * | | | * | * | * | * | * | |
| (1u0s)CheA | * | | | * | | | | | | | | | * | | * | | |
| (1u8t)CheY | | * | | | * | * | * | * | * | | * | | | | | | |
| (1vlt)Tar | | | | | | | | | | | | * | | | * | | |
| (2b1j)CheY | | | * | | | | | | * | * | | | | | | | |
| (2ch4)CheA | | | | * | | | * | | * | * | * | | | * | | * | |
| (2d4u)Tsr | | | * | * | | * | | | | | | * | | * | * | * | |
| (2pl9)CheY | * | * | * | * | | * | | | | | | | * | | | | |
| (2pmc)CheY | | | * | | * | | | | | | | | | * | | | * |

(b)

| | A | R | W | Y | Z | Tar | Tsr | FliM |
|---|---|---|---|---|---|---|---|---|
| CheA | * | - | - | - | - | - | - | - |
| CheR | | | - | - | - | - | - | - |
| CheW | * | | | - | - | - | - | - |
| CheY | * | * | * | * | - | - | - | - |
| CheZ | ■ | | * | * | ■ | - | - | - |
| Tar | * | * | | * | * | * | - | - |
| Tsr | * | * | * | * | * | ■ | * | - |
| FliM | * | * | | * | * | | | |

(c)

**図 2** 予測されたタンパク質間相互作用．(a)17 × 17 の全組み合わせにおける相互作用予測結果．予測された相互作用を*印で示す．(b) タンパク質の種類ごとに予測結果をまとめたもの．(c) 予測された未知相互作用の一つ（CheA–CheZ）に関する予測されたドッキング構造．左側が CheA，右側が CheZ である．

値は，検出結果の真陽性率と偽陽性率の差が最大になった値である．

図 2(a) は，17 × 17 の全ての組み合わせについての検出結果である．複合体構造の再ドッキングを検出すべき対角要素を網掛けで示している．

(b) は，(a) をタンパク質の種類ごとに整理したものである．(a) において，一つでも相互作用が検出された組み合わせを陽性としている．複合体の立体構造が得られていない組み合わせも含めて，実際に相互作用が確認されている組み合わせを (a) と同様の薄色の網掛けで示した．濃色の網掛け部分は，生物学的に相互作用の可能性が示唆されている組み合わせである．このような例の一つとして得た CheA と CheZ のドッキング予測構造は (c) のようになっていた．現時点で CheA と CheZ の相互作用は確認されていないが，CheA の short form である $\mathrm{CheA}_s$ は CheZ と相互作用することが知られている．$\mathrm{CheA}_s$ と CheZ の結合に必要とされる Cys 残基[6] を黒色で示した．

### 3.2 相互作用対検出の精度

図 3 に，走化性系とベンチマークデータの両者にドッキング後処理システムを適用した際の検出性能評価結果を示す．対照実験として，後処理のクラスタリングを行わず，ZDOCK の 2000 個の出力のうち最大のスコアについて，2000 候補の中での Z 値を評価値にしたデータも示している（図 3, “maxscore” データ系列）．図 2(a) の対角要素を検出した場合を真陽性，それ以外の要素を陽性と予測した場合を偽陽性として計算した F 値で精度を調べた．

走化性系の結果では，閾値 $m^* = 0.0$ とした時に，それぞれのタンパク質の組み合わせについて $E^* = 7.2$ で相互作用可能と判定した場合に F 値が最大（0.32）となった．評価結果は与えたパラメータの値によって異なるが，多くの場合，閾値 $E^*$ =6.5〜7.5 の場合に高い精度を得ていた．

ベンチマークの結果では，F 値のピークがより緩やかで，$m^* = 2.5$, $E^* = 7.8$ とした際に F 値が最大（0.44）になった．また，走化性系に比べて，対照実験よりも良い精度を得られる傾向があった．

二つのデータセットで，最良の精度を得られるパラメータ値は異なっていた．両者に共通してある程度の精度を得られるパラメータとしては，例えば $m^* = 1.0$, $E^* = 7.5$ の条件で判定した場合，走化性系で F 値 0.24，ベンチマークで F 値 0.36 となる．

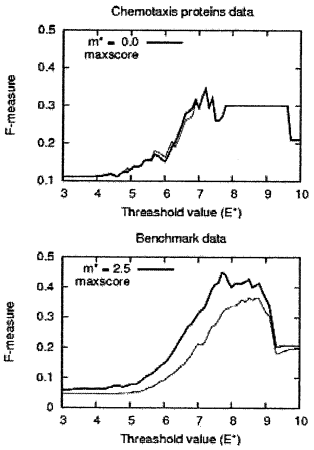



**図 3** ドッキング後処理システムの性能評価．上に走化性系 (17 × 17)，下にベンチマーク (43 × 43) の結果を示す．縦軸が相互作用対検出の F 値，横軸は相互作用判定を行った際の閾値 $E^*$ である．“maxscore” としたデータはクラスタリングを行わない対照実験．

## 4. 考　　察

### 4.1 検出された相互作用

予測された相互作用の中には，CheA–CheZ，CheZ–CheZ の結合のように，生物学的な役割が完全に明らかではないものの，結合する可能性が示唆されているものが含まれていた（図 2(b) の濃い網掛け部分）．例えば CheA の short form($\mathrm{CheA}_s$) は，CheZ と結合し，受容体付近において何らかのシグナル伝達制御を行っている可能性が報告されている[6]．また，CheZ は自身のオリゴマーを形成し，CheY のリン酸化状態のより複雑な制御を行うのではないかともいわれている[7]．実際には図 2(c) に示した CheA–CheZ のドッキング構造箇所は，$\mathrm{CheA}_s$ が CheZ と結合するために必要とされる Cys 残基の位置とは少し異

なるが，この構造を出発点にして実際に機能する結合構造を予測するという応用も考えられる．

今後はこのような，ドッキングと後処理によって示唆された未知相互作用の検証についても視野に入れて解析を行いたい．

### 4.2 実際のシグナル伝達系への適用における問題点

実際の生物系にドッキング後処理システムを適用するにあたり，以下のような問題点があった．

**データの属性**

ホモロガスなタンパク質であっても，変異株や，生理条件，存在場所，種などの違いがあるため，解析の目的によっては，これらの条件を慎重に吟味して解析対象のデータを決定する必要がある．そのため，機械的な検索で相互作用検出の候補となるタンパク質の PDB 構造データを集めることは簡単ではない．

**相互作用検出精度の評価基準**

相互作用対検出の正誤について，図 3 では複合体構造の再ドッキングを検出した際に正，それ以外の組み合わせを予測した際には誤とし，予測精度（F 値）を示している．走化性系については，誤りと判定した予測の中に，実際には結合が可能であると考えられる組み合わせが多く含まれているため，予測精度はベンチマークデータと比較して低くなっていた．

実際には，ホモロガスなタンパク質であれば別の複合体由来のタンパク質であっても結合可能である場合もある．しかし，今回用いたデータセットには異なる生物種や生理的条件が混在していたため，ドッキング後処理システムの性能評価の観点からは，再ドッキング以外を安易に正解と見なすことを避けた．

一方で，ドッキング後処理を利用した未知相互作用の発見という応用を考えると，対象とするデータについての既知の結合可能性を加味して，後処理システムの精度を評価する方が望ましい．

**三量体以上のタンパク質複合体**

今回の解析では，1 対 1 のタンパク質のドッキングのみを考慮し，三つ以上のタンパク質で構成される複合体を考慮していない．しかし，実際に機能しているタンパク質複合体には，三量体以上のものも多い．今後はこのような事例も扱えるように手法を拡張することが望ましい．

例えば，あるタンパク質の組み合わせについて得た 1 対 1 のドッキング予測を利用し，別のタンパク質との制約充足的なドッキング可能性を推定するという改良が考えられる．この改良により，競合的に二量体を形成する相互作用対，三量体以上の複合体を形成し得るタンパク質群などを区別して予測でき，生物学的解析への応用可能性を広げることができる．

### 4.3 相互作用対検出の精度

我々は，ドッキング予測を用いたタンパク質間相互作用の解析において，実際の結合状態を正確に予測するドッキング問題と，多数のタンパク質の組み合わせから相互作用の有無を検出する相互作用対検出問題の二点に取り組んできた．

評価手法やパラメータにもよるが，ドッキング問題ではウォード法によるクラスタリングが最適であるという結果を得たのに対し，相互作用対検出問題では現時点では群平均法のほうが精度が良いという結果を得ている[1)]．

これらの二つの問題は本来別の性質をもつと考えられる．相互作用対の検出においては，結合構造を正確に導くよりも，どの程度二つのタンパク質が相互作用しやすいかという親和性の判定に重点をおいた手法の改良が必要になると思われる．

今回提案した相互作用対の検出手法は，走化性系とベンチマークデータとで，高い精度を得られる評価パラメータの範囲が一致しなかった．データセットが異なる場合でもロバストに利用できる解析手法についても議論していく必要がある．

## 5. 結　　論

細菌の走化性に関連する 17 の複合体データをもとに，$17 \times 17$ の組み合わせの中から相互作用するタンパク質を検出するプログラムを適用した．タンパク質の立体構造をパスウェイ解析に用いる手法は今後重要な研究方法の一つになると期待される．生物学的研究への応用に際しては，一つのタンパク質に二つ以上のタンパク質が結合するなど，本システムでは想定していなかった状況が課題となる．ドッキングから制約充足的に三量体以上の複合体形成の可能性を検討するなどの改良が考えられる．

**謝辞**　本研究の一部は，文部科学省事業「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」の支援により実施されている．

## 参 考 文 献

1) 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，第 13 回バイオ情報学研究会予稿集, BIO-13-5 (同時発表) (2008).
2) Falke, J.J., Bass, R.B., Butler, S.L., Chervitz, S.A. and Danielson, M.A.: The two-component signaling pathway of bacterial chemotaxis: a molecular view of signal transduction by receptors, kinases, and adaptation enzymes, *Annu Rev Cell Dev Biol*, Vol.13, pp.457–512 (1997).
3) 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).
4) 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, pp.214–216 (2005).
5) 松崎裕介：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，東京工業大学学士研究論文 (2008 年 2 月).
6) O'Connor, C. and Matsumura, P.: The accessibility of cys-120 in CheA(S) is important for the binding of CheZ and enhancement of CheZ phosphatase activity, *Biochemistry*, Vol. 43, pp. 6909–6916 (2004).
7) Blat, Y. and Eisenbach, M.: Oligomerization of the phosphatase CheZ upon interaction with the phosphorylated form of CheY. The signal protein of bacterial chemotaxis, *J Biol Chem*, Vol.271, No.2, pp.1226–1231 (1996).