NAIST-IS-DD0461033

博士論文

複合体立体構造モデルを用いた
タンパク質間相互作用予測

福原 直志

2007 年 6 月 1 日

奈良先端科学技術大学院大学
情報科学研究科 情報生命科学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
博士(理学) 授与の要件として提出した博士論文である。

福原 直志

審査委員：

箱嶋 敏雄 教授　　(主指導教員)
小笠原 直毅 教授　　(副指導教員)
川端 猛 准教授　　(副指導教員)

複合体立体構造モデルを用いた

タンパク質間相互作用予測*

福原直志

内容梗概

タンパク質間相互作用は多くの生体内プロセスを支えており、相同タンパク質の中から特異的に相互作用する相手タンパク質を見つけることは、シグナル伝達などの細胞機能の解明に重要である。よって、タンパク質間相互作用を決定するために様々なハイスループットの実験手法が開発され、多量のデータが得られてきた。しかしながら、これらの実験手法が適用された生物種は少なく、その実験精度にも限界があるため、計算機を用いてアミノ酸配列や立体構造情報からタンパク質間相互作用を予測する手法の開発が望まれている。

本研究では、ホモロジーモデリングされた複合体立体構造に基づいて、相互作用するタンパク質を予測する手法を開発した。これは、本質的には相同なタンパク質ペアの中から特異的に相互作用するタンパク質ペアを予測することを意味する。先行研究で用いられてきたコンタクトエネルギーに加え、新たにシンプルな静電エネルギーとテンプレート構造との配列類似度の二つのスコアを導入し、これらの単独および結合スコアを計算することで、各複合体立体構造モデルの妥当性を推定した。予測結果については、DIP データベースに登録されているタンパク質ペアを相互作用するペア、登録されていないペアを相互作用しないペアとする新たな基準を設け、両者の識別力によって評価を行った。

本手法を酵母のヘテロの全タンパク質ペアに適用した。10,325 個のタンパク質の二量体モデル構造が作成され、そのうち 417 個が相互作用するタンパク質ペアとして DIP に登録されていた。Recall-Precision プロットと F-measure

の最大値による評価の結果、配列類似度スコアは立体構造に基づくスコアよりも高い識別力を有していることが分かった。しかし、配列類似度にコンタクトエネルギーを結合したスコアを用いることで識別力を有意に改善することができた。

これらの結果を踏まえ、いくつかの事項に関して検討を行った。まず、複合体立体構造モデルに基づく先行研究(Davis *et al.*, *Nucleic Acids Res.*, **34**, 2943-2952 (2006))の手法と本手法の性能比較を行った。先行研究はコンタクトエネルギーのみならず細胞内局在情報も特徴量として用いているため、本研究のコンタクトエネルギーよりは優れていたが、本研究の配列類似度には及ばないことが分かった。また、これまで評価データとして用いてきた相互作用するタンパク質ペアと相互作用しないペアの中から、細胞内局在情報を導入し新たにより信頼性の高い評価データを選択した。この高信頼性データを用いて、再度、評価を行ったところ特徴量の識別力の優劣に変化は見られず、配列類似度とコンタクトエネルギーを結合した場合に最良の性能が得られるという、本研究の結論が評価データの信頼性に依存しないことを確認することができた。

最後に、本研究手法を他の研究者が自由に利用できるように、ウェブサーバを開発した。これは、二つの標的配列を入力するとテンプレート構造の候補を列挙し、ユーザーが指定したテンプレートに基づいて作成した複合体立体構造モデルをエネルギー値とともに表示するものである。酵母以外のタンパク質配列に対しても適用可能であり、有用性の高いツールが構築できたと考えている。

キーワード

タンパク質間相互作用，ホモロジーモデリング，結合特異性，配列類似度，コンタクトエネルギー

---

*奈良先端科学技術大学院大学 情報科学研究科 情報生命科学専攻 博士論文,

NAIST-IS-DD0461033, 2007 年 6 月 1 日.

Prediction of Protein-Protein Interaction Using Homology-modeled Complex Structure *

Naoshi Fukuhara

Abstract

Protein-protein interactions support most biological processes, and it is important to find specifically interacting partner proteins among homologous proteins for elucidating cell functions such as signal transduction systems. Various high-throughput experimental methods for identifying these interactions have been invented, and used to generate a huge amount of data. Because these experiments have been applied to only a few organisms, and their accuracy is believed to be limited, it is desired to develop computational methods for predicting protein-protein interactions from their amino acid sequences or tertiary structural information.

In this study, we describe a prediction method of specific interacting proteins based on homology-modeled complex structures. In other words, we predicted specific interacting protein pairs among homologous protein pairs. We employed the statistical residue-residue contact energy used in the previous study, and two types of the new scores, simple electrostatic energy and sequence similarity between target sequences and template structures. The validity of each protein-protein complex model was measured using their single and combined scores. We evaluated our method in this study by discriminating interacting and non-interacting protein pairs. We defined yeast protein pairs registered in DIP as interacting, and not registered as non-interacting.

We applied our method to all the hetero protein pairs of *Saccharomyces cerevisiae*. In total, 10,325 protein dimer models of the protein pairs were generated, 417 pairs of them were registered in DIP as interacting protein pairs. Recall-precision plots and maximum F-measures show that sequence similarity

has a much higher discrimination power than the other structure-based scores, but using contact energy results in significant improvement over predictions using sequence similarity alone.

Based on the results, we did additional analyses to confirm the performance of our prediction method. First, we compared our predictions with the previous study (Davis *et al.*, *Nucleic Acids Res.*, **34**, 2943-2952 (2006)) which is based on contact energy, subcellular localization data and functional annotation. Their performance was better than that of our contact energy, nonetheless, it was much inferior to that of the sequence similarity. Second, we chose a reliable dataset of interacting and non-interacting protein pairs by experimentally determined subcellular localization data. The F-measure using the reliable dataset showed that the rank of the features did not change; the combined score of the sequence similarity and contact energy still had the best discrimination power.

Finally, we developed a WWW server to make our prediction service freely available to other researchers. The server accepts two target protein sequences from users, it lists the candidates of the template structures, and shows the homology-modeled complex structures with their energy values generated based on the templates selected by users. As our server is also applicable to protein sequences of any organisms, we believe that it is a useful tool for many researchers.

Keywords:

protein-protein interaction, homology-modeling, binding specificity, sequence similarity, contact energy

*Doctoral Dissertation, Department of Bioinformatics and Genomics, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD0461033, June 1, 2007.

目次

第１章　序論

１.１　タンパク質間相互作用解析

タンパク質間相互作用は、シグナル伝達、酵素活性、複製、転写、翻訳などの多くの重要な細胞機能を支えている。例えば、G タンパク質は GEF や GAP と相互作用することで活性が ON、OFF に調節される分子スイッチとして機能し、下流のシグナル伝達分子を制御している。また、サイクリンはサイクリン依存性キナーゼ（Cdk）と相互作用することで Cdk の活性を調節し、細胞周期の複製装置を起動させる（図１－１）。

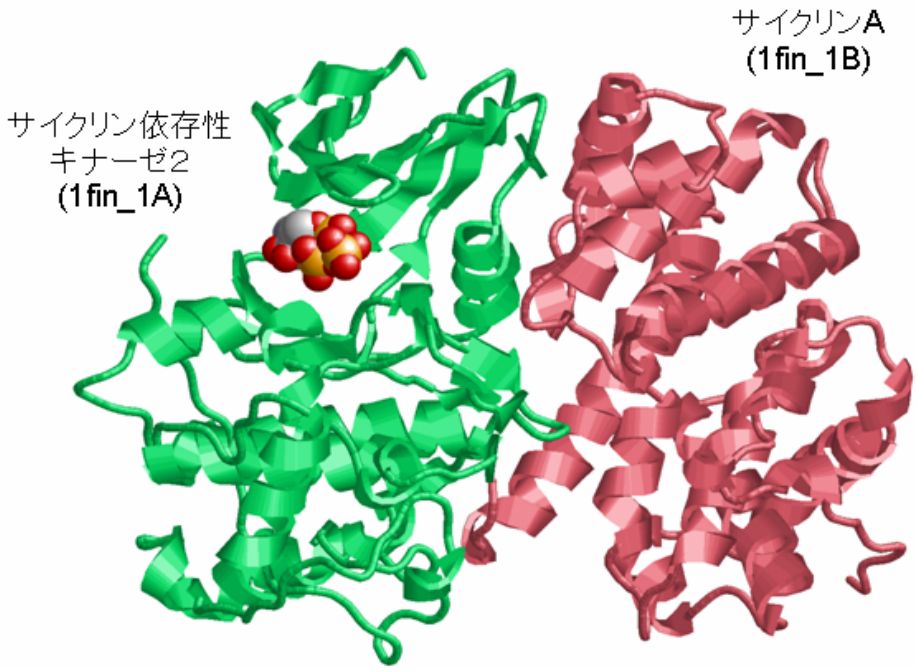


図１－１　サイクリン A とサイクリン依存性キナーゼ２の相互作用

相同なタンパク質の中から特異的に相互作用する相手タンパク質を見つけることは細胞機能を支える上で極めて重要である。細胞内には多くの相同なタンパク質ドメインが存在し、それらは自分自身と特異的に相互作用する一連の相手タンパク質を有している（図１－２）。

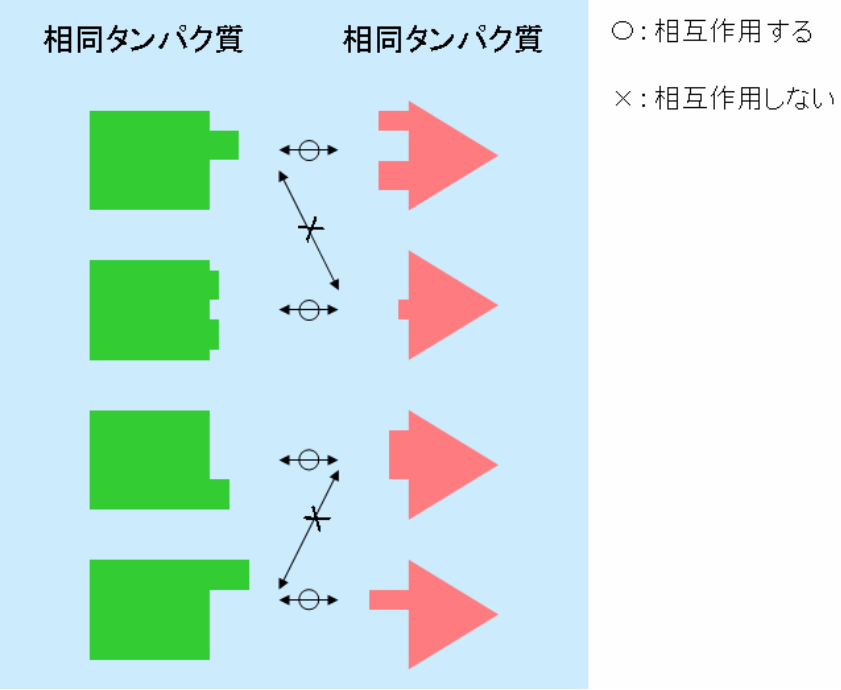


図１－２　タンパク質間相互作用の特異性の概念図

特に、タンパクキナーゼ、G タンパク質、転写因子のようなシグナル伝達系のタンパク質は細胞内に多くの類似する相同タンパク質を持つ(Rubin *et al.*, 2000)。これらのタンパク質の結合特異性は細胞内の複雑かつ強固なシグナル伝達系の基盤となっている(Gomperts *et al.*, 2002)。

このように様々な細胞機能を理解する上で、タンパク質間相互作用を解析することは極めて重要である。プロテオミクスにおけるタンパク質間相互作用解析の手法として、酵母２ハイブリッド法やタンデムアフィニティ精製（TAP）などが開発された。

酵母２ハイブリッド法は、酵母の転写因子 GAL4 のドメイン構造を利用して、レポーター遺伝子の発現から標的タンパク質間の相互作用を検出する方法である（図１－３）。

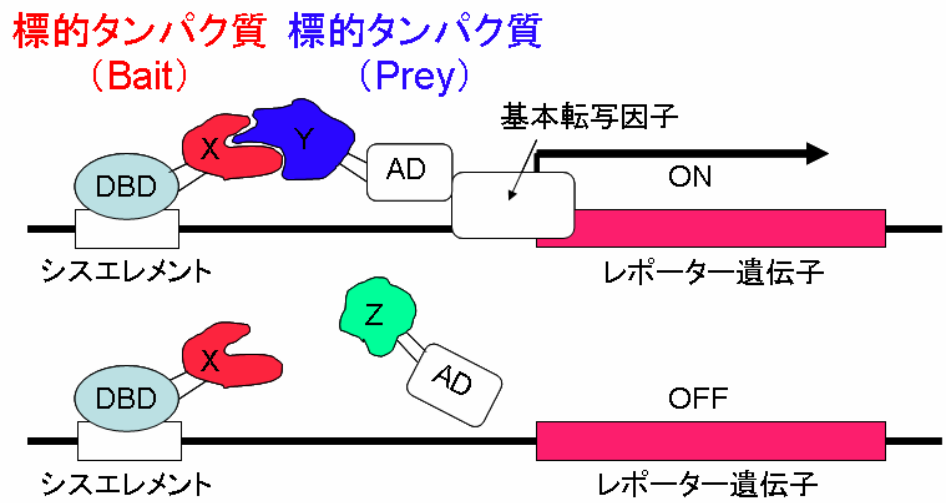


図１－３　酵母２ハイブリッド法の原理

二つの標的配列 X (Bait)と Y (Prey)が相互作用すると、この結合を介して DBD と AD がプロモーター上で一体化し、転写因子の機能が構築され、レポーター遺伝子の発現が誘導される。レポーター遺伝子を必須遺伝子とすれば、コロニーの生成を通して標的タンパク質間の相互作用を検出することができる。この手法の利点としては、*in vivo* での相互作用を解析できること、標的タンパク質の結合領域の同定にも応用できることなどが挙げられる。しかしながら、標的タンパク質のドメインをどのように切断するかによって結果が異なる場合があることや、標的タンパク質が直接基本転写因子を活性化することで偽陽性が検出される場合があるなどの問題点がある。

TAP においては、標的タンパク質にカルモジュリン結合ペプチド、TEV プロテアーゼ切断部位およびプロテイン A を融合したタンパク質を宿主細胞に発現させる。この融合タンパク質に相互作用するタンパク質を会合させて複合体を形成させる。タグをアフィニティクロマトグラフィーの担体に結合させ複

合体を精製し、複合体を構成するタンパク質を MS / MS で同定する（図１－４）。

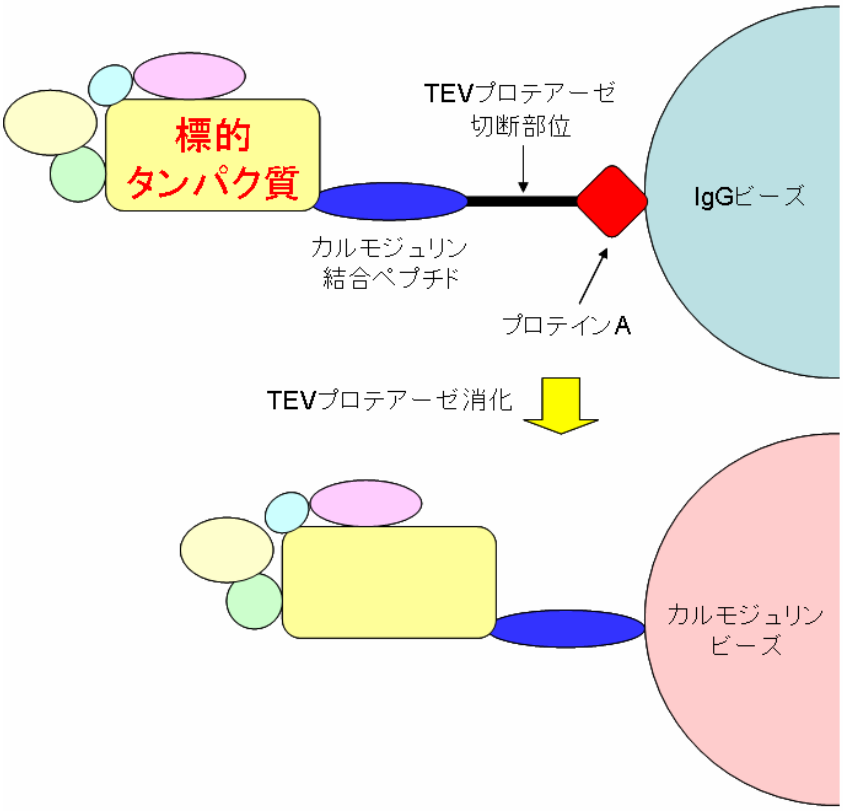


図１－４　タンデムアフィニティ精製（TAP）

カルモジュリン結合ペプチドとプロテイン A の二種類の性質の異なるタグを用いることによって、標的タンパク質の精製度が飛躍的に高まる。

酵母２ハイブリッド法も TAP もゲノムワイドに解析が可能であり、近年、高速かつ自動的に相互作用を検出できるハイスループットの手法へと改良され、多量のタンパク質間相互作用データを生み出した(Uetz *et al.*, 2000; Ito *et al.*, 2001; Gavin *et al.*, 2002; Ho *et al.*, 2002; Gavin *et al.*, 2006; Krogan *et al.*, 2006)。これらの相互作用データは DIP、MIPS、BIND などのデータベースに登録されている(Bader *et al.*, 2003; Salwinski *et al.*, 2004; Guldener *et al.*, 2006)。DIP データベース(Database of Interacting Proteins) [http://dip.doe-mbi.ucla.edu/]には、ショウジョウバエ、出芽酵母、大腸菌、線虫、ヒト、ピロリ菌などのタンパク質間相互

作用データが登録されており、酵母に関しては 2007 年 1 月時点で、18,243 個の相互作用するタンパク質ペアが登録されている（図１－５）。

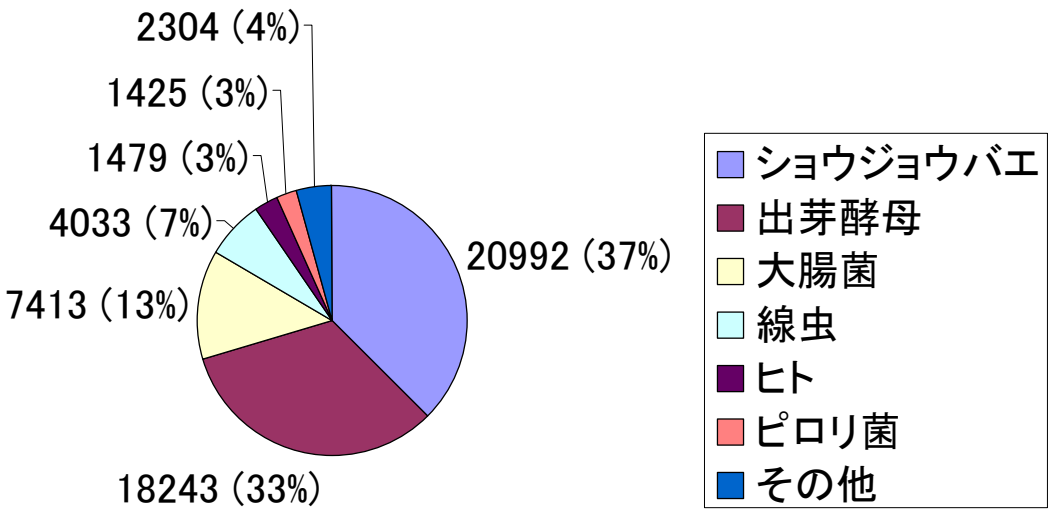


図１－５　DIP データベースに登録されている生物種の統計

本研究では、データ数が多く、多くのグループによって研究されている出芽酵母の DIP データを対象とすることにした。DIP データベースに登録されている出芽酵母のタンパク質間相互作用を検出した実験手法の統計を取ると、２ハイブリッド法、TAP、免疫沈降などのハイスループットの手法が大半を占めていることが分かる（図１－６）。

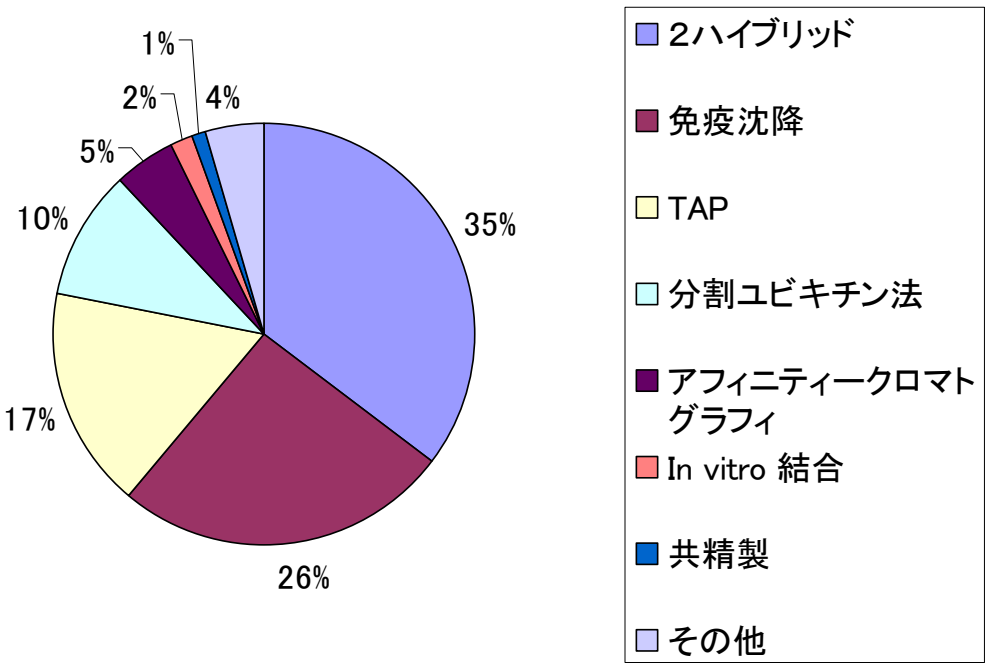


図１－６　DIP データベースに登録されている相互作用を検出した実験手法

一方、DIP データベースにはハイスループットのデータだけでなく、古典的な手法で一つ一つのタンパク質間相互作用を検出したロースループットのデータも含まれている。ロースループットの実験には、試験管などの人為的にコントロールされた環境下で標的タンパク質を結合させる *In vitro* 結合、レーザー光の共鳴状態の変化から相互作用を検出する表面プラズモン共鳴、X 線結晶解析で直接複合体の立体構造を解く方法など、様々な手法が存在する。

ハイスループットのゲノムワイドスクリーニング法は大規模な解析であるため、４００種以上の生物種の完全長ゲノム配列が決定されている今日でも、少数の生物種にしか適用されていない。大量のゲノム配列データと比較的適用範囲の狭い相互作用データとの間のギャップを埋めるため、計算機を用いて、アミノ酸配列情報からタンパク質間相互作用を予測する手法の開発が取り組まれてきた(Salwinski and Eisenberg, 2003; Bork et al., 2004)。

１.２　先行研究

タンパク質間相互作用を予測するために、遺伝子融合法(Enright *et al.*, 1999; Marcotte *et al.*, 1999)、系統プロファイル法(Pellegrini *et al.*, 1999)、進化学的手法(Matthews *et al.*, 2001; Wojcik and Schachter, 2001; Wojcik *et al.*, 2002; McDermott and Samudrala, 2004)などの様々なアミノ酸配列情報に基づくタンパク質間相互作用予測の手法が開発されてきた。

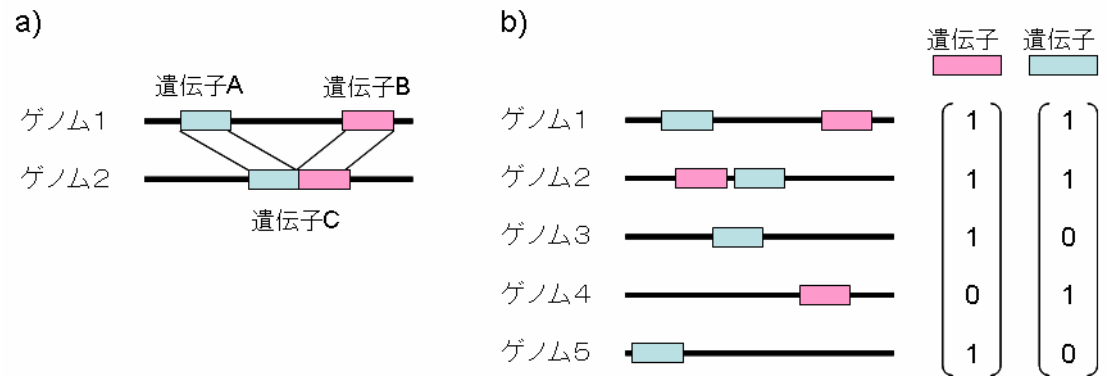


図１－７　遺伝子融合法と系統プロファイル法

遺伝子融合法とは、ある生物種のゲノムでは、それぞれ別の遺伝子としてコードされている二つのタンパク質が、別の生物種のゲノムでは融合した一つの遺伝子として見出される場合、その二つの遺伝子産物は機能的にリンクしていると推測する手法である（図１－７(a)）。この方法は、遺伝子融合で関係づけられるものでなければ適用できないという制限がある。系統プロファイル法では、N 本の完全長ゲノムが与えられた場合を考え、各遺伝子について、N 次元のベクトルを考える（図１－７(b)）。*i* 番目のゲノムにその遺伝子がコードされていれば１、コードされていなければ０の値が入る。この N 次元ベクトルがその遺伝子の系統プロファイルであり、この構成が類似する遺伝子は、進化の過程でのゲノム上における有無の挙動が相関することから機能的にリンクしていると推測する手法である。この方法は、原核生物のように完全長ゲノム配列が決定された生物種が大量に存在する場合でなければ適用できず、真核生物には適用が難しいという問題点がある。

最近では、タンパク質間複合体の立体構造に基づく予測手法も提案されてきた(Aloy and Russell, 2002; Lu *et al.*, 2002; Lu *et al.*, 2003; Davis *et al.*, 2006; Grigoryan and Keating, 2006)。これらの立体構造に基づく手法には、共通の手順が用いられている（図１－８）。

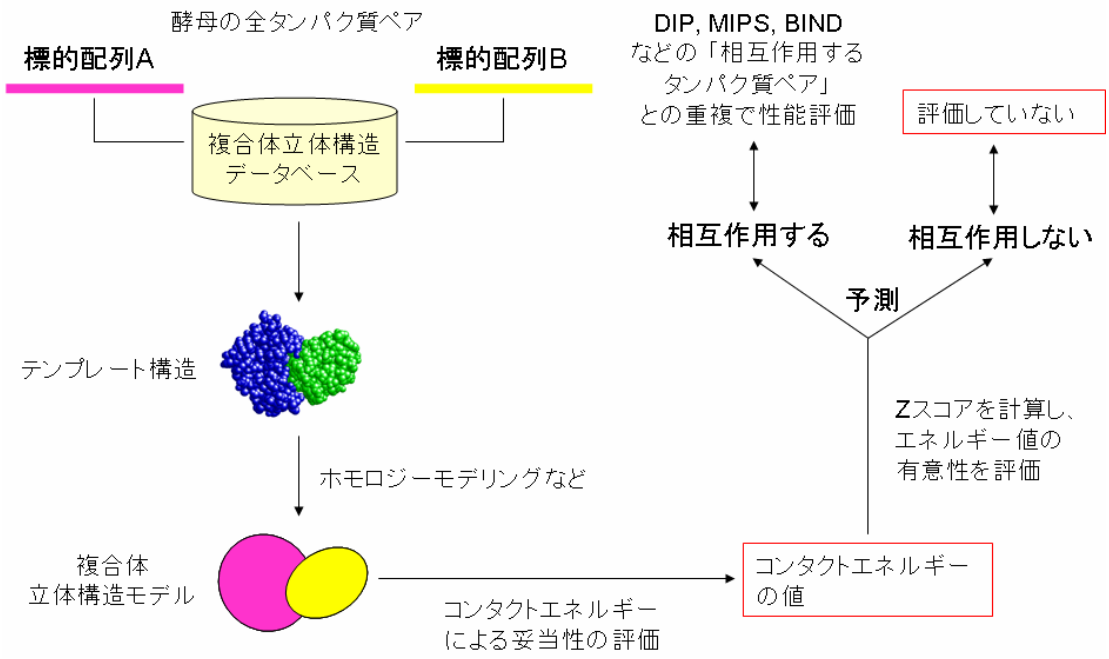


図１－８　立体構造に基づく先行研究で用いられた共通の手順

最初に、酵母の全タンパク質ペアから二つの標的配列を取り出し、それらの複合体モデル構造をホモロジーモデリングまたはスレッディングによって生成する。ホモロジーモデリングは、標的配列と相同性の高い既知のタンパク質立体構造をテンプレートとしてモデル構造を得る方法であり、スレッディングは、標的配列が既知の構造データベース内のどの構造と類似しているかを検出する方法である。Alloy & Russell はホモロジーモデリングのテンプレート構造を検索するために BLAST を用い、Lu らは多量体のモデリング専用に独自に開発したスレッディングプログラムを用いている。これらの二つの研究は残基単位のモデルを採用している。それに対し、Davis らは MODBASE(Pieper *et al.*, 2006)から得た全原子モデルを採用している。

次に、モデル構造の妥当性は相互作用エネルギーによって評価される。これらの三つの研究ではコンタクトエネルギーが用いられている。コンタクトエネルギーとは、異なるポリペプチド鎖に属し、コンタクトするアミノ酸残基のエネルギーを表すものである。最後に、相互作用エネルギーの値は様々な統計スコアによって検定される。Alloy & Russell と Davis らはランダム配列を基準としたZスコアを採用している。Zスコアとは、基準となる分布の平均を引き、標準偏差で割る変換を行うことで正規化されたスコアのことをいい、このZスコアを用いれば標準正規分布中の位置が示されるので有意性の検定や異なる分布間相互の比較を行うことができるようになる。Lu らも Z スコアを用いているが、ライブラリの全テンプレート構造を基準としている。

１.３　本研究の目的

本研究の目的は、複合体モデル構造の中から「相互作用するタンパク質ペア」と「相互作用しないタンパク質ペア」を識別する手法を開発することである。このように研究目的を設定した根拠は、相同なタンパク質ペアの中での相互作用の特異性に着目することが重要だと考えたからである。

前節で述べた先行研究の予測精度は、相互作用すると予測されたタンパク質ペアが DIP、MIPS、BIND などの実験で決定されたタンパク質ペアをどの程度カバーできたかによって評価されているが、相互作用しないと予測されたタンパク質ペアについては、あまり詳細に評価されていない。本研究では、図１－８の赤枠に示す部分を改良することにした。まず先行研究で用いられたコンタクトエネルギー以外の特徴量を導入しようと考え、静電エネルギーとテンプレート構造との配列類似度の二つの特徴量を新たに導入することにした。さらに相互作用しないと予測されたタンパク質ペアについても評価するべきだと考え、特徴量の識別力を「相互作用するタンパク質ペア」と「相互作用しないタンパク質ペア」の分離度で評価することとした。

研究を行うにあたり、問題となるのは正解とする「相互作用するタンパク質ペア」と「相互作用しないタンパク質ペア」の実験データの定義が定まっていないことである。このような問題が生じる一つの原因は、タンパク質間相互作用のハイスループットの実験には信頼性が低いデータが含まれていると考え

られるからである(Deane *et al.*, 2002; von Mering *et al.*, 2002; Sprinzak *et al.*, 2003)。Sprinzak らが、酵母２ハイブリッドによる相互作用データの信頼度を細胞内局在や細胞機能の一致度を用いて推定したところ、Uetz らのデータ(Uetz *et al.*, 2000)の信頼度は 50%程度、Ito らのデータ(Ito *et al.*, 2001)の信頼度も 50%程度であり、約半数が偽陽性であるとみなされた。また、相互作用しないタンパク質ペアを定義することはさらに難しい。

以上の問題点を踏まえた上で、本研究の手法の評価を行うために、まず DIP には全タンパク質間相互作用が登録されていると仮定した。このように想定しても、出芽酵母に関してはそれほど問題がないと考えられる。なぜなら、酵母はタンパク質間相互作用において最もよく知られたモデル生物であり、これまでに大量の実験データが蓄積されてきたからである。複合体立体構造モデルが構築できる酵母の全タンパク質ペアにおいて、DIP データベースに登録されているペアを「相互作用するタンパク質ペア」、登録されていないペアを「相互作用しないタンパク質ペア」とラベル付けした。その上で、特徴量の識別力を相互作用するタンパク質ペアと相互作用しないペアの分離度で評価することにした。しかしながら、DIP データベースに登録されていないが細胞内で起こりうる相互作用はまだ存在していると考えられ、特に相互作用しないタンパク質ペアの信頼性が低いことは無視できない。これを踏まえ、相互作用するタンパク質ペアと相互作用しないペアの中から、細胞内局在情報を導入し新たにより信頼性の高い評価データを選択して、再評価を行うこととした。

１.４　本論文の概要

本論文の概要は次のとおりである。第２章では、最初に予測法の概要について述べ、本研究で用いた複合体立体構造のデータセットおよび酵母のヘテロのタンパク質ペアのモデル構造の作成方法について解説する。次に、予測法で用いた三種類の特徴量の、コンタクトエネルギー、静電エネルギーおよびテンプレート構造との配列類似度について解説する。第３章では、まず複合体立体構造モデルの精度の検証を行い、その後、本手法を出芽酵母の全てのヘテロのタンパク質ペアに適用した結果について述べる。予測結果については、相互作用するタンパク質ペアと相互作用しないペアの間の識別力によって特徴量の性

能を評価した。具体的には、Recall-Precision プロットと F-measure の最大値を用いてそれぞれの三つのスコアとそれらの結合スコアによる性能比較を行った。第４章では、これまでの研究結果に対して議論を行う。先行研究との性能比較や研究結果に対する多面的な考察、DIP データベースに含まれていない相互作用の検出などを検討し、最後に、細胞内局在データを用いたより信頼性の高い評価基準へと本手法を適用した結果について報告する。第５章では、本手法のウェブサーバを開発したことについて報告し、第６章にて結論をまとめる。

第２章　手法

２.１　予測法の概要

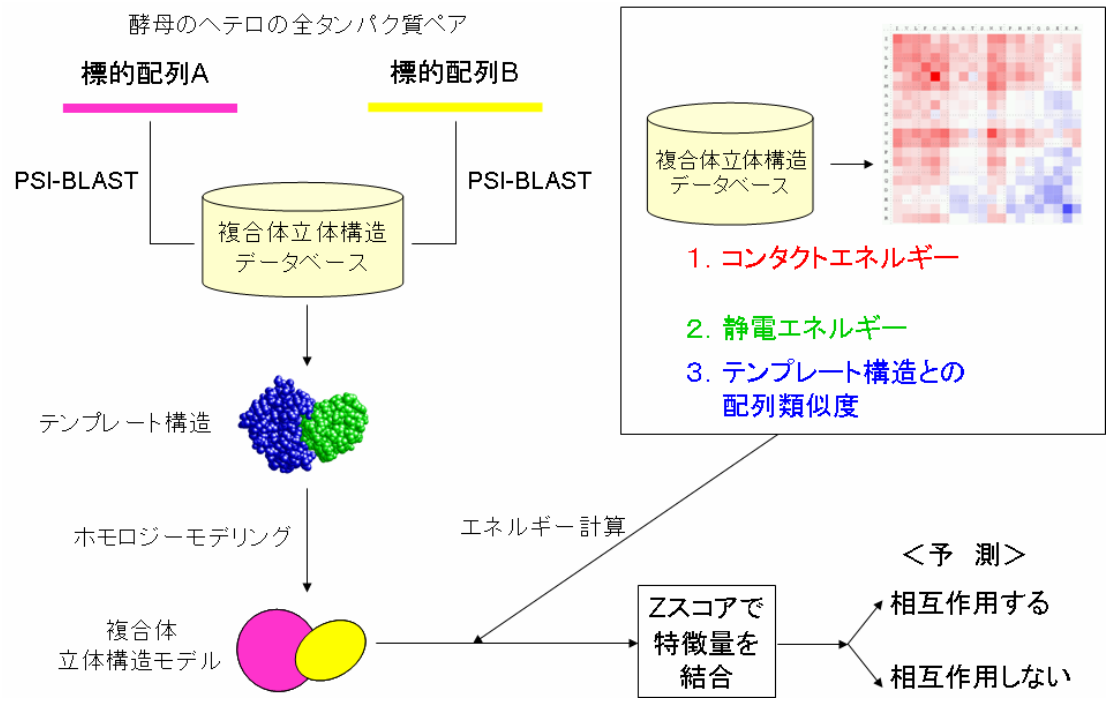


図２－１　予測法の概要

予測法の概要を図２－１に示す。最初に、ホモロジーモデリングによって二つの標的タンパク質の複合体立体構造を予測する。既知の二量体立体構造を構成するタンパク質配列をライブラリ、二つの標的タンパク質配列をクエリとし

て相同性検索を行う。それぞれの標的タンパク質に対して相同な二つのタンパク質からなるテンプレートの複合体立体構造が見つかったら、そのテンプレートに基づいて標的タンパク質の複合体立体構造をモデリングする。

モデル構造の妥当性を評価するために三種類のスコアを用いた。第一に、コンタクトエネルギーであり、これは序論で述べた三つの先行研究に用いられたものである。第二に、シンプルな静電エネルギーを導入した。これは、タンパク質間の長距離相互作用および結合特異性は静電相互作用によって与えられると考えられるからである。第三に、テンプレート構造との配列類似度に基づくスコアを用いた。相互作用することが知られているタンパク質ペアとの配列類似度は、配列ベースの予測にしばしば用いられてきた(Matthews *et al.*, 2001; Wojcik and Schachter, 2001; Wojcik *et al.*, 2002; McDermott and Samudrala, 2004)。しかしながら、この配列類似度は今まで立体構造の特徴量と組み合わせて用いられることはなかった。本研究ではこれらの三つのスコアを単独あるいは結合スコアとして用いる。結合スコアとするため、各特徴量はランダム配列を基準としたZスコアへと変換する。先行研究とは異なり、本研究ではエネルギーの平均と分散を解析的に推定した。計算されたZスコアが閾値を超えたタンパク質ペアについては相互作用すると予測し、閾値を超えなかったものについては相互作用しないと予測する。

２.２　複合体立体構造のデータセット

テンプレート構造のライブラリやコンタクトエネルギーの値の推定に用いるため、複合体立体構造のデータセットを準備した。これらのセットは PQS サーバ(Henrick and Thornton, 1998)から得た冗長でないヘテロダイマーの代表立体構造データからなり、以下の手順に従って生成した。最初に、PQS サーバに含まれる全ての多量体を二量体に分割する。PQS サーバは PDB に登録されている立体構造を結晶学的対称群に従って、全ての候補となる複合体を生成し、その中からクリスタルパッキング（天然には存在せず結晶化によって生成する相互作用）を除いた複合体の立体構造のみを抽出し、データベース化したものである（図２－２）。

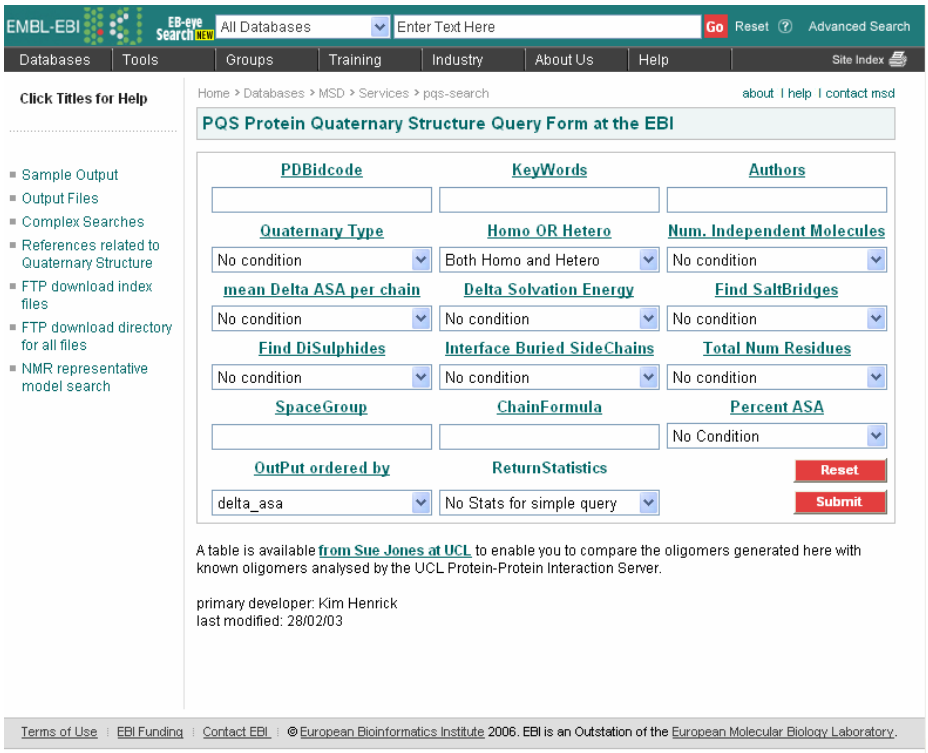


図２－２　PQS サーバのトップページ（http://pqs.ebi.ac.uk/）

PQS データは 2006 年 4 月 14 日にダウンロードしたものを用い、相互作用残基の数が 5 個より少ない二量体はデータセットから除去した。これらはヘテロダイマー（構成タンパク質間の同一残基率は 50%以下）からなり、相互作用残基は、他方のタンパク質鎖の重原子から 4 Å 以内に位置する重原子を少なくとも一つ持つアミノ酸残基として定義した。次に、これらの二量体をシングル・リンケージ・クラスタリング(Johnson and Wichern, 1998)によって、類似度に応じてクラスターにまとめた。二量体間の類似度は、対応するタンパク質間の二つの同一残基率のうち低い方の同一残基率として定義した。それぞれのクラスターの中から最も多くの相互作用残基を持つ二量体を代表として一つ抽出し、それらを集めて代表データセットとした。複合体の類似度の閾値を変えることで二つのタイプの代表データセットを準備した。複合体の類似度の閾値はそれぞれ 40%と 95%である。前者は 1,687 個のヘテロダイマーで、コンタクトエネルギーの計算のためのデータセットとして用いられる。後者は 2,635 個のヘ

テロダイマーで、ホモロジーモデリングのテンプレート構造ライブラリとして用いられる。

２.２.１　酵母のヘテロのタンパク質ペアのモデル構造の作成

UniProt データベース（ver. 49.4）(Wu *et al.*, 2006)から、5,314 個の出芽酵母のアミノ酸配列を抽出した。5,314 個の酵母のタンパク質配列を用いて全てのヘテロのタンパク質ペアを作成し、相互作用するかどうかの予測対象とする。それぞれの酵母のアミノ酸配列の配列プロファイルを生成するために、2006 年 9 月 22 日にダウンロードした nr データベースに対して PSI-BLAST (Altschul *et al.*, 1997)を実行した。E-value の閾値は 0.001 にセットし、繰り返しの数は三回とした。次に、生成された配列プロファイルを用いて、上記のテンプレート構造ライブラリに対して PSI-BLAST を実行した。それぞれの標的タンパク質ペアを構成するタンパク質に対して相同である二つのタンパク質が存在し、それらが結合した複合体がデータベース内に存在していたら、それをテンプレート構造の候補とする。さらに、標的タンパク質配列とテンプレート構造間の二つのアラインメントにおいて、アラインメントされた相互作用残基の割合は 50%以上であり、アラインメントされた相互作用残基の数は 10 以上でなければならない（図２－３）。

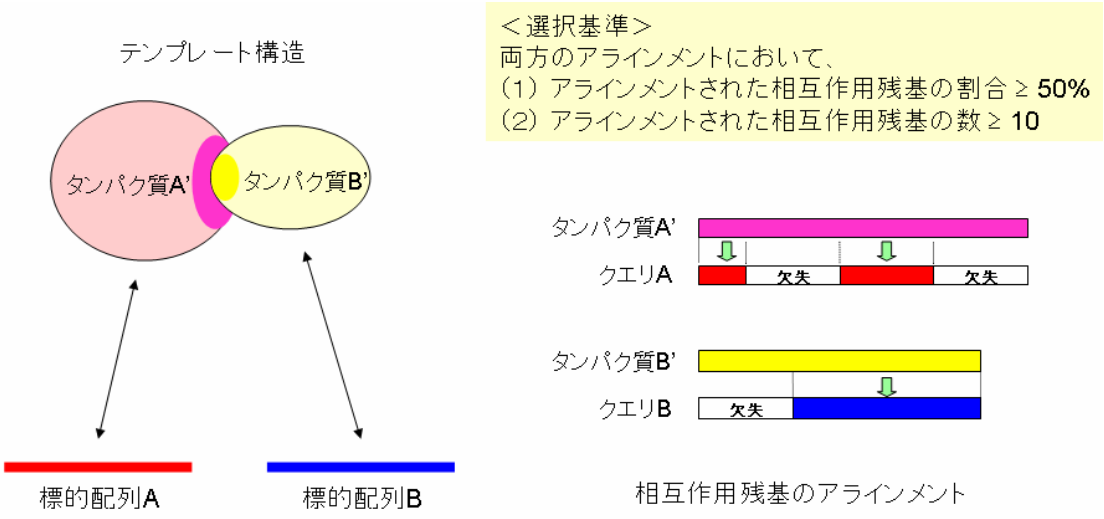


図２－３　テンプレート構造を選択する基準

これらの基準を満たすテンプレート複合体立体構造が複数見つかった場合は、小さい方の配列類似度が最大となるテンプレートを選んだ。

複合体立体構造のモデリングに関しては、MODELLER (Marti-Renom *et al.*, 2000) を用いて挿入残基の構築、置換側鎖のモデル構築を行うことも不可能ではなかったが、本研究では多くのタンパク質ペアを扱える計算負荷の小さいモデリング法が必要であったため、テンプレート構造のコンフォメーションをそのまま用い、アラインメントで挿入されたアミノ酸残基を無視し、置換後のアミノ酸残基の側鎖の原子をモデル構造に組み入れないこととした。こうした簡易なモデルを扱って複合体の妥当性を検証するため、後述する２つの立体構造を用いた相互作用エネルギーにおいても、コンタクトエネルギーは Cβ原子座標のみを、静電エネルギーについてはテンプレート構造の原子座標を用いることとした。

２.２.２　相互作用が報告されているペアと報告されていないペア

生成された複合体立体構造のモデルを、そのタンパク質ペアが DIP データベースに含まれているかどうかで「相互作用するタンパク質ペア」と「相互作用しないタンパク質ペア」の二つに分割した。DIP データは 2006 年 1 月 16 日にダウンロードしたものを用いた。DIP データベースは大量のタンパク質間相互作用データを含んでいるが、最新の実験結果の中にはまだ登録されていないものも存在している。もし標的タンパク質ペアに対してほとんど同一の（95%以上の同一残基率）複合体結晶構造が存在していたら、これらのペアはＸ線結晶解析という実験的な根拠があると考えられるため、DIP データベースに登録されていなくても相互作用するタンパク質ペアだとみなすこととした。

２.３　コンタクトエネルギー

コンタクトエネルギーは、元々タンパク質のフォールディングやスレディングのための残基単位のモデルの安定性を評価するために開発されたものである(Miyazawa and Jernigan, 1985; Sippl, 1990; Jones *et al.*, 1992)。これは、Amber (Case *et al.*, 2005) や CHARMM (Brooks *et al.*, 1983)のポテンシャルエネルギーとは異なり、PDB のデータベースの統計頻度から得られるエネルギー関数である。最近では、タンパク質間相互作用の評価にコンタクトエネルギーが適用されている(Keskin *et al.*, 1998; Moont *et al.*, 1999; Glaser *et al.*, 2001; Ofran and Rost, 2003; Lu *et al.*, 2003)。本研究では、コンタクトエネルギーの値を抽出するために一般的なログ・オッズの式を用いた。異なるポリペプチド鎖に属し、コンタクトするアミノ酸残基 a, b のコンタクトエネルギー$e_{con}(a, b)$ は以下の式で推定される。

$$e_{con}(a,b) = -\log \frac{Q(a,b)}{P(a)P(b)} \qquad (1)$$

ここで、$P(a)$, $P(b)$はアミノ酸 a, b が表面に現れる確率であり、$Q(a, b)$はアミノ酸 a, b がタンパク質間相互作用面において互いにコンタクトする確率である。タンパク質の表面残基は、溶媒露出度が 35%以上となるアミノ酸残基として定義し、コンタクトする残基ペアは、異なる鎖に属し Cβ原子が互いに 7 Å 以内に位置しているアミノ酸残基ペアとして定義する。これらの確率はコンタクトエネルギーライブラリ（2.2 節参照）の統計頻度から推定する。アミノ酸残基 a, b 間のコンタクトが相互作用面に多く見られるほど、$e_{con}(a, b)$の値は大きな負の値となる。

推定されたエネルギー値を図２－４に示す。

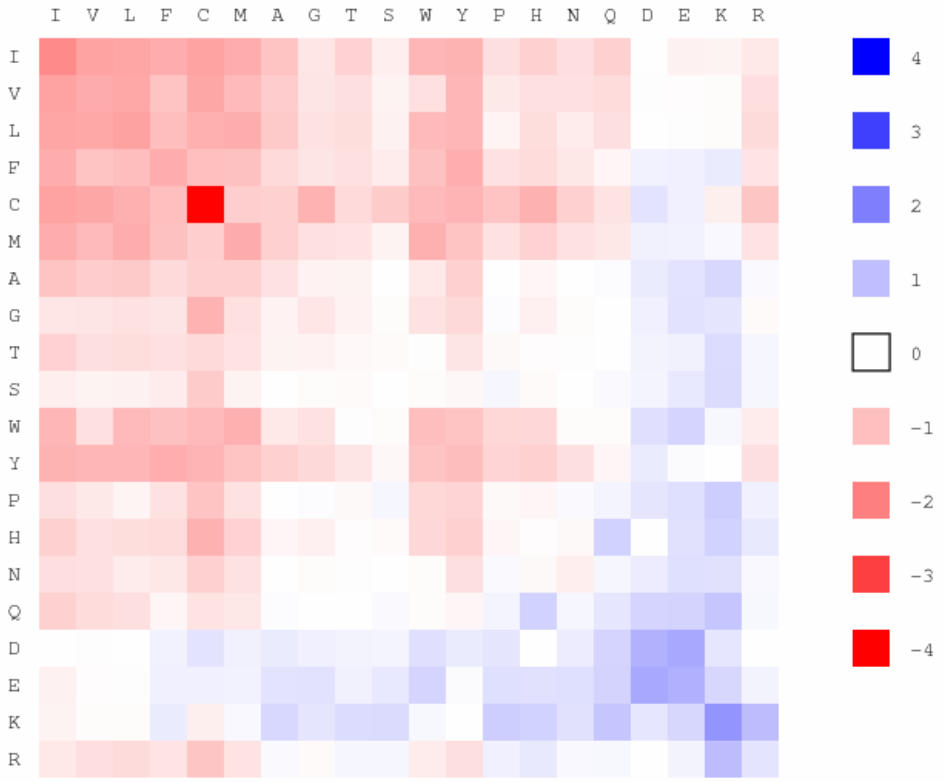


図２－４　コンタクトエネルギー

縦軸と横軸には、２０種類のアミノ酸を疎水性の降順に並べた。エネルギー値は赤（低エネルギー）から青（高エネルギー）で表されている。疎水性残基は互いに引き合い、特にシステイン－システインペアの場合に顕著である。しか

しながら、親水性残基はほとんどの場合反発し、アルギニンーグルタミン酸ペアのような異符号の荷電性残基ペアでさえも反発している。これらの特徴は、先行研究(Moont *et al.*, 1999; Lu *et al.*, 2003)の結果と類似している。

コンタクトエネルギーの総和 E_{con} は、表面残基と非表面残基の両方について全てのコンタクトしているアミノ酸残基ペアについての e_{con} の和であり、次の通りである。

$$E_{con} = \sum_{i,j(i \text{ contacts with } j)}^{N,M} e_{con}(a_i, a_j) \qquad (2)$$

ここで、N, M はそれぞれのタンパク質のアミノ酸残基の総数であり、a_i, a_j は残基 i, j のアミノ酸である。

２.４ 静電エネルギー

静電相互作用は、タンパク質間相互作用に重要な役割を果たしていると言われている(Sheinerman *et al.*, 2000)。複合体モデルの妥当性を確認するために、Shaul & Schreiber によって提案された次式のようなシンプルな静電エネルギーを採用した(Shaul and Schreiber, 2005)。電荷 q_1, q_2 の間の静電エネルギーe_{ele} は、デバイーヒュッケル理論に基づく次の式で計算される。

$$e_{ele}(r, q_1, q_2) = \frac{1}{4\pi\varepsilon_0\varepsilon_r} \frac{q_1 q_2}{r} \frac{e^{-\kappa(r-a)}}{1+\kappa a} \qquad (3)$$

ここで、ε_r は水の比誘電率(= 80)である。変数 r は電荷 q_1, q_2 間の距離であり、κ はデバイーヒュッケル遮蔽パラメータ(= 0.488 Å^{-1})である。パラメータ a は 6 Å とした。

静電エネルギーの総和 E_{ele} は、全ての荷電原子についての e_{ele} の和である。

$$E_{ele} = \sum_{i}^{N} \sum_{j}^{M} \sum_{s \in Q_i} \sum_{t \in Q_j} e_{ele}\left(r_{st}, q_s(a_i), q_t(a_j)\right) \qquad (4)$$

ここで i, j は異なるタンパク質に含まれるアミノ酸残基である。 N, M はアミノ酸残基の総数であり、Q_i, Q_j は残基 i, j に属する荷電原子の集合である。変数 r_{st} は原子 s, t 間の距離である。変数 $q_s(a_i)$はアミノ酸 a_i の原子 s の電荷である。

電荷は単純な形式電荷を用いる。すなわち、電荷-1 をアスパラギン酸やグルタミン酸、電荷+1 をリシンやアルギニンに割り当てる。本研究では置換され

たアミノ酸の側鎖の原子を作り直していないので、モデル構造のどこに電荷を置くかは自明ではない。モデル構造の原子への具体的な電荷の割り当て方については、Shaul & Schreiber によって提案された電荷則を採用した(Shaul and Schreiber, 2005)。標的配列におけるアミノ酸残基の総電荷を、テンプレート構造上の対応する残基のあらかじめ決められた原子の位置に均等に割り当てる。各アミノ酸側鎖の疑似電荷を置く原子は図２－５に赤色で示されている。

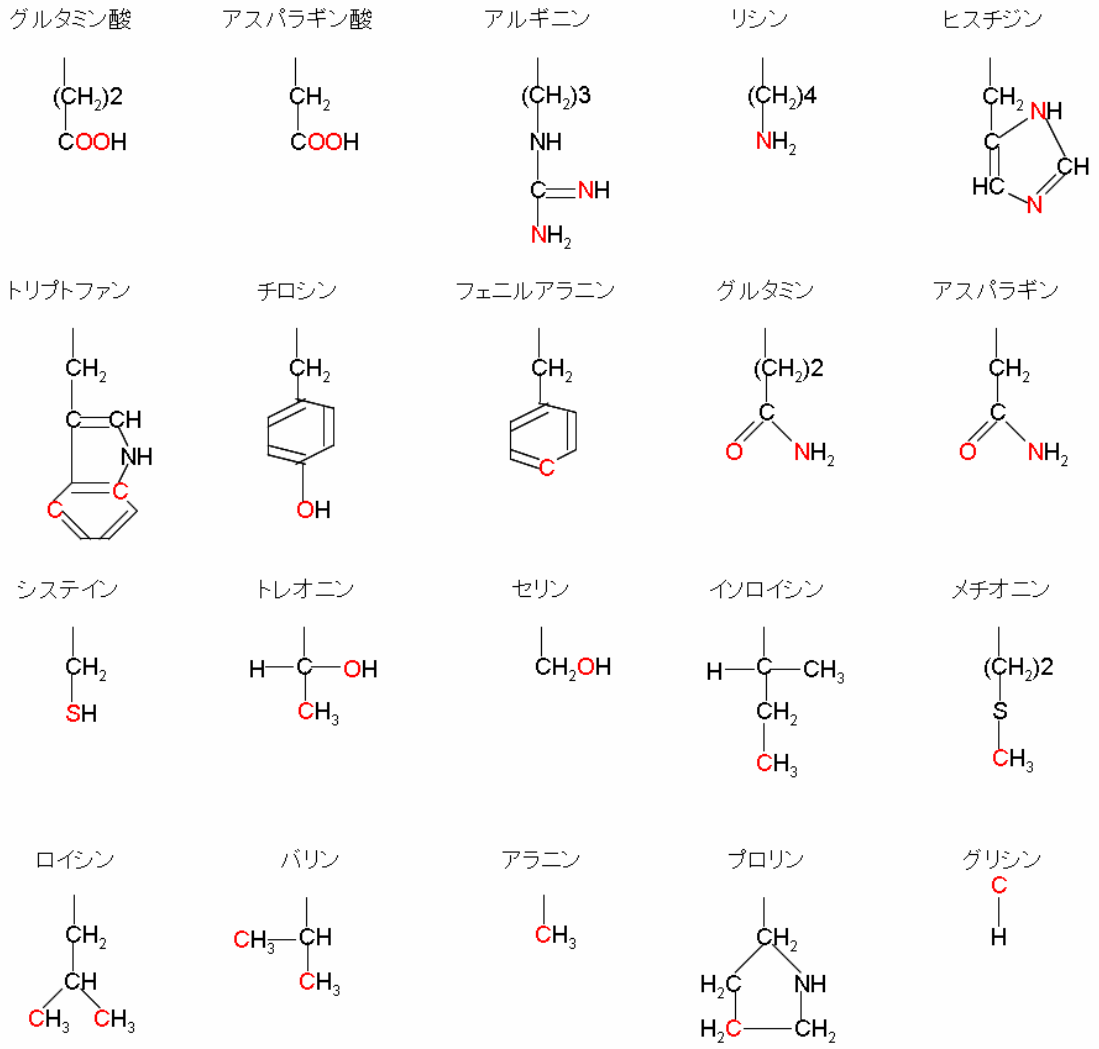


図２－５　電荷を置くアミノ酸の原子

大部分の原子は Shaul & Schreiber の電荷則(Shaul and Schreiber, 2005)から採用したものであるが、プロリンとグリシンの原子を新たに追加し、OXT 原子（酸素末端）と N 末の原子は電荷を置かないことにした。例えば、標的タンパク質のアミノ酸がグルタミン酸でテンプレート構造の対応するアミノ酸がトレオニンである場合、電荷-0.5 をトレオニン残基の原子 OG1 と CG2 の両方に割り当てることになる。

２.５　エネルギーの正規化

コンタクトおよび静電エネルギーを正規化するために、先行研究(Aloy and Russell, 2002; Lu *et al.*, 2002; Lu *et al.*, 2003; Davis *et al.*, 2006)に基づいて Z スコアを導入した。エネルギーE の Z スコアは次のように定義される。

$$Z(E) = \frac{E - Mean[E]}{\sqrt{Var[E]}} \tag{5}$$

ここで、$Mean[E]$と $Var[E]$はそれぞれ、アミノ酸の構成が同じである配列をランダムにシャッフルした場合の E の平均と分散である。平均と分散の値は、先行研究では多数のランダム配列を実際に生成することで求めていたが、本研究では次節で説明するように、表面アミノ酸の頻度から解析的に求める。

２.５.１　ランダム配列に対するコンタクトエネルギーの平均と分散

異なる二つのタンパク質の表面からランダムに一つずつアミノ酸を取り出すことにより、ランダムにコンタクトするアミノ酸ペアを生成したと仮定する。この過程を繰り返してコンタクトするアミノ酸のセットを作った場合、そのコンタクトエネルギーの平均μ_{con} と分散σ_{con}^2 は次のように計算される。

$$\mu_{con} = \sum_{a \in A} \sum_{b \in A} \{ e_{con}(a,b) \cdot P(a) \cdot P(b) \}, \tag{6}$$

$$\sigma_{con}^2 = \sum_{a \in A} \sum_{b \in A} \{ e_{con}^2(a,b) \cdot P(a) \cdot P(b) \} - \mu_{con}^2 \tag{7}$$

ここで、$P(a), P(b)$は各タンパク質の表面残基におけるアミノ酸 a, b の割合であり、A は２０種類のアミノ酸の集合である。$N_{contact}$ 個のコンタクトしているタンパク質ペアがシャッフリングの過程で独立に生成されたと仮定すると、コンタクトエネルギーの総和 E_{con} の平均と分散は次のように計算される。

$$Mean[E_{con}] = \mu_{con} \cdot N_{contact}, \tag{8}$$

$$Var[E_{con}] = \sigma_{con}^2 \cdot N_{contact} \tag{9}$$

ここで、$N_{contact}$ はコンタクトしているアミノ酸残基の総数である。

２.５.２　ランダム配列に対する静電エネルギーの平均と分散

静電エネルギーの平均と分散も、コンタクトエネルギーの場合と同様の考え方で計算することができる。異なる二つのタンパク質の表面からランダムに一つずつアミノ酸を取り出すことにより、タンパク質の i 番目と j 番目の位置でランダムにコンタクトするアミノ酸ペアを生成したと仮定する。この過程を繰り返してコンタクトするアミノ酸のセットを作り、その静電エネルギーの平均 $\mu_{ele}(i,j)$と分散$\sigma_{ele}^2(i,j)$は次のように計算される。

$$\mu_{ele}(i,j) = \sum_{a\in A}\sum_{b\in A} P(a)P(b) \sum_{s\in Q_i}\sum_{t\in Q_j} e_{ele}\left(r_{st}, q_s(a), q_t(b)\right), \tag{10}$$

$$\sigma_{ele}^2(i,j) = \left(\sum_{a\in A}\sum_{b\in A} P(a)P(b) \sum_{s\in Q_i}\sum_{t\in Q_j} e_{ele}^2\left(r_{st}, q_s(a), q_t(b)\right) \right) - \mu_{ele}^2(i,j) \tag{11}$$

ここで、変数 r_{st} は原子 s, t 間の距離である。変数 $q_s(a), q_t(b)$は、残基 i, j がアミノ酸 a, b に置き換わったときの原子 s, t の電荷である。$P(a), P(b)$は各タンパク質の表面残基におけるアミノ酸 a, b の頻度である。変数 s, t は原子であり、Q_i, Q_j は残基 i, j に属する荷電原子の集合である。シャッフリングの過程において、全てのタンパク質ペアが独立に生成されたと仮定すると、静電エネルギーの総和 E_{ele} の平均と分散は各アミノ酸ペアの平均の和と分散の和で計算される。

$$Mean[E_{ele}] = \sum_{i}^{N}\sum_{j}^{M} \mu_{ele}(i,j), \tag{12}$$

$$Var[E_{ele}] = \sum_{i}^{N}\sum_{j}^{M}\sigma_{ele}^{2}(i,j) \qquad (13)$$

ここで、N と M は各タンパク質のアミノ酸残基の総数である。

２.６ テンプレート構造との配列類似度

相互作用するタンパク質を見つけるためのもう一つの特徴量として、テンプレート構造との配列類似度を採用した。二つの標的配列と相同タンパク質間の複合体のアミノ酸配列がよく似ていたら、その二つの標的配列は相互作用する可能性が高い。ここでもまた、配列類似度を計算するために Z スコアを導入する。アラインメントにおける同一残基の数 N_{iden} を、ランダム配列のセットについての平均と分散を用いて次式のように正規化する。

$$Z(N_{iden}, N_{comp}) = -\frac{N_{iden} - N_{comp}p}{\sqrt{N_{comp}p(1-p)}} \qquad (14)$$

N_{comp} はギャップを除いたアラインメントにおいて比較された残基の数である。アミノ酸の均一分布（p は 1/20 とする）を用いてランダムシャッフリングを行い、同一残基の数 N_{iden} が二項分布に従うと仮定すると、ランダム配列のセットについての平均（$N_{comp}p$）と分散（$N_{comp}p(1-p)$）を求めることができる。コンタクトおよび静電エネルギーの Z スコアと符号をそろえるために配列類似度の Z スコアに-1 を乗じた。二量体の立体構造のモデリングでは、一つのタンパク質ペアにつき二つの異なる配列類似度が得られるが、二つの配列類似度のうちスコアの悪い方（つまり、配列類似度の低い方）を識別のために用いることにした。

配列類似度のランダムシャッフリングの過程は、コンタクトおよび静電エネルギーの場合と若干異なる。コンタクトおよび静電エネルギーのシャッフリングでは、異なるタンパク質表面からランダムにそれぞれ一つずつアミノ酸が取り出す。配列類似度のシャッフリングでは、テンプレートタンパク質の配列を固定し、アミノ酸の均一分布に従って、ランダムに標的タンパク質の配列を生成する。

２．７ Recall-Precision プロットによる評価

相互作用するタンパク質ペアと相互作用しないペアの間の識別力を評価するために、Recall-Precision プロットを生成した。Recall（再現率）とは、正解タンパク質ペアのうち閾値以上のスコアを示したペアの割合を表し、Precision（適合率）とは、閾値以上のスコアを示すタンパク質ペアのうち正解ペアの割合を表す。Recall と Precision は次のように定義される。

$$Recall(S) = \frac{N_{tp}(S)}{N_t}, \tag{15}$$

$$Precision(S) = \frac{N_{tp}(S)}{N_p(S)} \tag{16}$$

ここで、$N_{tp}(S)$は S よりもよいスコアを持つ相互作用が報告されているタンパク質ペアの数であり、N_tは相互作用するタンパク質ペアの数、$N_p(S)$は S よりもよいスコアを持つタンパク質ペアの数である。Recall と Precision を全てのスコア S について計算し、平面上に曲線で結んでプロットする。Recall と Precision はトレードオフの関係にあり、例えば Recall（再現率）を重視して予測すると Precision（適合率）が低下し、逆に Precision（適合率）を重視して予測すると Recall（再現率）が低下するため、平面上に右下がりの曲線がプロットされることになる。より右上にプロットされた曲線は、左下のものよりも識別力がよいことを表している。

Recall と Precision の間のバランスをうまくとるために F-measure の最大値を用いた。F-measure の $F(S)$は Recall と Precision の調和平均として定義され、F-measure の最大値 $F_{\max}$ は、全てのスコアの中で F-measure が最大となるときのその値である。

$$F(S) = 2\left(\frac{1}{Recall(S)} + \frac{1}{Precision(S)}\right)^{-1} = \frac{2N_{tp}(S)}{N_t + N_p(S)}, \tag{17}$$

$$F_{\max} = \max_s \left[F(S)\right] \tag{18}$$

第３章　結果

３.１　複合体立体構造モデルの精度の検証

本研究は、複合体立体構造モデルのエネルギーを計算することで、タンパク質間相互作用が起こり得るかどうかを予測するものである。したがって、その予測精度は、(i) 複合体立体構造モデルの精度、 (ii) エネルギー計算の精度 の二つの精度に左右されることになる。本節では、(i)について検証した結果を報告する。一般にホモロジーモデリングの予測精度はテンプレート構造との配列類似度に依存することが知られている。ヘテロダイマーとホモダイマーについて配列類似度と相互作用残基の予測精度の関係を調査することで、複合体立体構造モデルの精度について検証した。

2.1 節の「複合体立体構造のデータセット」で作成した 1,687 個の代表ヘテロダイマーを標的構造とし、その立体構造とモデル構造の間で相互作用残基がどれだけ重なっているかを評価するため、両者の相関係数を計算した。相互作用残基は、異なる鎖に Cβ原子が 7 Å 以内に位置する表面残基が存在しているような表面残基として定義し、表面残基は、溶媒露出度が 35%以上となるアミノ酸残基として定義した。相関係数は Matthews がタンパク質の二次構造予測で用いた点相関係数を採用した(Matthews, 1975)。すなわち、標的構造とモデル構造の両方に含まれる相互作用残基の数を a、モデル構造のみに含まれる相互作用残基の数を b、標的構造のみに含まれる相互作用残基の数を c、残りの表面残基の数を d とすると、相関係数 *corr.coef.*は以下の式で計算される。

$$corr.coef. = \frac{ad - bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}} \qquad (-1 \le corr.coef. \le 1) \qquad (19)$$

相関係数が 1 に近づくほど相互作用残基を相互作用残基と予測する傾向が高く、0 では完全に無相関なランダムな予測であることを示し、-1 に近づくほど非相互作用残基を相互作用残基と予測する傾向が高いことを示す。全体集合を標的構造上の表面残基とした理由は、全残基とすると相互作用残基の一致度を過大評価してしまうからである。アミノ酸残基が相互作用するためには表面に存在していることが前提であると考えられるので、評価対象を表面残基に限定することでより妥当性の高い評価を行った。標的構造とモデル構造間の相互作用残基の相関係数を縦軸、両者の同一残基率を横軸として得られた散布図を示す（図３－１）。

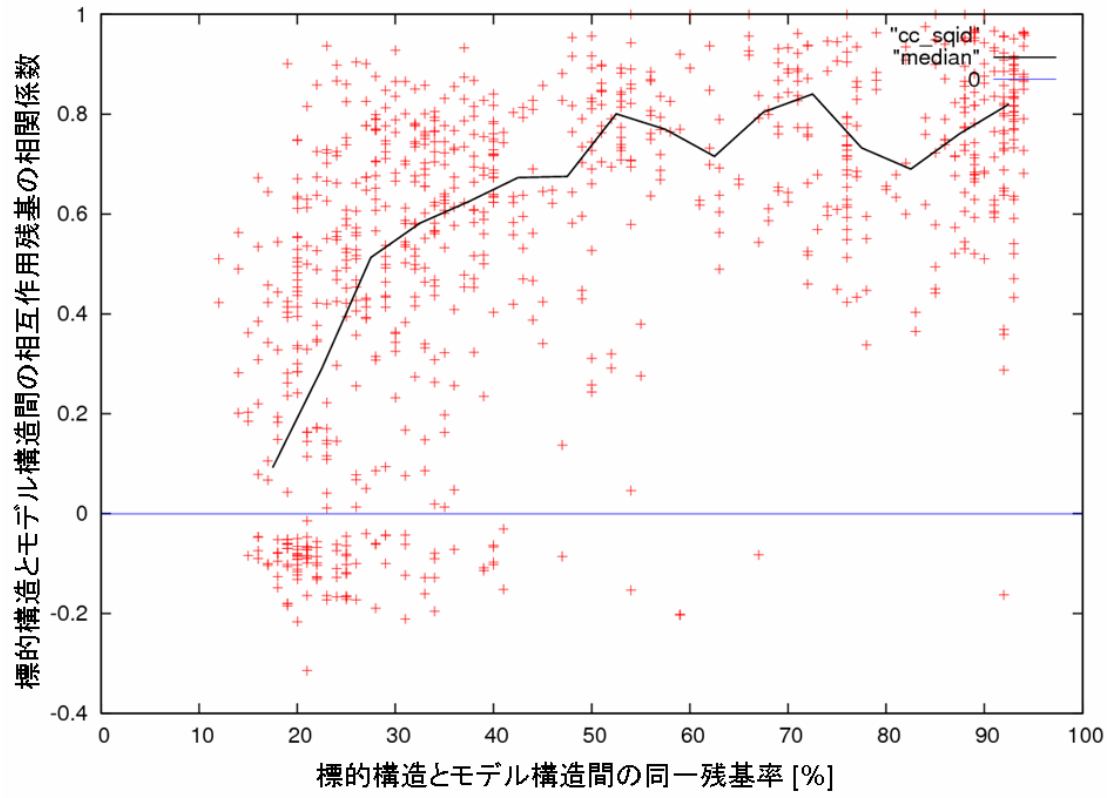


図３－１　ヘテロダイマーの立体構造モデルの精度

1,687 個の代表ヘテロダイマーのうち、428 個についてモデル構造の相互作用残基が計算でき、これらのヘテロダイマーを構成しているタンパク質（856 個）

ごとに相関係数と同一残基率を計算してプロットした。同一残基率は複合体一つあたり二つ存在するが、小さい方を用いてプロットした。また、メジアンの位置を黒の実線で表示した。856 個中 643 個（75%）が相関係数 0.4 以上、486 個（57%）が相関係数 0.6 以上となり、相関係数が負の値をとるものもあった（112 個、13%）。同一残基率が 40%を下回ったあたりから次第に相関係数が低下する傾向が見られる。このことは、同一残基率が 30-40%を境に相互作用面がずれるという Aloy らの報告（Aloy *et al.*, 2003）と見解が一致する。

相関係数が低い値をとった立体構造を詳細に観察すると、配列上ヘテロダイマーであってもその構成タンパク質どうしが相同な関係にあり、事実上ホモダイマー的な複合体が多く含まれることが分かった。例えば、プロテアソームのような巨大複合体には、同一残基率 50%以下の相同な関係にある構成タンパク質ペアが複数含まれる。これらはプロテアソーム中での相対的な位置関係が様々に異なるため、標的構造およびテンプレート構造として用いると相互作用残基が全く異なったモデル構造を生成してしまうことになる。図３－２において、標的タンパク質とテンプレートタンパク質の対応はそれぞれ青と緑のタンパク質で表されており、相対的な位置関係が全く異なっている。この場合、青のタンパク質ではなく、赤（1rypJ）のタンパク質をテンプレートとすればより適切なモデル構造が得られていたことになる。

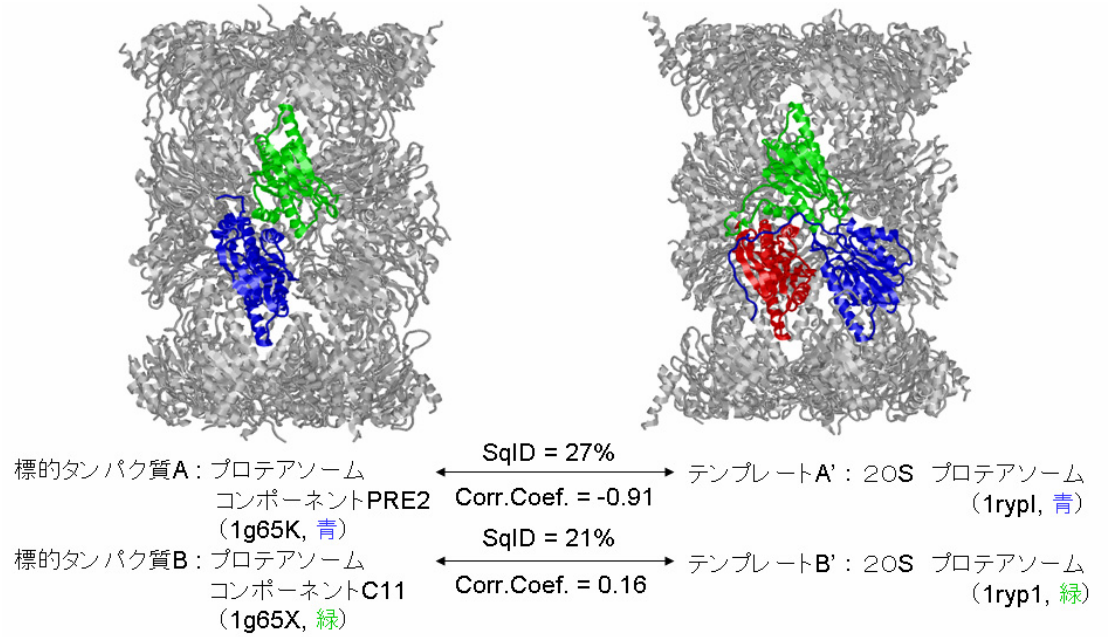


図３－２　ヘテロダイマーの場合の不適切なテンプレート構造の例

ホモダイマーについても同様の調査を行うため、2.1 節の「複合体立体構造のデータセット」と同様の手順で 4,130 個のホモダイマーを作成した。このうち 2,182 個についてモデル構造の相互作用残基が計算でき、これらのホモダイマーを構成しているタンパク質（4,364 個）ごとに相関係数と同一残基率を計算してプロットした（図３－３）。

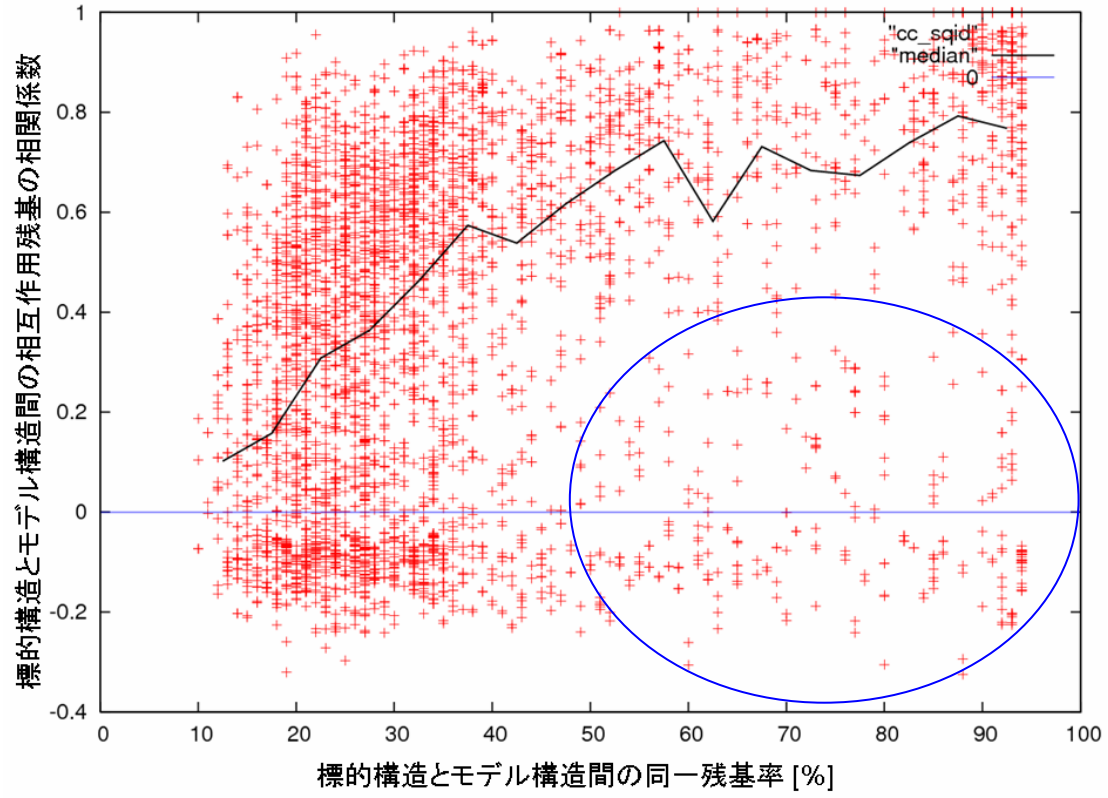


図３－３　ホモダイマーの立体構造モデルの精度

同一残基率が 40%を下回ったあたりから、次第に相関係数が低下する傾向が見られる点でヘテロダイマーと共通し、4,364 個中 2,270 個（52%）が相関係数 0.4 以上、1,471 個（34%）が相関係数 0.6 以上、相関係数が負の値をとるものが 1,011 個（23%）存在した。ヘテロダイマーの場合と大きく異なる点は、ホモダイマーの場合、同一残基率が 50%以上の高い領域においても相関係数が 0.4 以下のプロットが多く存在していることである（図３－３の青枠内の領域）。標的配列が天然のホモダイマーで、テンプレート構造がクリスタルパッキングである例を示す（図３－４）。

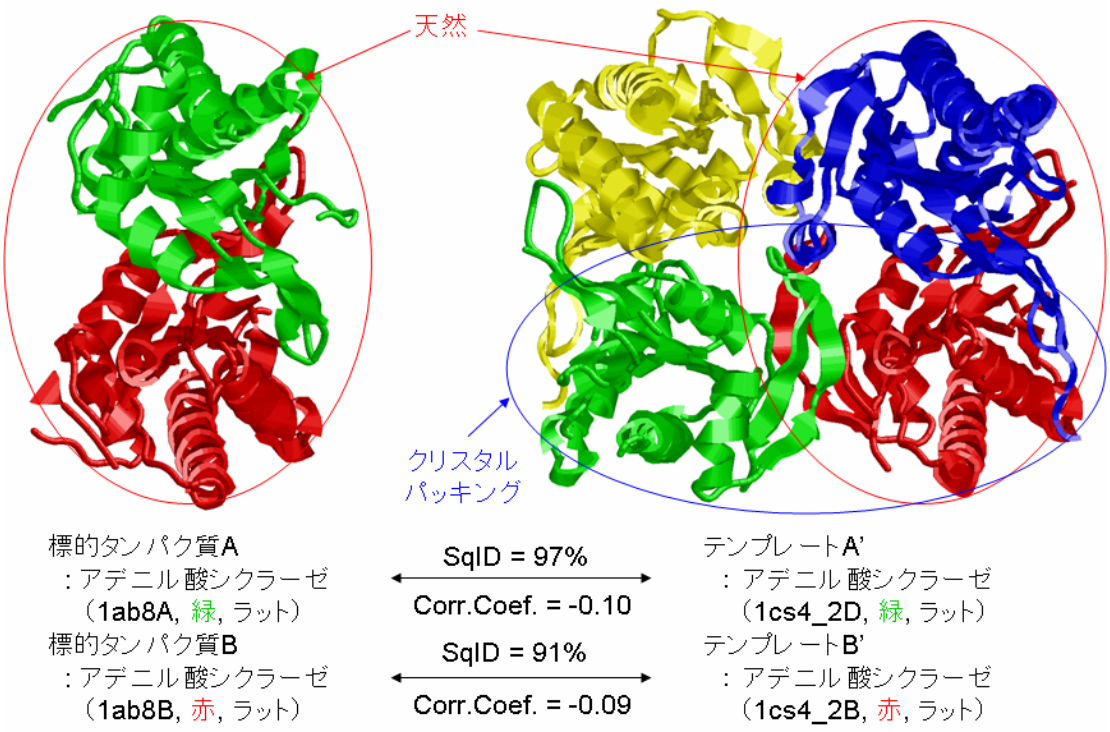


図３－４　ホモダイマーの場合の不適切なテンプレート構造の例

図３－４において標的配列 AB に対し、テンプレート構造を緑(1cs4_2D)と赤(1cs4_2B)ではなく、青(1cs4_2A)と赤(1cs4_2B)にしていれば、複合体立体構造モデルは高い精度で作成されていたことになる。

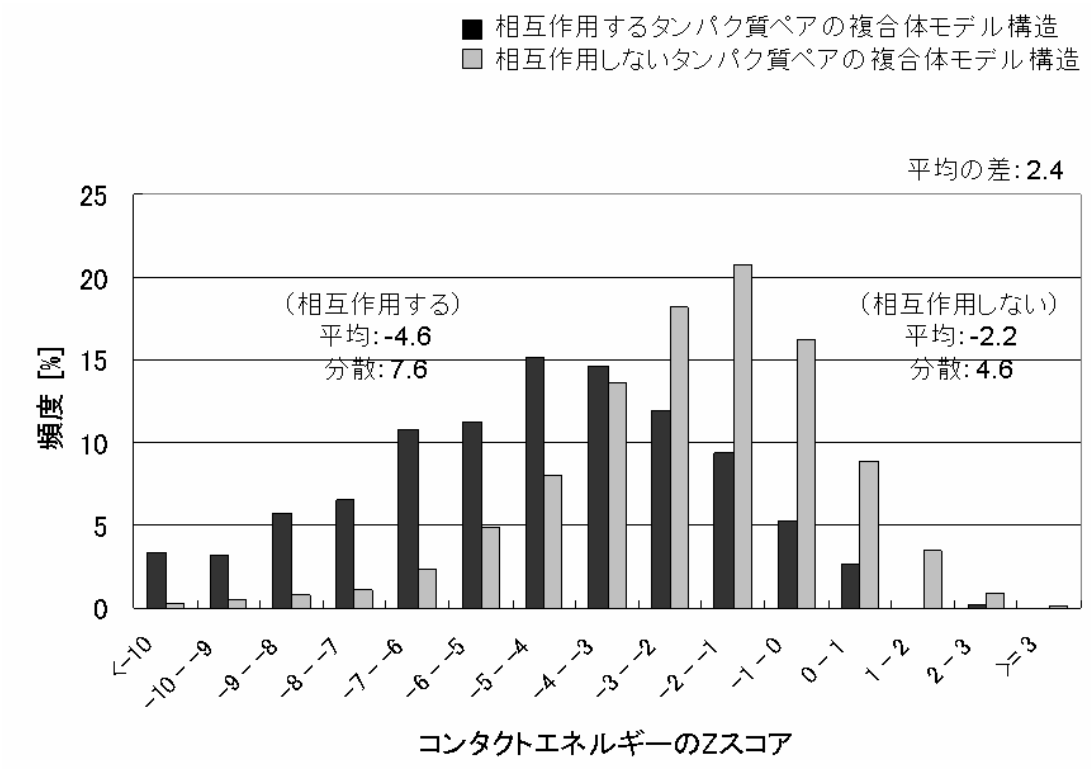
以上の点を踏まえ、ホモのタンパク質ペアの場合、同一残基率が高い領域においてもモデル構造の信頼性が低くなるものが含まれると予想されるため、本研究ではホモのタンパク質ペアを除くことにし、ヘテロのタンパク質ペアのみを予測対象とすることにした。ヘテロのタンパク質ペアであっても、同一残基率が 40%を下回ると信頼性が低くなるものが含まれると予想されるが、このようなモデル構造を除外してしまうと予測可能なタンパク質ペアが大幅に限定されてしまうため、あえて同一残基率に制限を設けないこことした。

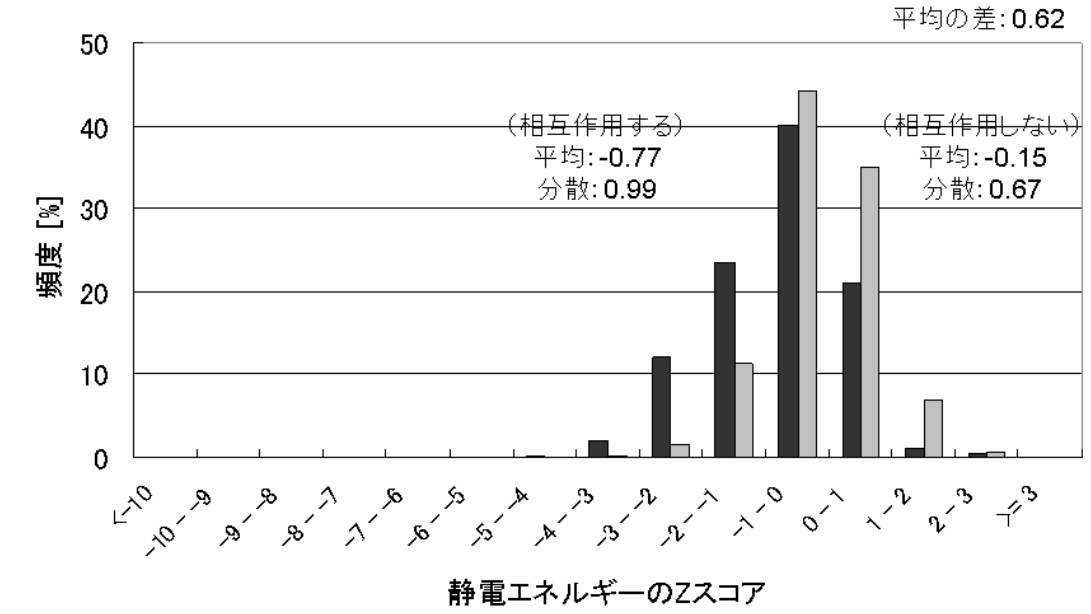
３.２　複合体モデル構造の作成

ホモロジーモデリングを用いて、出芽酵母のヘテロのタンパク質ペアの複合体立体構造をモデリングした。まず酵母の 5,314 個のタンパク質配列について総当りのヘテロのペアを考えると、(5,314×5,313) / 2 = 14,116,641 ペア作ることが可能であるが、そのうち、複合体モデル構造を作成できたペアは 10,325 ペアであった。これら 10,325 個のペアを、DIP データベースに含まれていれば「相互作用するタンパク質ペア」、DIP データベースに含まれていなければ「相互作用しないタンパク質ペア」とラベル付けすると、相互作用するタンパク質ペアは 417 個、相互作用しないタンパク質ペアは 9,908 個となった。

３.３　各特徴量のスコア分布

本研究で採用した３つの特徴量が「相互作用するペア」と「相互作用しないペア」を識別する能力があるか調べるために、３つの特徴量（コンタクトエネルギー、静電エネルギー、テンプレート構造との配列類似度）の Z スコア分布を図３―５に示す。


■ 相互作用するタンパク質ペアの複合体モデル構造
□ 相互作用しないタンパク質ペアの複合体モデル構造
平均の差: 2.4
25
20
15
10
5
0
頻度 [%]
（相互作用する）
平均: -4.6
分散: 7.6
（相互作用しない）
平均: -2.2
分散: 4.6
<-10
-10 - -9
-9 - -8
-8 - -7
-7 - -6
-6 - -5
-5 - -4
-4 - -3
-3 - -2
-2 - -1
-1 - 0
0 - 1
1 - 2
2 - 3
>=3
コンタクトエネルギーのZスコア


平均の差: 0.62
50
40
30
20
10
0
頻度 [%]
（相互作用する）
平均: -0.77
分散: 0.99
（相互作用しない）
平均: -0.15
分散: 0.67
<-10
-10 - -9
-9 - -8
-8 - -7
-7 - -6
-6 - -5
-5 - -4
-4 - -3
-3 - -2
-2 - -1
-1 - 0
0 - 1
1 - 2
2 - 3
>=3
静電エネルギーのZスコア

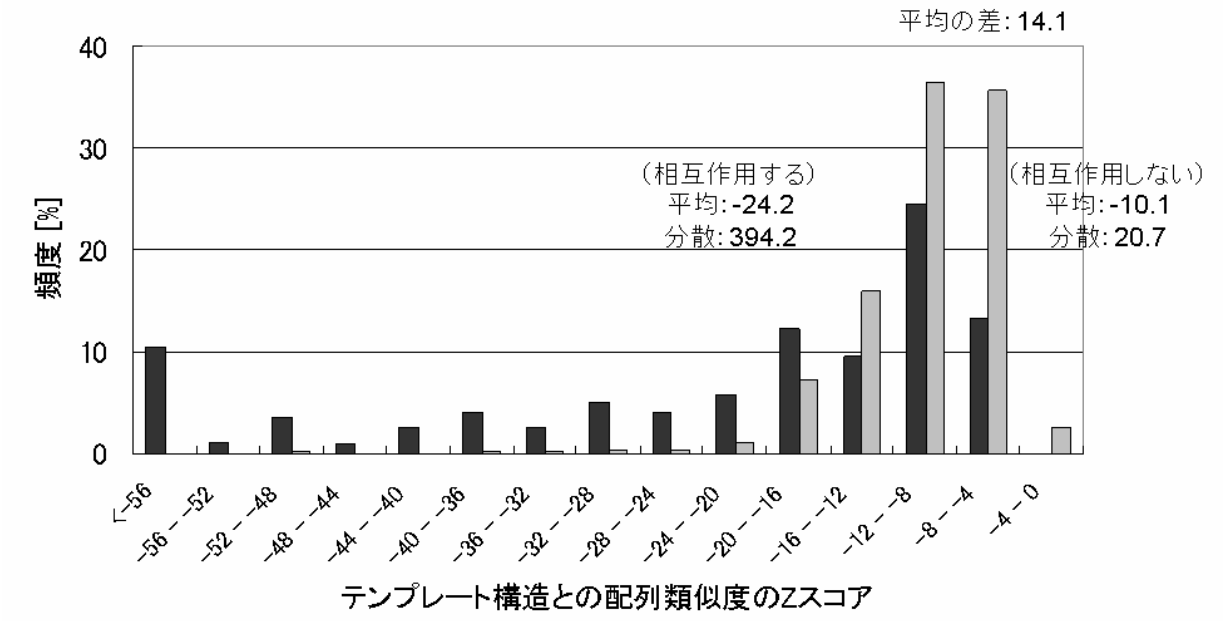


図３－５　３つの特徴量のＺスコア分布

コンタクトエネルギーと静電エネルギーのＺスコアは、同一のシャッフリング過程を用いているので互いに比較可能である。コンタクトエネルギーのＺスコアは、静電エネルギーのＺスコアよりも低い値で広い範囲に分布している。相互作用するタンパク質ペアの複合体モデル構造（「相互作用する」で表示）と相互作用しないタンパク質ペアの複合体モデル構造（「相互作用しない」で表示）について、コンタクトエネルギーのＺスコアの平均値はそれぞれ-4.6 と-2.2 であるが、静電エネルギーのＺスコアの平均値はそれぞれ-0.77 と-0.15 である。コンタクトエネルギーの分散は 7.6（相互作用する）と 4.6（相互作用しない）で、静電エネルギーの分散は 0.99（相互作用する）と 0.67（相互作用しない）である。相互作用するタンパク質ペアと相互作用しないペアの間の平均値の差は 2.4（コンタクトエネルギー）と 0.62（静電エネルギー）なので、コンタクトエネルギーは静電エネルギーよりも識別力が高いと考えられる。

テンプレート構造との配列類似度の分布は、コンタクトエネルギーや静電エネルギーのようなベル型ではなく、左側に裾が非対称に伸びた形状をしている。相互作用するタンパク質ペアは、相互作用しないペアよりも広い範囲に分

布しており、分布の分散は 394.2（相互作用する）と 20.7（相互作用しない）である。相互作用するタンパク質ペアと相互作用しないペアの間の平均値の差は 14.1 なので、非常に高い識別力を有していると考えられる。

３．４　Recall-Precision プロット

識別力をより厳密に評価するために、全ての三つの Z スコアとその結合スコアに対して Recall-Precision プロットを作成した。結合スコアの生成においては、二つあるいは三つの Z スコアを重みづけせずにそのまま足し合わせた。フィッシャーの線形判別法などの様々な重みづけも試みたが、性能は有意に改善しなかった。Recall-Precision プロットを図３－６に示し、Recall-Precision プロットの F-measure の最大値は図３－７にまとめた。

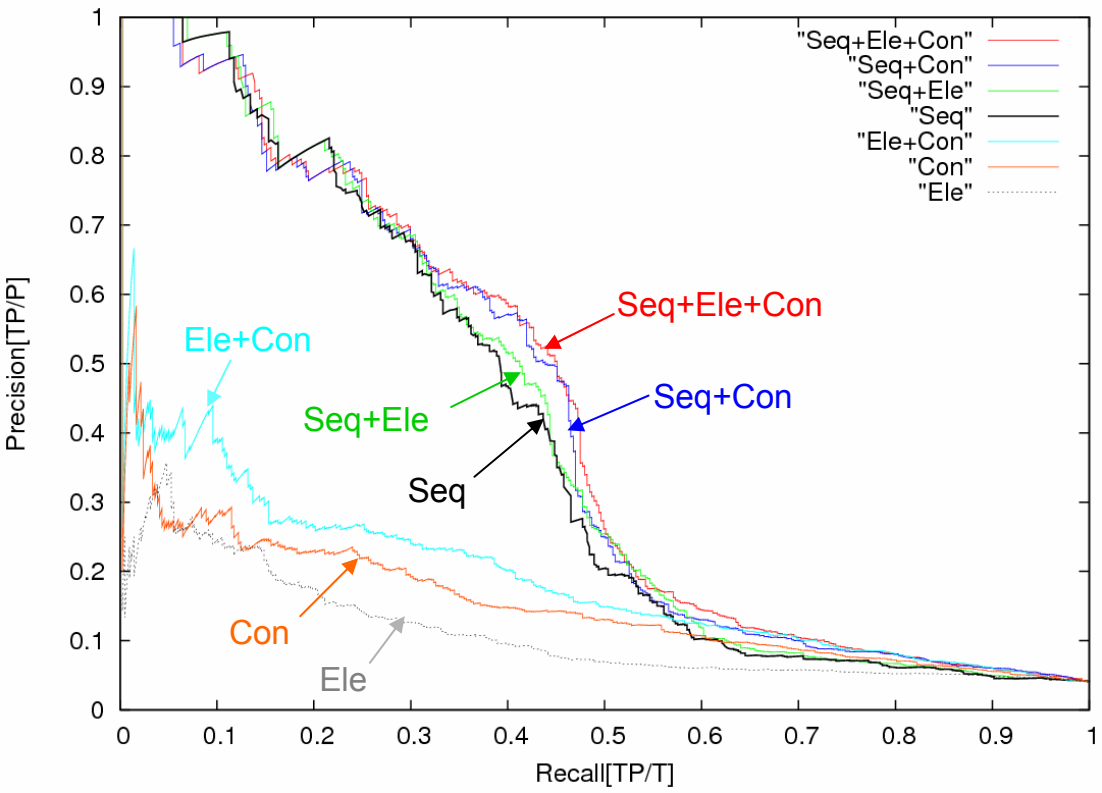


図３－６　単独および結合 Z スコアを用いた、報告ありと報告なしのタンパク質ペアの識別の Recall-Precision プロット

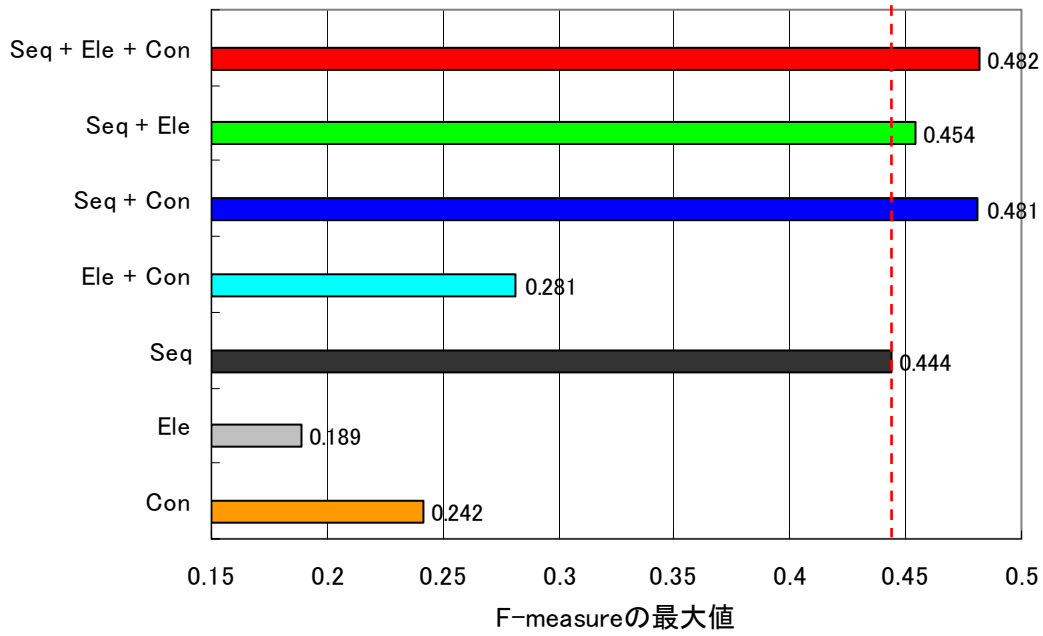


図３－７　単独および結合 Z スコアを用いた、各 Recall-Precision プロットにおける F-measure の最大値

図３－６、図３－７において、スコアの省略形は次のとおりである。Con：コンタクトエネルギー、Ele：静電エネルギー、Seq：テンプレート構造との配列類似度。Ele＋Con、Seq＋Con、Seq＋Ele、Seq＋Ele＋Con は結合 Z スコア。図３－７において、破線は配列類似度の場合の F-measure の最大値を表す。

まず、コンタクトエネルギーは静電エネルギーよりも識別力が高いことがわかる。しかし、配列類似度の識別力は、コンタクトおよび静電エネルギーよりも大幅に高い。この配列類似度の性能の高さは、配列類似度を用いた他の研究(Matthews *et al.*, 2001; Wojcik and Schachter, 2001; Wojcik *et al.*, 2002; McDermott and Samudrala, 2004)と見解が一致している。さらに、コンタクトエネルギーと静電エネルギーを配列類似度に結合すると、F-measure の最大値は 0.010 から 0.038 までの範囲で改善した。したがって、相互作用するタンパク質ペアの複合体モデル構造と相互作用しないペアのモデルを識別するためには配列情報

が最も効果的な特徴量であるが、構造情報を加えることで予測性能を有意に改善できるということが示された。

これらの改善の統計的有意性を確認するため、ブートストラップサンプリングテストを行った。ブートストラップサンプリングとは、N 個のデータの中から重複を許して N 個のデータを取り出す試行を M 回繰り返し、M 個のセットを作成する統計的手法であり、M 個のセットのうち目的の結果が得られた回数を調べることによって結果の安定性を評価することができる。10,325 個の複合体モデル構造の中からブートストラップサンプリングされた複合体モデル構造を用いて、F-measure の最大値を再計算した。サンプリングは 1,000 回繰り返し、1,000 個の異なる F-measure の最大値を得た。そのうち、配列類似度とコンタクトエネルギーの結合スコア(Seq + Con)、および全特徴量の結合スコア(Seq + Con + Ele)を用いた場合に、1,000 個全ての F-measure の最大値が配列類似度(Seq)の場合を超えた。しかしながら、配列類似度と静電エネルギーの結合スコア(Seq + Ele)の場合、984 個の F-measure の最大値だけが配列類似度(Seq)の場合を超えた。したがって、配列類似度に対し、コンタクトエネルギー結合後の識別力の改善は統計有意であるといえるが($p < 0.01$)、静電エネルギー結合後の改善は有意とは言えないことが分かった。

第４章　議論

４.１　検討事項

第３章で報告した研究結果について、以下の５つの点について検討を行うことにした。

１．本研究と先行研究の性能比較（4.2 節）
２．相互作用するタンパク質ペアと相互作用しないペアのファミリーの偏りの調査（4.3 節）
３．配列類似度にコンタクトエネルギーを加えて改善されたタンパク質ペアのファミリーの調査（4.4 節）
４．　DIP データベースに含まれていない相互作用を検出（4.5 節）
５．より信頼性の高い評価基準への適用（4.6 節）

検討事項１に関しては、タンパク質間複合体の立体構造に基づく予測手法(Davis et al., 2006)と本手法との性能比較を行うことで、本手法の性能をより客観的に評価するためである。

検討事項２に関しては、複合体モデル構造が構築できるタンパク質ペアには特定のファミリーが大量に含まれている可能性があるからである。

検討事項３に関しては、コンタクトエネルギーを加えて改善されたタンパク質ペアのファミリーを具体的に調査し、その原因を探るとともに、改善に寄与したタンパク質ペアにもファミリーの偏りがあるかどうかを確認する必要性があると考えたからである。

検討事項４に関しては、本研究では、DIP データベースには全タンパク質間相互作用が登録されていると仮定したが、1.3 節で述べたようにハイスループットの実験データの信頼性が低いことを鑑みると、DIP に含まれていないが実際には相互作用するタンパク質ペア（偽陰性）が存在していると考えられるからである。

検討事項５に関しては、本研究の評価データ（特に相互作用しないタンパク質ペア）の信頼性に問題があることに鑑み、細胞内局在情報を用いてより信頼性の高い評価データを作成し、再計算を行った。

４.２　本研究と先行研究の性能比較

2006 年に Davis らによって報告されたタンパク質間複合体の立体構造に基づく予測手法（Davis *et al.*, 2006）と本手法との性能比較を行うことにした。Davis らも特徴量としてコンタクトエネルギーを採用したが、相互作用する原子数の比による連続的なコンタクトの度合いの統計をベースとしているため、定式化がより複雑である。また彼らは、接触する残基ペアを計算する原子の種類を主鎖－主鎖、主鎖－側鎖、側鎖－側鎖、全原子－全原子の４通りに分類し、その接触もドメイン間接触とドメイン内接触の２通りに分類、コンタクトしているかどうかの閾値も 4, 6, 8 Å の３通りに分類し、合計２４通りのコンタクトエネルギーを作成した。これらのエネルギーに対してベンチマークテストを行った結果、側鎖－側鎖の原子間、ドメイン間接触、コンタクトの閾値 8 Å で作成したコンタクトエネルギーが最も性能が高いことを確認し、その条件で酵母の全タンパク質ペア（ヘテロだけではなく、ホモのタンパク質ペアも含む）の中から相互作用するタンパク質ペアの予測を行った。

彼らの予測法の概略は図１－７で説明したとおりであるが、単量体構造モデルとして MODBASE（Pieper *et al.*, 2006）に登録された全原子モデルを採用し、単量体モデルと PIBASE に登録されたテンプレート複合体構造を構造アラインメントすることで複合体構造モデルを得ている。その複合体立体構造モデルに対して上記のコンタクトエネルギーの計算を行い、得られたエネルギー値をランダム配列を基準とした Z スコアに変換し、有意性を評価している。Z スコアが-1.7 以下となるタンパク質ペアを予測すると、1,390 個の酵母のタンパク質

で構成される 12,867 ペアが予測されている。このタンパク質ペアの中から、さらに信頼性のある相互作用を予測するために、ペアを構成する二つのタンパク質について、機能（ジーンオントロジーのアノテーション）が一致しているもの、および細胞内局在が一致しているものに限定すると、最終的に 3,387 個の相互作用するタンパク質ペアが予測されている。

Davis らも酵母の全タンパク質ペアに自らの手法を適用しているため、評価基準を本手法に合わせることで本手法との性能評価を行うことは可能であると考えられる。すなわち、Davis らによって予測されたタンパク質ペア（3,387 個）と、本手法の評価に用いた相互作用するタンパク質ペアおよび相互作用しないタンパク質ペアとの重複で評価することにした。まず、Davis らによって予測された 3,387 個のタンパク質ペアからヘテロ（同一残基率 50%以下）のタンパク質ペアを抽出すると、2,520 個のヘテロのペアが得られた。これらのタンパク質ペアと、本研究で作成した相互作用するタンパク質ペア（417 個）と相互作用しないタンパク質ペア（9,908 個）との重複を調べると、前者については 417 個中 84 個が重複し、後者については 9,908 個中 216 個が重複していることが分かった。したがって、Davis らの Recall と Precision の値は式(15)と式(16)から、

$$Recall(S) = \frac{N_{tp}(S)}{N_t} = \frac{84}{417} = 0.201,$$

$$Precision(S) = \frac{N_{tp}(S)}{N_p(S)} = \frac{84}{84+216} = 0.280$$

となり、これらの値は図４－１の紫の三角で示される位置にプロットされる。

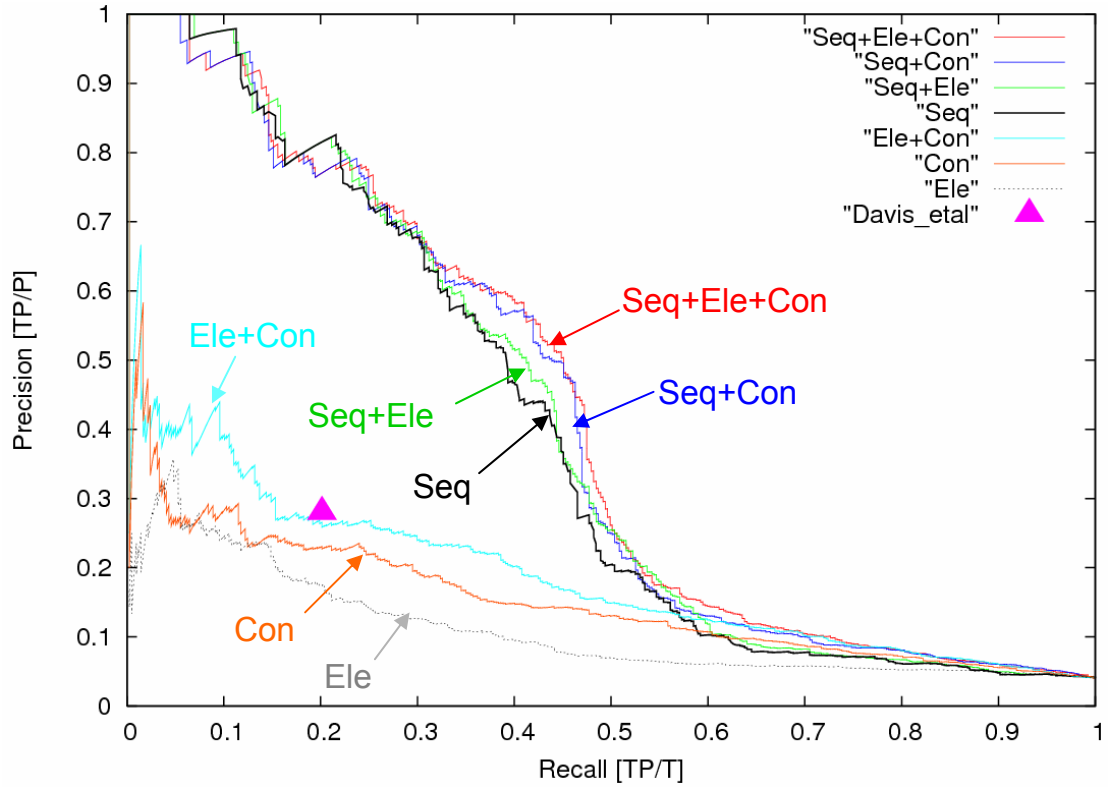


図４－１　Davis らの手法と本手法との性能比較

図４－１において紫の三角のプロットから、Davis らの手法は本手法のコンタクトエネルギーよりも優れており、本手法のコンタクトエネルギーと静電エネルギーの結合スコアよりも若干優れているといえる。これは、Davis らの手法は、本手法のコンタクトエネルギーよりも精密であること、構成タンパク質の機能や細胞内局在が一致しているものだけを予測していることなどが原因として考えられる。しかしながら、彼らの手法も本手法の配列類似度には及ばない。したがって、配列類似度は様々な立体構造に基づく特徴量よりも高い予測性能を持ち、タンパク質間相互作用予測には配列情報をベースとした上で立体構造情報を加味した特徴量が極めて有効であるといえる。

４.３　相互作用するペアと相互作用しないペアのファミリーの偏り

複合体モデル構造を構築できるタンパク質ペアの全体像をつかむために、これらのペアのネットワークを描いた（図４－２）。

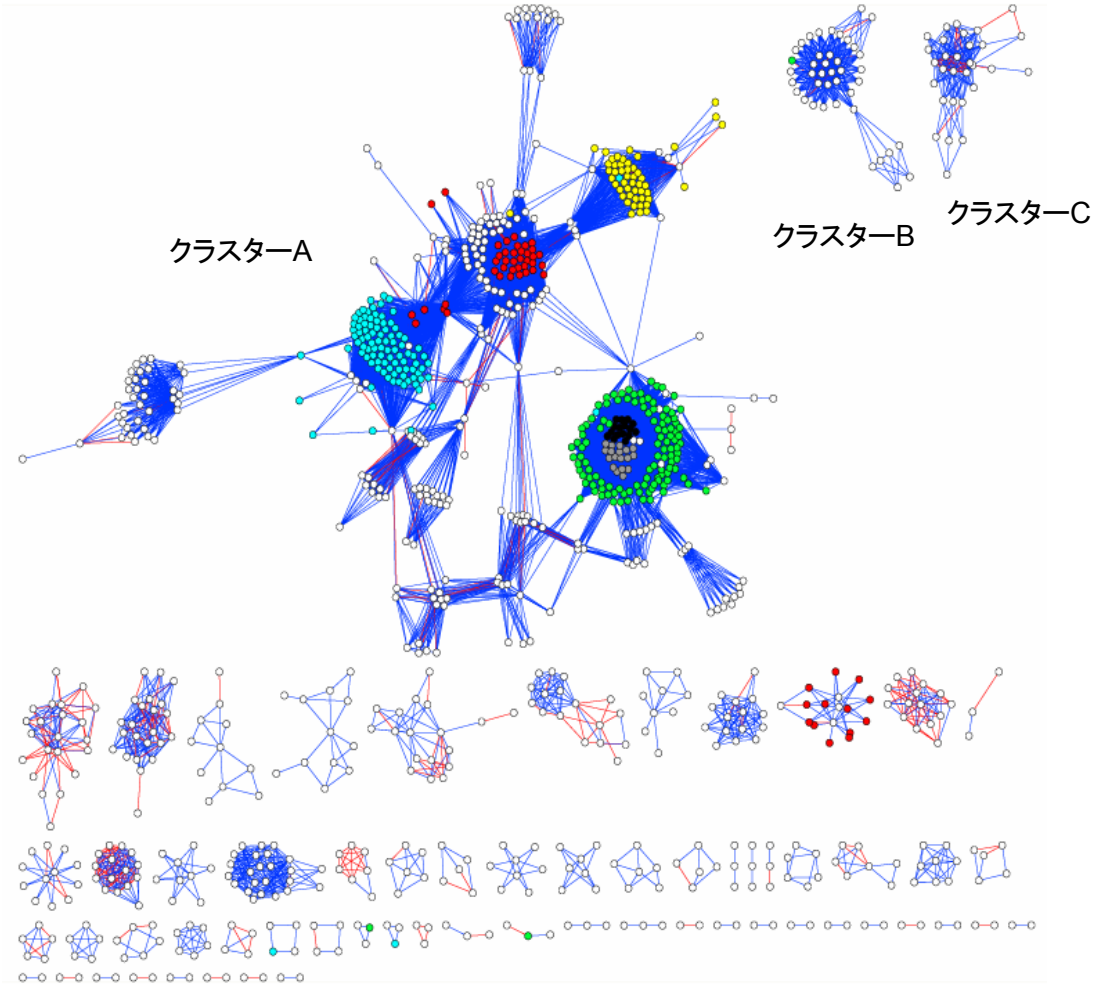


図４－２　複合体モデル構造を構築できるタンパク質ペアのネットワーク

このグラフは Cytoscape (Shannon *et al.*, 2003)を用いて可視化した。ノードは標的タンパク質に対応し、エッジは複合体立体構造がモデリング可能な標的タン

パク質のペアに対応する。ノードの数は 1,036 個、エッジの数は 10,325 個であった。相互作用するタンパク質ペアは赤で示し、相互作用しないペアは青で示す。相互作用しないタンパク質ペアの数は相互作用するペアの数の約２４倍であるので、大部分のエッジは青で色づけされている。このネットワークはシングル・リンケージ・クラスタリングでは６４個のクラスターから成るとされ、実験で決定されたタンパク質間相互作用ネットワーク(Gavin *et al.*, 2002; Krogan *et al.*, 2006)よりも分割の程度が大きい。この理由は、研究対象をホモロジーモデリング可能なタンパク質ペアに限定したからであると考えられる。最も大きなクラスター（クラスターA）には 573 個のタンパク質が含まれ、二番目(Cluster B)と三番目(Cluster C)に大きなクラスターはそれぞれ 41 個、30 個のタンパク質を含む。クラスターA に含まれている標的タンパク質に着目し、クラスターA に含まれる主要なドメインによってネットワークのノードを色づけした。クラスターA はシグナル伝達系として働くタンパク質を含んでいる。タンパクキナーゼの触媒サブユニット（緑)、WD40 リピート（シアン)、G タンパク質（赤)、カノニカル RBD（黄)、アンキリンリピート（グレー)、サイクリン（黒）のドメインを含む標的タンパク質の数は、それぞれ 119、97、55、50、18、16 個であった。標的タンパク質がこれらの六種類のドメインのうち二種類以上のドメインを含んでいる場合は、N 末に最も近いドメインよってノードを色づけした。ドメインの分類には SCOP データベース(Andreeva *et al.*, 2004)を用いた。クラスターB はユビキチン化のタンパク質を含み、主に RING フィンガードメイン C3HC4（17 ドメイン)、およびユビキチン結合酵素 UBC（14 ドメイン）の二つのファミリーからなる。クラスターC は DNA 複製系として働くタンパク質を含み、拡張 AAA-ATP アーゼ（23 ドメイン)、および DNA ポリメラーゼ III クランプローダーサブユニットの C 末ドメイン（７ドメイン）が存在する。

図４－２のネットワークにおいて色づけした六種類のファミリーの存在からも示唆されるように、複合体モデル構造が構築できるタンパク質ペアを構成しているタンパク質には大きなファミリーの偏りがあるといえる。構成タンパク質のみならず、タンパク質ペア単位でもファミリーの偏りが生じているかどうかを調べるため、相互作用するタンパク質ペアと相互作用しないペアについ

てテンプレート構造によく現れるファミリーのペアの統計を取った（表４－１）。

表４－１　テンプレート構造によく現れるファミリーのペア

		タンパク質ペアの数	
テンプレート構造のファミリーのペア	テンプレート	相互作用する	相互作用しない
相互作用するタンパク質ペアによく現れるファミリー			
1. b.38.1.1 / b.38.1.1	1b34A / 1b34B	33 (15 / 15)	24 (15 / 15)
2. d.153.1.4 / d.153.1.4	1g65J / 1g65K	30 (14 / 14)	44 (14 / 14)
3. h.1.15.1 / h.1.15.1	1gl2B / 1gl2C	20 (15 / 15)	80 (19 / 19)
4. c.37.1.20 - a.80.1.1 / c. 37.1.20 - a.80.1.1	1sxjC / 1sxjB	19 (8 / 8)	95 (20 / 20)
5. a.74.1.1 - a.74.1.1 / d.144.1.7	1finB / 1finA	18 (13 / 7)	1662 (15 / 114)
6. c.3.1.3 - d.16.1.6 - c.3.1.3 / c.37.1.8	1ukvG / 1ukvY	13 (2 / 10)	61 (2 / 34)
7. d.144.1.7 / d.211.1.1	1bi7A / 1bi7B	10 (8 / 5)	1912 (108 / 18)
8. a.22.1.1 / a.22.1.1	1id3A / 1id3F	9 (6 / 6)	12 (9 / 9)
9. i.1.1.1 / i.1.1.1	1s1hJ / 1s1hN	8 (13 / 13)	16 (21 / 21)
10. a.116.1.1 / c.37.1.8	1ow3A / 1ow3B	6 (4 / 4)	342 (10 / 35)
相互作用しないタンパク質ペアによく現れるファミリー			
1. d.144.1.7 / d.211.1.1	1g3nA / 1g3nB	10 (8 / 5)	1912 (108 / 18)
2. a.74.1.1 - a.74.1.1 / d.144.1.7	1oiuB / 1oiuC	18 (13 / 7)	1662 (15 / 114)
3. b.69.4.1 / c.37.1.8 - a.66.1.1 - c.37.1.8	1gotB / 1gotA	1 (1 / 1)	530 (89 / 6)
4. a.116.1.1 / c.37.1.8	1ow3A / 1ow3B	6 (4 / 4)	342 (10 / 35)
5. d.144.1.7 / j.66.1.1	1f3mC / 1f3mA	1 (1 / 1)	319 (108 / 3)
6. c.10.2.4 / d.58.7.1	1a9nA / 1a9nB	2 (1 / 2)	257 (6 / 44)
7. c.10.1.2 / c.37.1.8	1k5dC / 1k5dA	4 (2 / 4)	239 (7 / 35)
8. c.45.1.1 / d.144.1.7	1fq1A / 1fq1B	0 (0 / 0)	204 (2 / 105)
9. a.48.1.1 - a.39.1.7 - d.93.1.1 - g.44.1.1 / d.20.1.1	1fbvA / 1fbvC	3 (2 / 3)	189 (16 / 12)
10. a.118.1.1 / c.37.1.8	1qbkB / 1qbkC	4 (3 / 2)	184 (8 / 34)

表中の SCOP ID は次のとおりである。a.22.1.1：ヌクレオソームコアヒストン, a.39.1.7：マルチドメインタンパクの EF ハンドモジュール, a.48.1.1：cb1 の N 末ドメイン, a.66.1.1：トランスデューシンαサブニット挿入ドメイン, a.74.1.1：

サイクリン, a.80.1.1：DNA ポリメラーゼ III クランプローダーサブユニット C 末ドメイン, a.116.1.1 : BCR 相同 GTP アーゼ活性化ドメイン, a.118.1.1： アルマジロリピート, b.38.1.1： 核内低分子リボ核タンパク質の Sm モチーフ, b.69.4.1 : WD40 リピート, c.3.1.3：GDI 様 N 末ドメイン, c.10.1.2：Rna1p N 末ドメイン, c.10.2.4 : U2A'様, c.37.1.8 : G タンパク質, c.37.1.20： 拡張 AAA-ATP アーゼドメイン, c.45.1.1： 二重特異性フォスファターゼ様, d.16.1.6：GDI 様, d.20.1.1： ユビキチン結合酵素 UBC, d.58.7.1 ： カノニカル RBD, d.93.1.1 ： SH2 ドメイン, d.144.1.7 ： タンパクキナーゼ触媒サブユニット, d.153.1.4 ： プロテアソームサブユニット, d.211.1.1： アンキリンリピート, g.44.1.1 : RING フィンガードメイン C3HC4, h.1.15.1 ： SNARE 融合複合体, i.1.1.1 ： リボソーム複合体, j.66.1.1 ： pak1 自己調節ドメイン

表中、テンプレート複合体の PDB ID を併記した。図４－２で色づけした６種類のファミリーについては、表中の対応するファミリーに同様の色づけを行った。括弧内の数字はペアを形成するタンパク質の数を表す。例えば、テンプレートの SCOP ID が b.38.1.1 である 15 個のタンパク質が、33 個の相互作用するタンパク質ペアを形成し、またテンプレートの SCOP ID が b.38.1.1 である 15 個のタンパク質が、24 個の相互作用しないタンパク質ペアを形成している。

相互作用しないタンパク質ペア（「相互作用する」で表示）は、相互作用するペア（「相互作用しない」で表示）よりもファミリーの偏りが大きく、その偏りの主な原因は図４－２のネットワークで色づけした６種類のファミリーであることが分かった。例えば、15 個のサイクリンドメインと 114 個のタンパクキナーゼ触媒サブユニットドメインは1,662個もの相互作用しないタンパク質ペアを形成している。

参考のため、図４－２のネットワークで色づけした６種類のファミリーの立体構造を図４－３に示す。

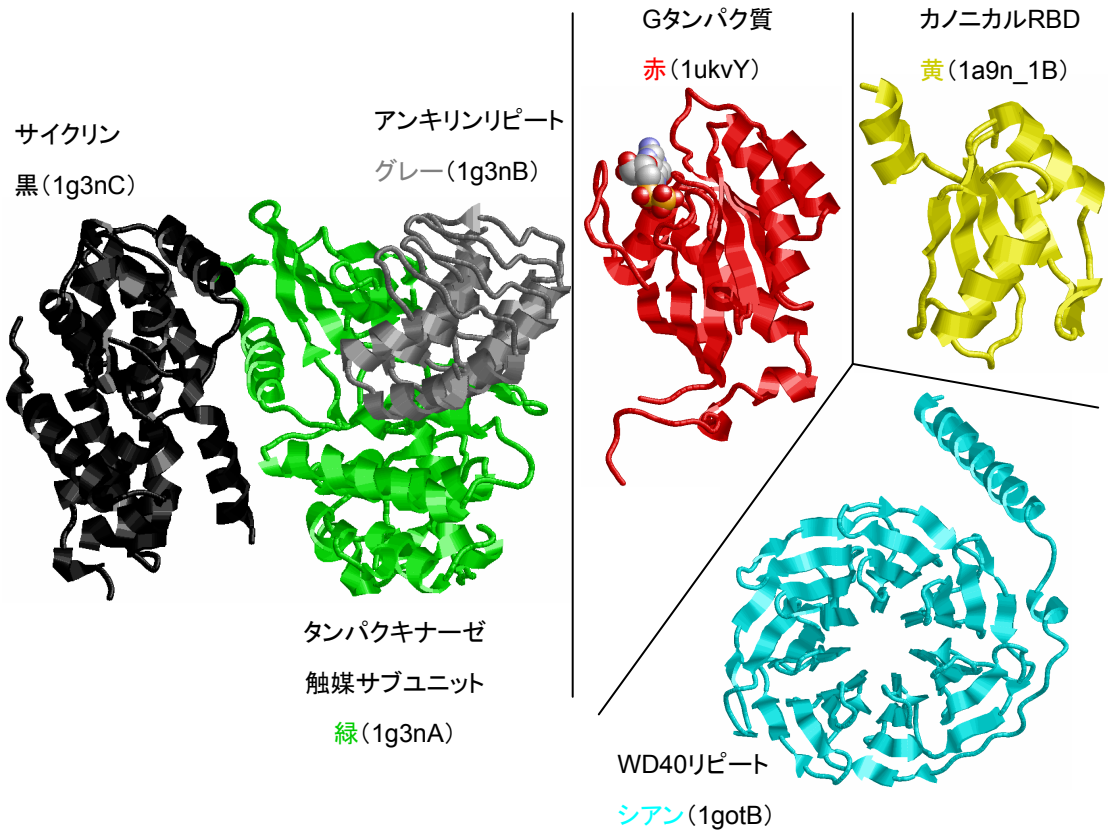


図４－３　相互作用するタンパク質ペアと相互作用しないペアに
よく現れる６種類のタンパク質ファミリーの立体構造

サイクリン、アンキリンリピートおよびタンパクキナーゼ触媒サブユニットに関しては三量体の立体構造が解かれており、アンキリンリピートはタンパクキナーゼ触媒サブユニットのインヒビターとして機能している。

４.４　コンタクトエネルギーを加えて改善されたペアのファミリー

コンタクトエネルギーを配列類似度に結合することで、実際にどのタイプのタンパク質ペアが予測されたのだろうか。異なるスコアを用いた予測を比較す

るために、F-measure が最大となるときの Z スコアを閾値とした。つまり、スコアがその閾値を下回ったら、そのタンパク質ペアは予測されたものとみなす。配列類似度のみ、および配列類似度とコンタクトエネルギーを結合したときのスコアの閾値はそれぞれ-20.8、-26.1 である。予測されたタンパク質ペアの数を表４－２にまとめた。

表４－２　予測されたタンパク質ペアの数

予測	DIP	Seq	Seq + Con
○	○	162	175
○	×	151	135
×	○	255	242
×	×	9757	9773

表４－２において、○は相互作用する、×は相互作用しないことを表す。例えば予測○、DIP○とは、複合体立体構造モデルが構築できるタンパク質ペアのうち、相互作用が DIP データベースに報告され、かつ予測されたものを表している。Seq は配列類似度による予測、Seq + Con は配列類似度とコンタクトエネルギーの結合スコアによる予測を表す。「予測○　DIP○」のタンパク質ペアの数は 162 ペアから 175 ペアに増加した。その差である 13 ペアの内訳は、20 ペアが増加し、7 ペアが減少している。その増加した 20 ペア、すなわち相互作用が DIP データベースに報告され、かつコンタクトエネルギーを加えたことで新たに予測されたタンパク質ペアを表４－３に示す。

表４－３　コンタクトエネルギーを加えたことで新たに予測された相互作用が DIP データベースに報告されているタンパク質ペア

タンパク質ペア	テンプレート	SqID [%]	Con	テンプレート構造のファミリー
MYO1 / MLC1	1br1_1A / 1br1_1B	47 / 39	-8.08	c.37.1.9 - b.34.3.1 / a.39.1.5
MYO2 / CALM	1w7jA / 1w7jB	52 / 43	-7.88	c.37.1.9 - b.34.3.1/ a.39.1.5
MYO3 / CALM	1w7jA / 1w7jB	38 / 43	-6.85	c.37.1.9 - b.34.3.1 / a.39.1.5
MYO4 / CALM	1w7jA / 1w7jB	50 / 43	-8.89	c.37.1.9 - b.34.3.1 / a.39.1.5
MYO4 / MLC1	1w7jA / 1w7jB	50 / 41	-9.21	c.37.1.9 - b.34.3.1 / a.39.1.5
MYO5 / CALM	1w7jA / 1w7jB	37 / 43	-6.78	c.37.1.9 - b.34.3.1 / a.39.1.5
VPS21 / GDI1	1ukvY / 1ukvG	34 / 100	-9.12	c.37.1.8 / c.3.1.3 - d.16.1.6 - c.3.1.3

YPT1 / RAEP	1ukvY / 1ukvG	100 / 26	-7.56	c.37.1.8 / c.3.1.3 - d.16.1.6 - c.3.1.3
YPT6 / GDI1	1ukvY / 1ukvG	35 / 100	-9.95	c.37.1.8 / c.3.1.3 - d.16.1.6 - c.3.1.3
YPT7 / GDI1	1ukvY / 1ukvG	35 / 100	-9.67	c.37.1.8 / c.3.1.3 - d.16.1.6 - c.3.1.3
YPT52 / GDI1	1ukvY / 1ukvG	32 / 100	-9.25	c.37.1.8 / c.3.1.3 - d.16.1.6 - c.3.1.3
CG22 / CDC28	1fin_1B / 1fin_1A	33 / 63	-5.81	a.74.1.1 - a.74.1.1 / d.144.1.7
CG23 / CDC28	1fin_1B / 1fin_1A	31 / 63	-7.57	a.74.1.1 - a.74.1.1 / d.144.1.7
CG24 / CDC28	1fin_1B / 1fin_1A	31 / 63	-7.56	a.74.1.1 - a.74.1.1 / d.144.1.7
SMC1 / SMC2	1xexB / 1xexA	39 / 40	-10.17	（未登録）
SMC1 / SMC3	1xexB / 1xexA	40 / 37	-10.55	（未登録）
SMC2 / SMC3	1xexA / 1xexB	40 / 39	-9.27	（未登録）
CAL1 / RAM2	1tny_2D / 1tny_2C	30 / 30	-7.44	（未登録）
RAM1 / RAM2	1tn7B / 1tn7A	40 / 30	-7.55	（未登録）
CAPZA / CAPZB	1izn_1A / 1izn_1B	32 / 56	-8.85	e.43.1.1 / e.43.1.2

タンパク質名は UniProt データベースのエントリ名であり、テンプレートの PDB コード、標的配列とテンプレート構造の間の同一残基率(SqID [%])、コンタクトエネルギーの Z スコア(Con)、を併記した。テンプレート構造のファミリーにおいて、SCOP ID は次のとおりである。a.39.1.5 : カルモジュリン様, a.74.1.1 : サイクリン, b.34.3.1 : ミオシン S1 フラグメント N 末ドメイン, c.3.1.3 : GDI 様 N 末ドメイン, c.37.1.8 : G タンパク質, c.37.1.9 : モータータンパク質, d.16.1.6 : GDI 様, d.144.1.7 : タンパクキナーゼ触媒サブユニット, e.43.1.1 : Capz α-1 サブユニット, e.43.1.2 : Capz β-1 サブユニット

表４－３によく現れる４つのファミリーのペアは次のとおりである。ミオシンとカルモジュリン様タンパク質のペア（６ペア)、G タンパク質と GDI 様タンパク質のペア（５ペア)、サイクリンとタンパクキナーゼのペア（３ペア)、染色体の構造維持(SMC)タンパク質間のペア（３ペア)。このことから、コンタクトエネルギーを加えて改善されたタンパク質ペアには細胞機能に関わる様々なファミリーが含まれており、一つのファミリーに偏っているわけではないことが分かった。

次に、表４－３に含まれている G タンパク質と GDI 様タンパク質の５ペアに注目し、コンタクトエネルギーによる改善の原因を探ることにした。GDI は GDP 解離インヒビターの略で、GDP が解離して GTP と交換されるのを阻害す

る、すなわち GEF のインヒビターである。G タンパク質と GDI 様タンパク質のテンプレートの立体構造を図４－４に示す。

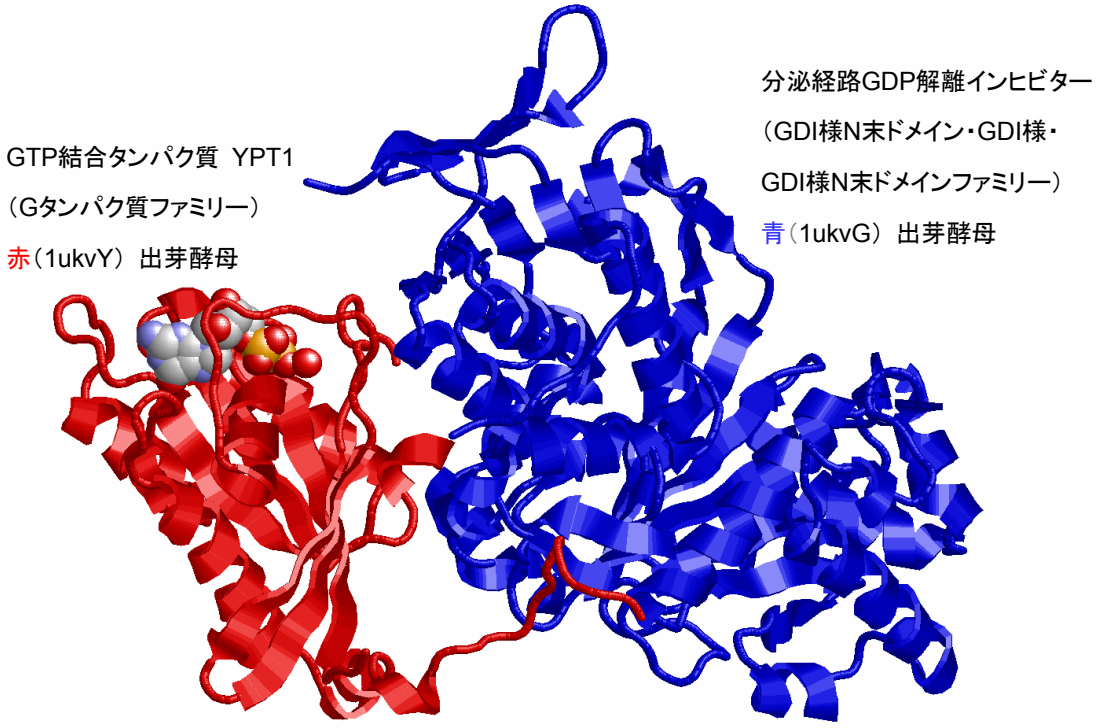


図４－４　G タンパク質と GDI 様タンパク質ペアのテンプレート構造

図４－４の立体構造を観察すると、G タンパク質の GDP は GDI 様タンパク質との相互作用面には位置していないことが分かる。さらに、その相互作用面は比較的面積が大きく、疎水的なアミノ酸が多く含まれており、コンタクトエネルギーのスコアが良くなるための要素が含まれていると考えられる。

最後に、表４－２において「予測○ DIP○」のタンパク質ペアだけでなく、それ以外の、「予測○ DIP×、予測× DIP○、予測× DIP×」のタンパク質ペアについてもコンタクトエネルギーによるスコアの変化を調査することにした。G タンパク質と GDI 様タンパク質のペアだけでなく、G タンパク質が関与するシグナル伝達系の全てのタンパク質ペアについても包括的に調べるため、G タンパク質とその相互作用する相手タンパク質の間の複合体モデル構造の Z

スコアマトリックスを作成した。配列類似度を用いた場合は図４－５、配列類似度とコンタクトエネルギーの結合スコアを用いた場合は図４－６に示す。

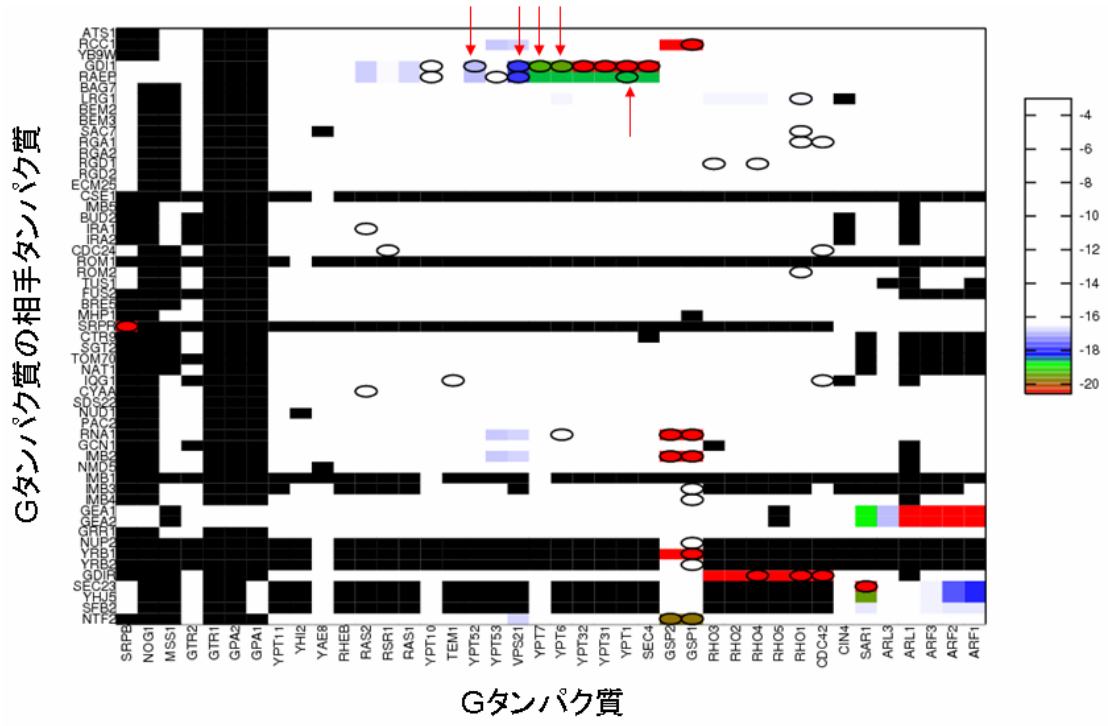


図４－５　配列類似度を用いた場合のＧタンパク質とその相手タンパク質の間の複合体モデル構造のＺスコアマトリックス

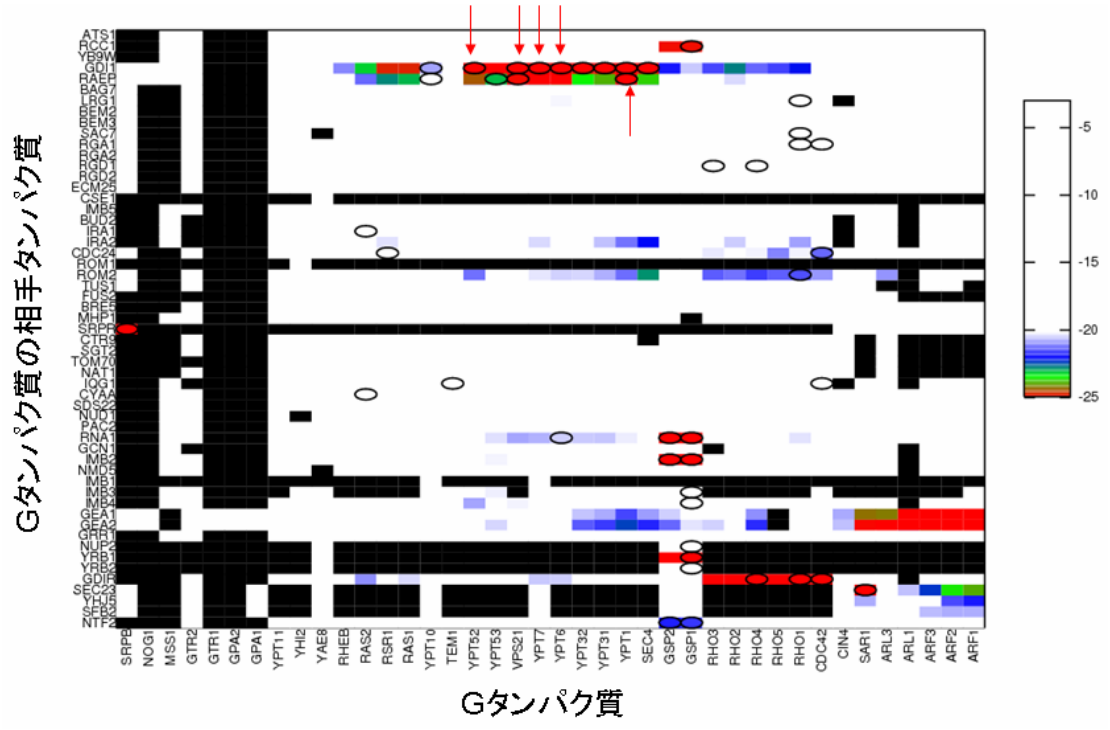


図４－６　配列類似度とコンタクトエネルギーの結合スコアを用いた場合の G タンパク質とその相手タンパク質の間の複合体モデル構造の Z スコアマトリックス

図４－５、４－６の縦軸と横軸のタンパク質名は、BLAST のスコアに基づいて生成した系統樹のタンパク質の順に並べた。それぞれの色は Z スコアを表し、白は Precision が 0.2 より小さいときの Z スコア、青、緑、赤はそれぞれ Precision が 0.3、0.4、0.5 に等しいときの Z スコアに対応している。黒く塗りつぶしたます目はホモロジーモデリングができないタンパク質ペアを表し、丸は相互作用が DIP データベースに報告されているタンパク質ペアを表している。赤の矢印はコンタクトエネルギーを加えて改善されたタンパク質ペア(YPT52 / GDI1, VPS21 / GDI1, YPT7 / GDI1, YPT6 / GDI1, YPT1 / RAEP)を示し、これらは表４－２に含まれていたものである。図４－５、４－６において丸がついていない赤います目がいくつか存在しているが、これらは相互作用が DIP データベースに報告されていないが、高い Precision 値で予測されたタンパク

質ペアである。これらのペアは、実験では相互作用が検出できなかったが実際には相互作用する可能性が高いタンパク質ペアであると考えられる。

４.５ DIP データベースに含まれていない相互作用の検出

ハイスループットの実験にはかなりの量の不正確なデータが含まれているとして、タンパク質間相互作用の実験の精度が問題視されている(Deane *et al.*, 2002; von Mering *et al.*, 2002; Sprinzak *et al.*, 2003)。本研究では、DIP データを正しいとみなしたが、その中に偽陽性が含まれている可能性はあり、また DIP データに含まれていないタンパク質ペアが細胞内で相互作用する可能性も否定できない。本手法のほとんどのパラメータは DIP データから独立して決定されており、DIP を用いて決定したのはスコアの閾値だけである。したがって、特に「偽陰性」と呼ばれているタンパク質ペア、つまり DIP データベースに登録されていないが細胞内で起こりうる相互作用を発見できるかもしれない。以下に示す手順で、相互作用する可能性のあるタンパク質ペアを本手法で検出できるかどうかを確認した。最近決定されたタンパク質間相互作用データで DIP データベースにまだ登録されていないものを用意し、DIP データや本手法で予測されたタンパク質ペアとどの程度重複しているのかを調査した。最新の実験で検出された相互作用なので、DIP データベースにまだ登録されていなくても信頼性の高い相互作用データであると考えられる。まず 2006 年に報告された Krogan らの研究(Krogan *et al.*, 2006)で得られた、7,123 個のタンパク質間相互作用のコアセットを準備した。本研究の予測と Krogan のコアセットとの重複の割合を図４－７に示す。

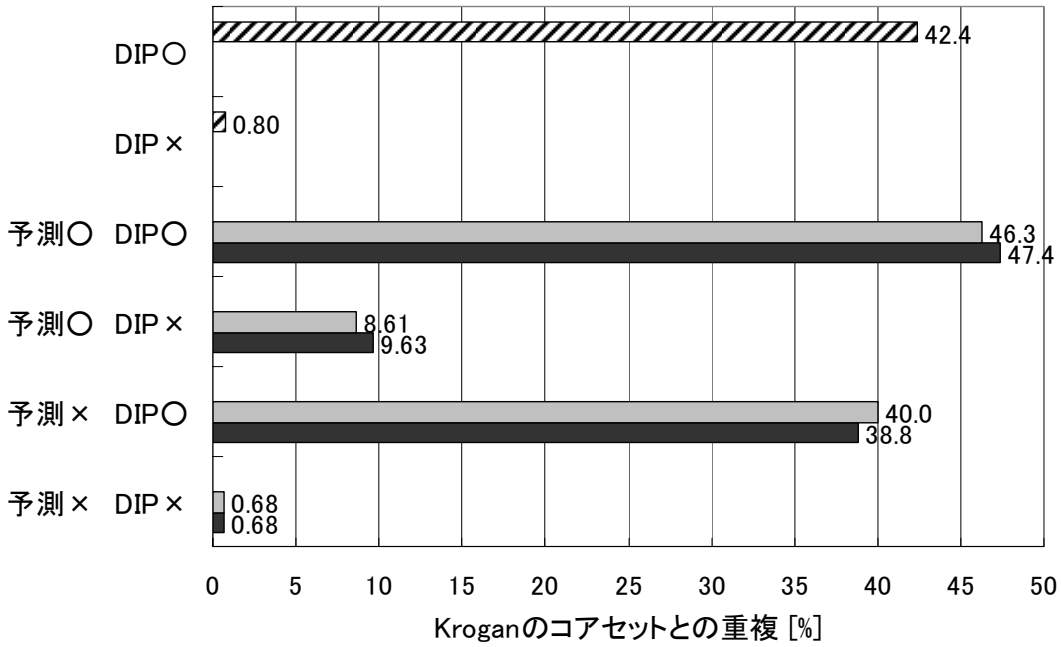


図４－７　本研究の予測と Krogan らの実験データセットとの比較

図４－７において、○は相互作用する、×は相互作用しないことを表す。例えば予測○、DIP○とは、相互作用が DIP データベースに報告され、かつ予測されたタンパク質ペアを表している。すなわち、斜線の棒グラフは複合体モデル構造が構築できたタンパク質ペア（3.2 節参照）のうち、相互作用が DIP データベースに報告されているもの（417 個）および報告されていないもの（9,908 個）と、Krogan のコアセットとの重複を表している。グレーと黒に塗りつぶした棒グラフは、表４－２に含まれている８タイプのタンパク質ペアと Krogan のコアセットとの重複を表している。グレーの棒グラフは配列類似度のみ、黒の棒グラフは配列類似度とコンタクトエネルギーの結合スコアを用いて予測した場合を表す。「DIP○」と Krogan のセットとの重複は 42.4%であり、「DIP×」との重複(0.80%)よりも大幅に大きい。次に、本研究で予測されたタンパク質ペア（表４－２で示した）と Krogan のコアセットとの重複を調べた。配列類似度の場合、「予測○　DIP×」のペアと Krogan のコアセットとの重複は 8.61%（１３ペア）であり、配列類似度とコンタクトエネルギーの結合スコアの場合、9.63%（１３ペア）であった。カイ二乗検定の結果、これらの重複（8.61%,

9.63%）は「DIP×」の重複 (0.80%)よりも有意に大きいことが示された($p < 0.01$)。したがって、複合体モデル構造が構築できるタンパク質ペアのうち相互作用が DIP に報告されていないタンパク質ペアであっても、本手法で相互作用すると予測されたものは予測されなかったペアよりも実際に相互作用する可能性が高いといえる。

４.６　より信頼性の高い評価基準への適用

本研究の評価データとして用いてきた「相互作用するタンパク質ペア」と「相互作用しないタンパク質ペア」は、DIP データベースに酵母の全てのタンパク質間相互作用が登録されていると仮定して作成したものである。しかしながら、4.5 節でも述べたように DIP データベースに登録されていないが細胞内で起こりうる相互作用はまだ存在していると考えられ、特に相互作用しないタンパク質ペアの信頼性が低いことは無視できない。先行研究においては、相互作用しないタンパク質ペアを、ランダムにシャッフルした配列ペアの生成、ランダムに選んだ配列ペアの生成、細胞内局在が異なるタンパク質ペアの生成などで行った例もある。近年、酵母を用いた大規模な細胞内局在を同定する実験が行われており、これらを用いてタンパク質間相互作用の信頼性を評価する試みがなされている(Kumar *et al.*, 2002; Ghaemmaghami *et al.*, 2003; Huh *et al.*, 2003) 。本研究でも、Ben-Hur & Noble, 2006 や Li *et al.*, 2006 の研究にならい、細胞内局在情報を用いてより信頼性の高い評価データだけを抽出し、同様の計算を行うことで結果の信頼性を検討することにした。

細胞内局在情報は、2005 年 11 月 14 日にダウンロードした MIPS データベースのデータを用いた(Guldener *et al.*, 2006)。酵母の場合、それぞれのタンパク質に１９種類の細胞内局在が三桁の番号で記載されている（図４－８）。

（７０１）細胞外	（７０５）芽	（７１０）細胞壁
（７１５）細胞表面	（７２０）原形質膜	（７２２）内膜
（７２５）細胞質	（７３０）細胞骨格	（７３５）小胞体
（７４０）ゴルジ体	（７４５）輸送小胞	（７５０）核
（７５５）ミトコンドリア	（７６０）ペルオキシソーム	

（７６５）エンドソーム　（７７０）液胞　（７７５）ミクロソーム
（７８０）脂質粒子　（７９９）その他の細胞内局在

図４－８　MIPS に登録されている酵母の細胞内局在の種類

タンパク質の中には細胞内局在が不明のものや複数の細胞内局在を持つものも存在し、MIPS に登録されている 5,211 個のタンパク質について 13,815 個の細胞内局在が決定されている。このタンパク質の細胞内局在（複数の細胞内局在を持つタンパク質についてはその中の一つ）の統計は図４－９のようになり、核、細胞質、ミトコンドリアが全体の 70%を占めていた。

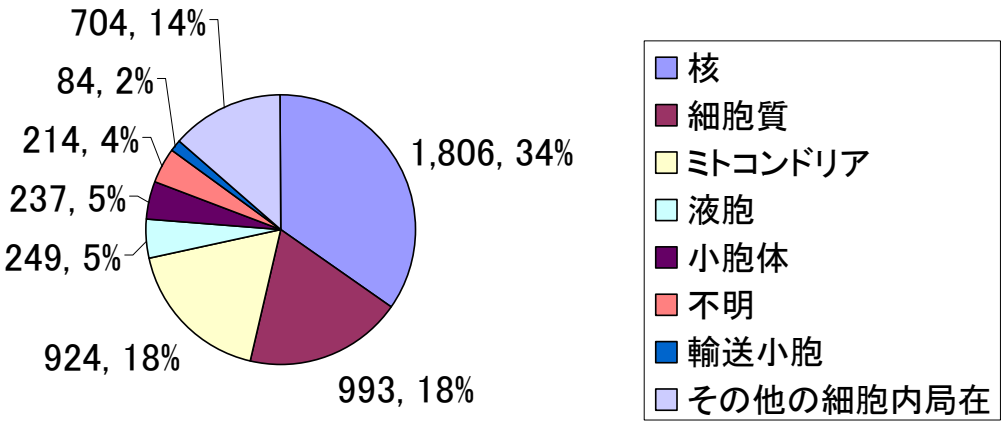


図４－９　MIPS に登録されている酵母の細胞内局在の統計

本研究の相互作用するタンパク質ペア（417 個）と相互作用しないタンパク質ペア（9,908 個）のそれぞれを、(i) ペアを構成する二つのタンパク質の細胞内局在が少なくとも一つ重複している場合、(ii) 少なくとも一つのタンパク質

の細胞内局在が不明である場合、(iii) 細胞内局在が全く重複していない場合の３種類に分類した（表４－４）。

表４－４　相互作用するタンパク質ペアと相互作用しないペアの細胞内局在情報による分類

	相互作用するペア	相互作用しないペア
(i) 細胞内局在が重複しているペア	380（正例）	5,536
(ii) 少なくとも一つのタンパク質の細胞内局在が不明であるペア	10	1,573
(iii) 細胞内局在が重複していないペア	27	2,799（負例）
計	417	9,908

本研究の相互作用するタンパク質ペアの中で細胞内局在が重複しているものについては相互作用するペアの中でも信頼性の高いペア（正例）だと考え、相互作用しないタンパク質ペアの中で細胞内局在が重複していないものについても信頼性の高いペア（負例）だとすると、正例が 380 個、負例が 2,799 個得られた。相互作用するペアについては細胞内局在が重複しているものに限定しても 37 ペアしか減少しなかったが、相互作用しないペアについては 7,109 ペアも減少した。

こうして抽出された高信頼性のデータセットにおける正例と負例の識別力を調べるために、本研究の各特徴量（コンタクトエネルギー、静電エネルギー、配列類似度）の単独および結合スコアから Recall-Precision プロットとその F-measure の最大値を計算すると図４－１０、図４－１１が得られた。

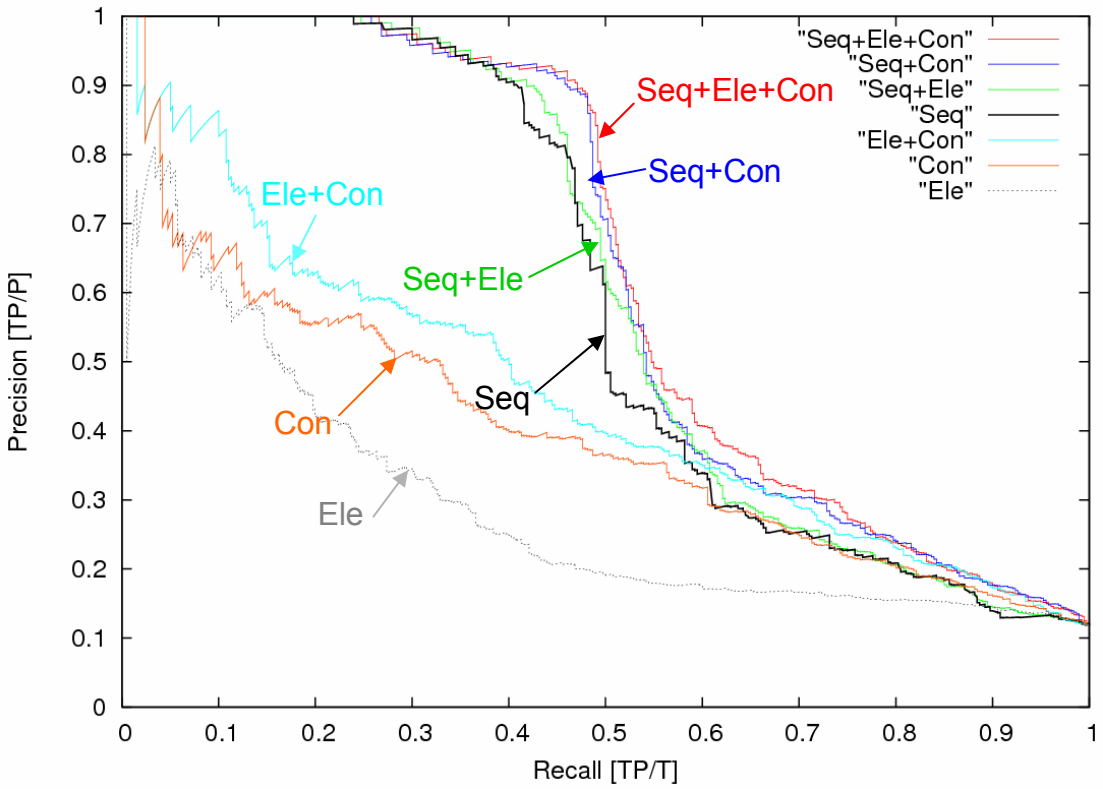


図４－１０　細胞内局在情報を用いて作成したより信頼性の高い評価データの識別の Recall-Precision プロット

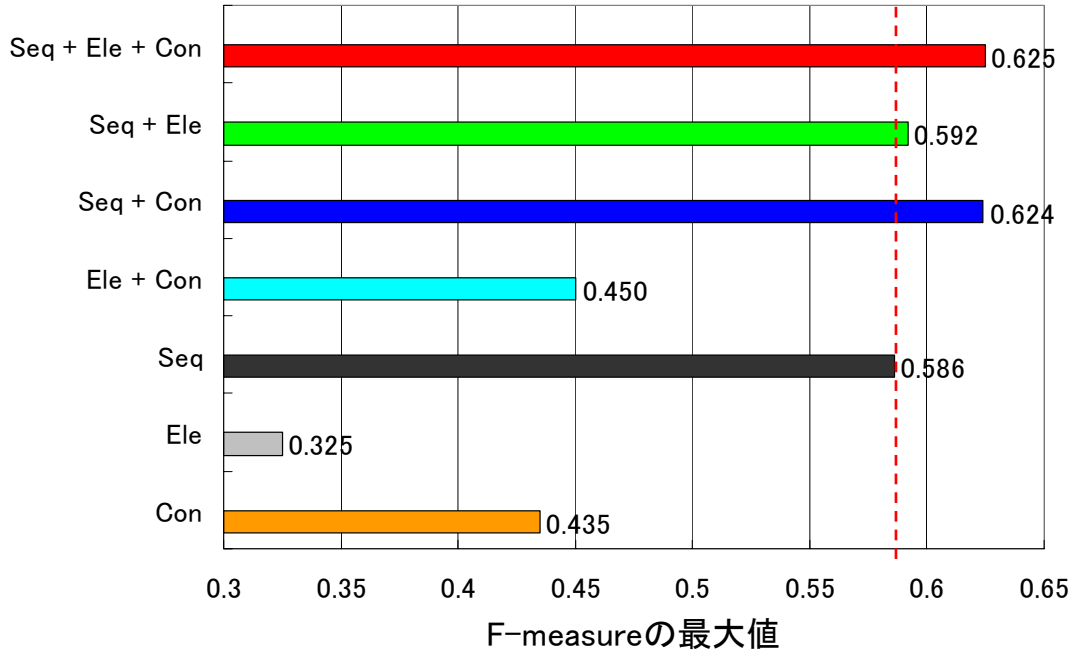


図４－１１　細胞内局在情報を用いて作成したより信頼性の高い評価データの識別の Recall-Precision プロットの F-measure の最大値

図３－６と図４－１０、図３－７と図４－１１をそれぞれ比較すると、各特徴量の単独および結合スコアの識別力の順位は変化していないことが分かる。図４－１１について統計的有意性を確認するため、再度ブートストラップサンプリングテストを行うと、配列類似度とコンタクトエネルギーの結合スコア(Seq + Con)および全特徴量の結合スコア(Seq + Con + Ele)を用いた場合に、1,000 個全ての F-measure の最大値が配列類似度の場合(Seq)を超え、配列類似度と静電エネルギーの結合スコアの場合(Seq + Ele)、809 個の F-measure の最大値だけが配列類似度(Seq)の場合を超えた。したがって、配列類似度にコンタクトエネルギーを結合したときと、全特徴量を結合したときに有意な改善($p < 0.01$)を示し、図３－７のときと同様の結果を得た。より信頼性の高い評価データに適用しても同様の結果が得られたことは、本研究の結論の信頼性が高いことを示すと考えられる。

尚、細胞内局在情報を用いて作成した評価データを用いたときの方が全体に特徴量の識別力が高い。これに関しては、細胞内局在情報を用いない評価デー

タのときは10,325個のタンパク質ペアの中から417個の相互作用するタンパク質ペアを予測しなければならなかったが、細胞内局在情報を用いた評価データのときは、相互作用しないタンパク質ペアの数が約七千個減少したことにより3,179 個のタンパク質ペアの中から 380 個の正例を予測すればよいため、ランダムな予測での Precision が単純に増加したからだと考えられる。

第５章　ウェブサーバの開発

本予測手法を他の研究者が自由に利用できるように、タンパク質間相互作用予測サーバ HOMCOS (HOmology Modeling of protein COmplex Structure) [http://biunit.naist.jp/homcos]を開発した（図５－１）。

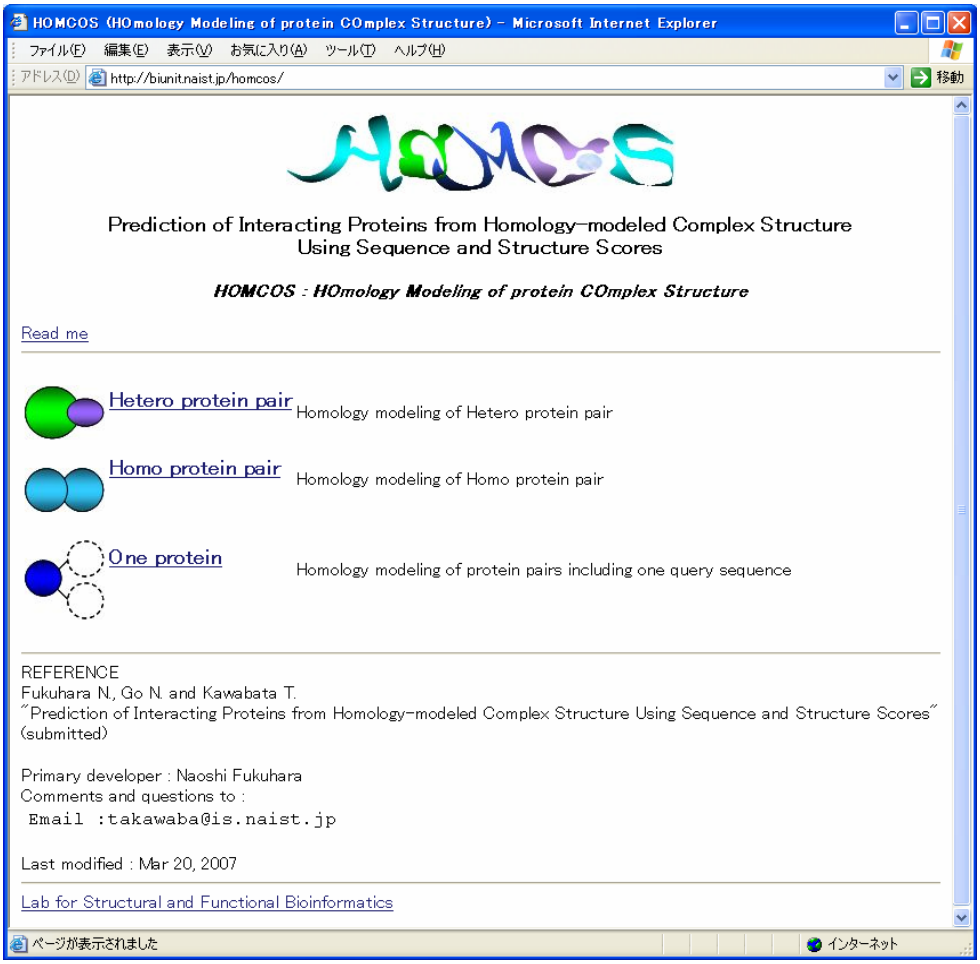


図５－１　HOMCOS のトップページ（http://biunit.naist.jp/homcos/）

HOMCOS は、タンパク質複合体のホモロジーモデリングを主要な機能とし、テンプレートの複合体立体構造の候補の列挙、アラインメントや相互作用残基の情報の表示、複合体モデル構造のエネルギー計算と表示、標的タンパク質配列の相互作用の予測などを行うことができる。

トップページからヘテロダイマーのモデリングのページに移動する（図５－２）。ここでは、二つの標的タンパク質配列をコピーして「SEARCH」をクリックする。このとき、複合体モデル構造の静電エネルギー計算も行いたい場合は「Calculate electrostatic energy」をチェックしてから「SEARCH」をクリックするが、静電エネルギーは全原子の計算が必要であるため計算時間が多くかかる点に注意が必要である。

図５－２　ヘテロダイマーのモデリングのページ

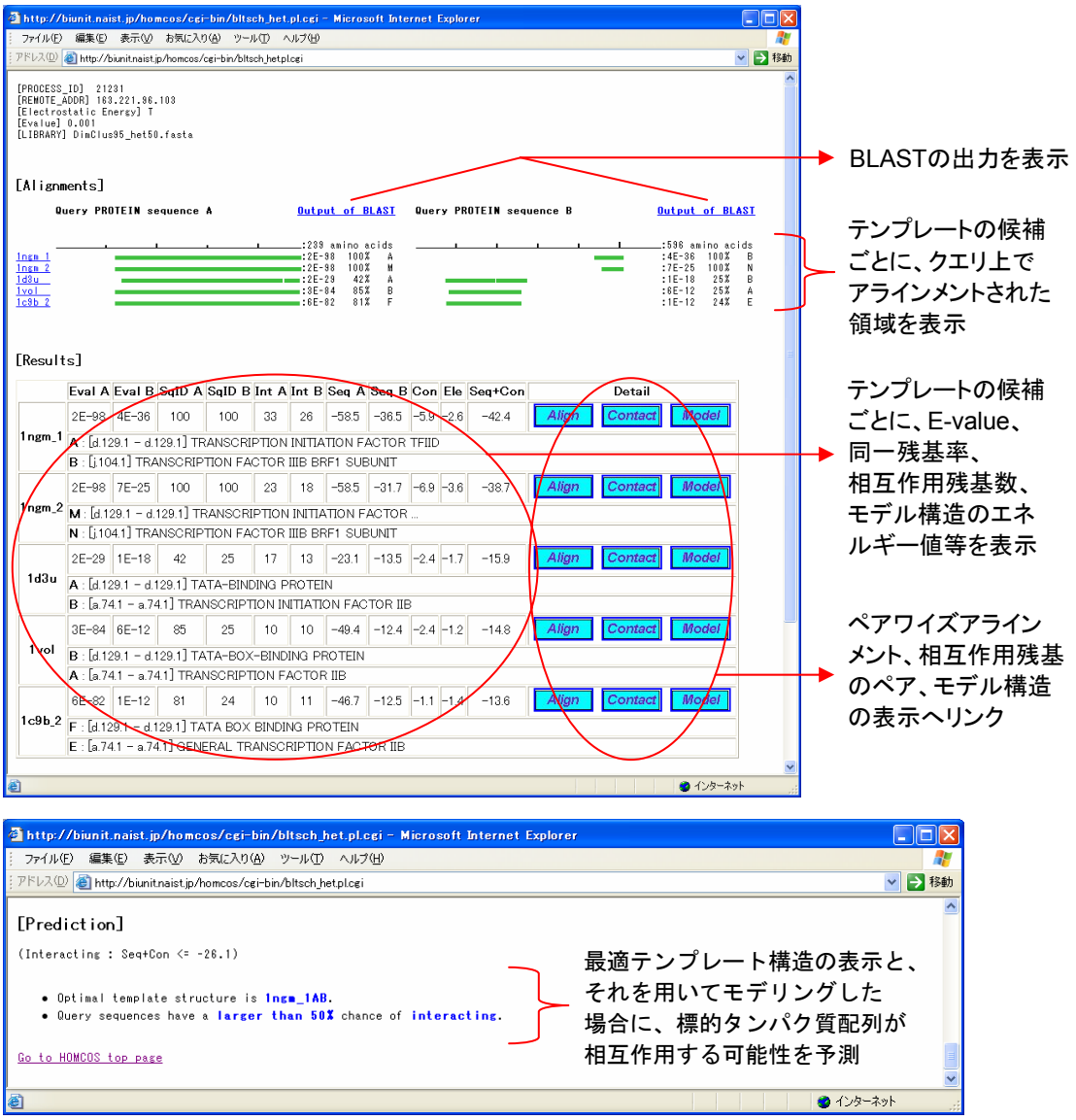


図５－３　結果の出力画面

エネルギー計算が終了すると、結果が出力される（図５－３）。ページの上部にはクエリ上でアラインメントされた領域がバーで表示され、中部にはテンプレート構造と標的タンパク質配列間の配列類似度（E-value と同一残基率）、相互作用残基数、モデル構造のエネルギー値のＺスコア（配列類似度、コンタクトエネルギー、配列類似度とコンタクトエネルギーの結合スコア）、テンプ

レートのタンパク質の機能が表の要素として表示される。これらは、各テンプレート構造の候補について、モデル構造のエネルギー値が安定な順に並べられている。ページの下部には、モデル構造のエネルギー値が最安定となるテンプレート構造が示され、それを用いてモデリングした場合に二つの標的タンパク質配列が相互作用する可能性が予測結果として出力される。

　ページ中部の表の右端の要素はモデリングの詳細な情報へのリンクが貼ってあり、ペアワイズアラインメント、相互作用残基のペア、複合体モデル構造が表示される（図５－４）。

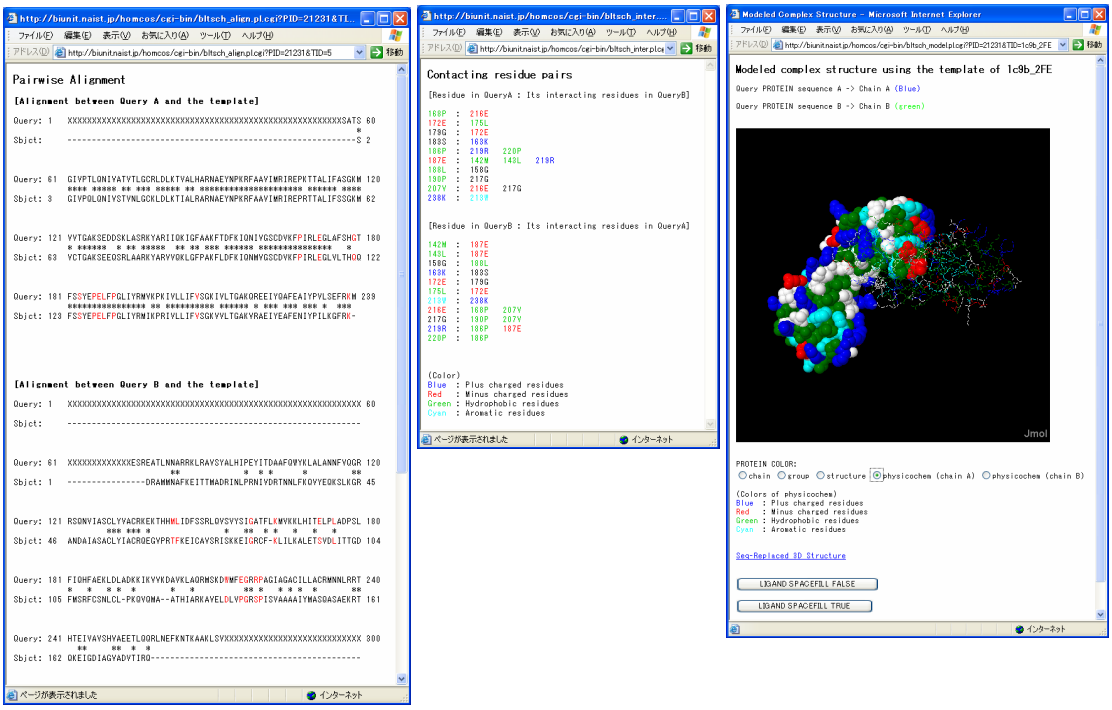


図５－４　モデリングの詳細な情報の表示

ペアワイズアラインメントでは相互作用残基が赤のアミノ酸で表示されており、相互作用残基のペアではアミノ酸が物理化学的な性質（疎水性、荷電性、芳香族性）に応じて色分けされている。複合体モデル構造の表示では、相互作用の様子が立体構造上で観察でき、テンプレート構造のアミノ酸残基を標的タンパク質配列の対応するアミノ酸残基に置き換えたモデル構造の立体構造ファイルもユーザーが取得できる。

トップページ（図５－１）からは、ホモダイマーのモデリングなどもでき、またあらゆる生物種のタンパク質配列に対しても広く適用可能であるため、有用性の高いツールである。今後はバイオの実験系の研究者の需要も考慮して、さらに付加機能を充実させていく方針である。

第６章　結論

本研究では、複合体モデル構造の中から「相互作用するタンパク質ペア」と「相互作用しないタンパク質ペア」を識別する手法を開発した。特徴量として、従来から用いられてきたコンタクトエネルギーに加え、新たに静電エネルギーおよびテンプレート構造との配列類似度を採用した。複合体がホモロジーモデリング可能であるタンパク質ペアに制限したので、本研究のアプローチは本質的に、相同タンパク質の中から特異的に相互作用するタンパク質を識別することを目的としている。

先行研究においては、相互作用すると予測されたタンパク質ペアと、実験で相互作用が報告されたペアの間の重複によって Recall（再現率）の性能評価がなされていた。しかしながら、Recall はスコアの閾値を緩めれば増加しうるものなので、これでは適切に評価されているとはいえない。本研究では、相互作用しないと予測されたペアについても評価を行うべきだと考え、Recall（再現率）のみならず Precision（適合率）を評価することで、初めて意味のある評価を行うことができた。

具体的な評価方法としては、本手法を酵母の全ヘテロタンパク質ペアに適用し、DIP データベースに報告されたタンパク質ペアと報告されていないペアをそれぞれ相互作用するペアと相互作用しないペアとみなして、両者の識別力で性能評価を行った。このように評価データを作成した根拠としては、酵母においては既に大量のタンパク質間相互作用データが存在しており、いわゆる「偽陰性」のデータ（DIP に登録されていないが、実際には細胞内で相互作用するもの）の数は、他の生物種よりも少ないと考えたからである。しかしながら、偽陰性のデータは相当数存在していると考えられ、本研究の相互作用しないタンパク質ペアの信頼性が低いことは無視できない。これを踏まえ、我々は細胞内局在情報を導入して信頼性の高い評価データを作成し、評価を行った。

Recall-Precision プロットによる評価から、テンプレート構造との配列類似度は立体構造に基づくスコア（コンタクトエネルギーと静電エネルギー）よりも大幅に識別力が大きいことが示された。また先行研究と性能比較を行うと、先行研究の手法は、本手法のコンタクトエネルギーよりやや識別力が高いが、テンプレート構造との配列類似度よりも大幅に識別力が小さいことが示された。しかし、配列類似度にコンタクトエネルギーを結合したスコアを用いることで識別力を有意に改善することができた。

この分野の過去の研究においては、ホモロジーモデリングの段階で配列情報が用いられており、生成したモデル構造の相互作用の評価についてはコンタクトエネルギーのような立体構造情報のみを用いればよいと考えられてきた。しかしながら、本研究によって、配列類似度にコンタクトエネルギーを結合したスコアが最も良い識別力を持っていることが示されたので、モデル構造の相互作用の評価においても、配列情報を立体構造情報に組み合わせて用いるべきであることが示された。したがって、本研究は複合体立体構造モデルの相互作用の評価に最適な特徴量の組み合わせを提供したものとして位置付けられるといえる。

本研究の適用範囲・限界は、(i) 複合体モデル構造が作成可能なタンパク質ペアのみが対象であること、(ii) モデル構造を扱うことによる予測精度の低下、の二点が挙げられる。(i) に関しては DIP データベースに登録されている酵母の約２万個のタンパク質ペアのうち、本研究でモデリング可能なものは 400 程度に留まっており、ホモロジーモデリングのテンプレート構造ライブラリの複合体の数（2,635 個）がネックになっていると考えられる。今後、タンパクキナーゼのように多くの相同タンパク質を持つタンパク質の複合体立体構造が解かれれば、モデリング可能なものの割合も急激に増加していくものと考えられる。(ii) に関しては、例えば本研究の静電エネルギーの性能の低下の原因として考えられ、置換後の残基に電荷が正確に置かれていないこと、ヌクレオチドや金属イオンのようなリガンドの電荷が考慮できていないことなどが挙げられる。静電相互作用はタンパク質間相互作用に重要な役割を果たしているはずなので、モデル構造に対しても有効に働く静電エネルギーへと改良していくことが必要であると考えられる。

最後に、本研究手法を他の研究者にも自由に利用してもらえるようにするため、複合体立体構造モデルの情報を提供するウェブサーバを開発した。本ウェブサーバを有効活用するためには、Recall-Precision プロットの結果からも分かるように、より多くのタンパク質間相互作用の候補を挙げることを目的とするのであれば、スコアの閾値を緩めて Recall を 0.4 程度まで引き上げればよいが、Precision が 0.5 程度に低下してしまうため、予測されたタンパク質ペアの相互作用を実験でも検証することが必要となると考えられる。逆に実験をしなくてもよい確実な相互作用を抽出したいのであればスコアの閾値をきつくして Precision を 0.8 程度まで引き上げることとなる。

本ウェブサーバをさらに改良するために、二つのタンパク質配列の相互作用を予測するだけではなく、一つのタンパク質配列を入力するとそれに結合する可能性のある全てのタンパク質を予測できるようにすることが必要であると考えられる。タンパク質間相互作用を明らかにしていく過程において、候補となる二つのタンパク質が初めから既知である場合はそれほど多くなく、むしろ一つの興味あるタンパク質から芋づる式にそれに結合するタンパク質を解明していくのが通常であると考えられるからである。

謝辞

本論文に記した一連の研究は、国立大学法人 奈良先端科学技術大学院大学 情報科学研究科 情報生命科学専攻 博士後期課程在学中に、人材養成ユニットおよび蛋白質機能予測学講座において行われたものです。本研究を進める上で多くの皆様のご支援とご教授を承りました。様々なサポートをして頂いた方々に感謝の意をここで表したいと思います。

はじめに、川端猛准教授に深く感謝いたします。５年間の研究活動全般にわたって多くの御指導を賜りました。川端准教授の御指導がなければ本論文の完成はなかったと思っております。

郷信広博士、由良敬博士、河野秀俊博士、中村建介博士をはじめとする日本原子力研究開発機構の研究員の皆様にはセミナーなどを通して熱心なディスカッションと励ましを頂きました。また、Bose Institute の Dr. Gautam Basu には人材養成ユニット時代に大変お世話になりました。

箱嶋敏雄教授には本論文の主査を、小笠原直毅教授には、審査を引き受けていただきありがとうございました。本研究全般にわたり、様々な御支援、御助言を頂きました。心より感謝致します。

人材養成ユニットおよび蛋白質機能予測学講座では、鷲尾尊規博士、Dr. Jitender Jit Singh Cheema, Dr. Saharuddin Bin Mohamad, 三森智裕博士、松田敬子博士、秘書の町田淳子さんに大変お世話になりました。及川雅隆君、山極裕道君、岩西雄大君、渡邉潤也さん、吉井悠喜君、宮久保博幸さんのおかげで楽しく研究をすることができました。吉井悠喜君には、タンパク質間相互作用予測サーバ HOMCOS のロゴマークを作成して頂きました。

真木寿治教授、真木智子助教、坪田智明博士をはじめとする本学バイオサイエンス研究科 原核生物分子遺伝学講座の皆様には、共同研究および投稿論文 *Journal of Biological Chemistry* の執筆に際し大変お世話になりました。

Dr. Andrej Sali, Dr. Fred P. Davis には、本研究と先行研究の性能比較に必要な予測された相互作用するタンパク質ペアのリストを頂きました。本論文のみならず投稿論文 *BIOPHYSICS* の重要データとして活用させて頂き、大変感謝しております。

在学中は、バイオ COE リサーチ・アシスタント、日本学生支援機構の奨学金制度、授業料免除などの奈良先端科学技術大学院大学の様々な経済支援に助けられてきました。本学で得たものは大変多く、入学してよかったと心から思っております。

最後に、これまで学生生活を見守り、支えてくれた両親に感謝致します。

引用文献

Aloy,P. and Russell,R.B. Interrogating protein interaction networks through structural biology. (2002) *Proc. Natl Acad. Sci. USA,* **99**, 5896-5901.

Aloy,P., Ceulemans,H., Stark,A. and Russell,R.B. The relationship between sequence and interaction divergence in proteins. (2003) *J. Mol. Biol.,* **332**, 989-998.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. (1997) *Nucleic Acids Res.,* **25**, 3389-3402.

Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A.G. SCOP database in 2004: refinements integrate structure and sequence family data. (2004) *Nucleic Acids Res.,* **32**, D226-D229.

Bader,G.D., Betel,D. and Hogue,C.W.V. BIND: the biomolecular interaction network database. (2003) *Nucleic Acids Res.,* **31**, 248-250.

Ben-Hur,A. and Noble,W.S. Choosing negative examples for the prediction of protein-protein interactions. (2006) *BMC Bioinformatics*, **7**, S2.

Bork,P., Jensen,L.J., Mering,C., Ramani,A.K., Lee,I. and Marcotte,E.M. Protein interaction networks from yeast to human. (2004) *Curr. Opin. Struct. Biol.,* **14**, 292-299.

Brooks,B.R., Bruccoleri,R.E., Olafson,B.D., States,D.J., Swaminathan,S., Karplus,M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. (1983) *J. Comp. Chem.,* **4**, 187-217.

Case,D.A., Cheatham,T.E.3rd, Darden,T., Gohlke,H., Luo,R., Merz,K.M.Jr., Onufriev,A., Simmerling,C., Wang,B., Woods,RJ. The Amber biomolecular simulation programs. (2005) *J. Computat. Chem.*, **26**, 1668-1688.

Davis,F.P., Braberg,H., Shen,M.Y., Pieper,U., Sali,A. and Madhusudhan,M.S. Protein complex compositions predicted by structural similarity. (2006) *Nucleic Acids Res.,* **34**, 2943-2952.

Deane,C.M., Salwinski,L., Xenarios,I. and Eisenberg,D. Protein interactions: two methods for assessment of the reliability of high throughput observations. (2002) *Mol. Cell. Prot.,* **1**, 349-356.

Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. Protein interaction maps for complete genomes based on gene fusion events. (1999) *Nature,* **402**, 86-90.

Gavin,A.C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. (2002) *Nature*, **415**, 141-147.

Gavin,A.C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. (2006) *Nature,* **440**, 631-636.

Ghaemmaghami,S., Huh,W.K., Bower,K., Howson,R.W. Belle,A. Dephoure,N., O'Shea,E.K. and Weissman,J.S. Global analysis of protein expression in yeast. (2003) *Nature*, **425**, 737-741.

Glaser,F., Steinberg,D.M., Vakser,I.A. and Ben-Tal,N. Residue frequencies and pairing preferences at protein-protein interfaces. (2001) *Proteins,* **43**, 89-102.

Gomperts,B.D., Kramer,I.M. and Tatham,P.E.R. Protein domains and signal transduction. In:Signal Transduction. (2002) Academic Press, San Diego, pp. 393-410.

Grigoryan,G. and Keating,A.E. Structure-based prediction of bZIP partnering specificity. (2006) *J. Mol. Biol.,* **355**, 1125-1142.

Guldener,U., Munsterkotter,M., Oesterheld,M., Pagel,P., Ruepp,A., Mewes,H.W. and Stumpflen,V. MPact: the MIPS protein interaction resource on yeast. (2006) *Nucleic Acids Res.,* **34**, D436-D441.

Henrick,K. and Thornton,J.M. PQS: a protein quaternary structure file server. (1998) *Trends Biochem. Sci.,* **23**, 358-361.

Ho,Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. (2002) *Nature,* **415**, 180-183.

Huh,W.K., Falvo,J.V., Gerke,L.C., Carroll,A.S., Howson,R.W., Weissman,J.S. and O'Shea,E.K. Global analysis of protein localization in budding yeast. (2003) *Nature*, **425**, 686-691.

Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. (2001) *Proc. Natl Acad. Sci. USA*, **98**, 4569-4574.

Johnson,R.A. and Wichern,D.W. Applied multivariate statistical analysis. (1998) Prentice-Hall, London, pp. 740.

Jones,D.T., Taylor,W.R. and Thornton,J.M. A new approach to protein fold recognition. (1992) *Nature,* **358**, 86-89.

Keskin,O., Bahar,I., Badretdinov,A.Y., Ptitsyn,O.B. and Jernigan,R.L. Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions. (1998) *Protein Sci.*, **7**, 2578-2586.

Krogan,N.J. *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. (2006) *Nature,* **440**, 637-643.

Kumar,A. *et al.* Subcellular localization of the yeast proteome. (2002) *Genes & Dev.*, **16**, 707-719.

Li,M.H., Wang,X.L., Lin,L. and Liu,T. Effect of example weights on prediction of protein-protein interactions. (2006) *Computat. Biol. Chem.*, **30**, 386-392.

Lu,H., Lu,L. and Skolnick,J. Development of unified statistical potentials describing protein-protein interactions. (2003) *Biophys. J.,* **84**, 1895-1901.

Lu,L., Lu,H. and Skolnick,J. MULTIPROSPECTOR: An algorithm for the prediction of protein-protein interactions by multimeric threading. (2002) *Proteins,* **49**, 350-364.

Lu,L., Arakaki,A.K., Lu,H. and Skolnick,J. Multimeric threading-based prediction of protein-protein interactions on a genomic scale: Application to the *Saccharomyces cerevisiae* proteome. (2003) *Genome Res.,* **13**, 1146-1154.

Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. Detecting protein function and protein-protein interactions from genome sequences. (1999) *Science,* **285**, 751-753.

Marti-Renom,M.A., Stuart,A., Fiser,A., Sanchez,R., Melo,F., Sali,A. Comparative protein structure modeling of genes and genomes. (2000) *Annu. Rev. Biophys. Biomol. Struct.,* **29**, 291-325.

Matthews,B.W., Comparison of the predicted and observed secondary structure of T4 phage lysozyme. (1975) *Biochim Biophys Acta,* **405**, 442-451.

Matthews,L.R., Vaglio,P., Reboul,J., Ge,H., Davis,B.P., Garrels,J., Vincent,S. and Vidal,M. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "Interologs". (2001) *Genome Res.,* **11**, 2120-2126.

McDermott,J. and Samudrala,R. Enhanced functional information from predicted protein networks. (2004) *Trends Biotechnol.,* **22**, 60-62.

Miyazawa,S. and Jernigan,R.L. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. (1985) *Macromolecules,* **18**, 534-552.

Moont,G., Gabb,H.A. and Sternberg,M.J.E. Use of pair potentials across protein interfaces in screening predicted docked complexes. (1999) *Proteins,* **35**, 364-373.

Ofran,Y. and Rost,B. Analysing six types of protein-protein interfaces. (2003) *J. Mol. Biol.,* **325**, 377-387.

Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. (1999) *Proc. Natl Acad. Sci. USA,* **96**, 4285-4288.

Pieper,U. *et al.* MODBASE: a database of annotated comparative protein structure models and associated resources. (2006) *Nucleic Acids Res.,* **34**, D291-D295.

Rubin,G.M. *et al.* Comparative genomics of the eukaryotes. (2000) *Science,* **287**, 2204-2215.

Salwinski,L. and Eisenberg,D. Computational methods of analysis of protein-protein interactions. (2003) *Curr. Opin. Struct. Biol.,* **13**, 377-382.

Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. The Database of Interacting Proteins: 2004 update. (2004) *Nucleic Acids Res.,* **32**, D449-D451.

Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. (2003) *Genome Res.,* **13**, 2498-2504.

Shaul,Y. and Schreiber,G. Exploring the charge space of protein-protein association: a proteomic study. (2005) *Proteins,* **60**, 341-352.

Sheinerman,F.B., Norel,R. and Honig,B. Electrostatic aspects of protein-protein interactions. (2000) *Curr. Opin. Struct. Biol.,* **10**, 153-159.

Sippl,M.J. Calculation of conformational ensembles from potentials of mean force: An approach to the knowledge-based prediction of local structures in globular proteins. (1990) *J. Mol. Biol.,* **213**, 859-883.

Sprinzak,E., Sattath,S. and Margalit,H. How reliable are experimental protein-protein interaction data? (2003) *J. Mol. Biol.,* **327**, 919-923.

Uetz,P. *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. (2000) *Nature*, **403**, 623-627.

von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. and Bork,P. Comparative assessment of large-scale data sets of protein-protein interactions. (2002) *Nature,* **417**, 399-403.

Wojcik,J. and Schachter,V. Protein-protein interaction map inference using interacting domain profile pairs. (2001) *Bioinformatics,* **17**, S296-S305.

Wojcik,J., Boneca,I.G. and Legrain,P. Prediction, assessment and validation of protein interaction maps in bacteria. (2002) *J. Mol. Biol.,* **323**, 763-770.

Wu,C.H. *et al.* The universal protein resource (UniProt): an expanding universe of protein information. (2006) *Nucleic Acids Res.,* **34**, D187-D191.

業績リスト

学術論文誌

1. Fukuhara N., Go N., Kawabata T. “Prediction of interacting proteins from homology-modeled complex structures using sequence and structure scores.” *BIOPHYSICS* in press
2. Tsubota T., Tajima R., Ode K., Kubota H., Fukuhara N., Kawabata T., Maki S., Maki H. (2006) “Double-stranded DNA binding, an unusual property of DNA polymerase ε, promotes epigenetic silencing in *Saccharomyces cerevisiae*.” *Journal of Biological Chemistry* **281**, 32898-32908.

国際会議議事録（査読なし）

1. Fukuhara N., Kawabata T., Go N. “Predicted protein-protein interaction networks based on homology-modeled complex structures.” The 5th East Asian Biophysics Symposium & The 44th Annual Meeting of the Biophysical Society of Japan (EABS & BSJ 2006), Okinawa Convention Center, Okinawa, Japan, 11.12-16, 2006.
2. Fukuhara N., Kawabata T., Go N. “Accuracy of protein-protein interface prediction using homology modeling.” The 10th Pacific Symposium on Biocomputing (PSB 2005), The Fairmont Orchid, Hawaii, USA, 1.4-8, 2005.
3. Fukuhara N., Kawabata T., Go N. “Selection of proper template complex for modeling protein-protein complex.” The 15th International Conference on Genome Informatics (GIW 2004), Pacifico Yokohama, Yokohama, Japan, 12.13-15, 2004.
4. Fukuhara N., Kawabata T., Go N. “Prediction of protein-protein interaction sites using residue interface propensity.” The 1st Pacific-Rim International Conference on Protein Science (PRICPS 2004), Pacifico Yokohama, Yokohama,

Japan, 4.14-18, 2004.

国内学会発表

1. 福原直志、川端猛、郷信広
   「残基間コンタクトポテンシャルと静電エネルギーを用いたタンパク質間相互作用予測」
   第６回日本蛋白質科学会年会, 国立京都国際会館, 2006 年 4 月 24 日～26 日, 京都
2. 福原直志、川端猛、郷信広
   「統計ポテンシャルを用いたタンパク質間相互作用予測」
   第 43 回日本生物物理学会年会, 札幌コンベンションセンター, 2005 年 11 月 23 日～25 日, 札幌
3. 福原直志、川端猛、郷信広
   「統計ポテンシャルを用いたタンパク質間相互作用予測」
   第５回日本蛋白質科学会年会, 福岡国際会議場, 2005 年 6 月 30 日～7 月 2 日, 福岡

研究会発表

1. 福原直志、郷信広、川端猛
   「複合体の立体構造のホモロジーモデリングに基づくタンパク質間相互作用ネットワークの予測」
   バイオ COE サマーキャンプ 2006, コンカレントセッションにて口頭発表, 淡路夢舞台国際会議場, 2006 年 9 月 10 日～12 日, 淡路
2. 福原直志、川端猛、郷信広
   「統計ポテンシャルを用いたタンパク質間相互作用予測」
   バイオ COE サマーキャンプ 2005, 口頭発表とポスター発表, 淡路夢舞台国際会議場, 2005 年 9 月 2 日～3 日, 淡路
3. Fukuhara N., Kawabata T., Go N. “Selection of proper template complex for modeling protein-protein complex.” NAIST 21st Century Bio-COE Program International Symposium, Oral Presentation & Poster Session, Nara-Ken New Public Hall, Nara, Japan, 1.17-19, 2005.

4. Fukuhara N., Kawabata T., Go N. “Prediction of protein-protein interaction sites using residue interface propensity.” UMN / NAIST Joint workshop “Systems analysis of biological processes”, University of Minnesota, Minnesota, USA, 7.22-23, 2004.