高精度タンパク質間相互作用予測システムの開発

東京大学大学院農学生命科学研究科
清水　謙多郎

Development of High-Accuracy Protein-Protein Interaction Prediction System

Kentaro Shimizu
Department of Biotechnology, The University of Tokyo

Protein interactions play an important role in a number of biological activities. We are developing a system for predicting protein interactions in the following approaches: (1) Protein-protein interaction prediction and its network prediction, (2) Protein-protein interaction site prediction and protein-ligand binding site prediction, (3) Protein-protein docking, (4) Analysis of physical interactions between protein and other molecules. In the protein-protein docking, we developed a high-speed algorithm that uses a series expansion of basis functions which are combinations of spherical harmonics and radial base polynomials. By using the protein-protein docking and molecular simulation, we analyzed the interactions between components of Rieske non-heme iron oxygenase and revealed the correlation of binding surface properties and the electron transfer ability.

1. はじめに

ゲノム解析、プロテオーム解析などを通して、生体の内外における様々な現象が多元的な相関をもって関連づけられるようになった。同時に複数の現象をつなぐ情報・エネルギー等の流れも明らかとなってきたが、それらの多くは、タンパク質と他の分子（タンパク質、リガンド、核酸など）の間の相互作用（以後、タンパク質間相互作用と総称する）を介したものである。しかしながら、生化学的研究手法やX線結晶構造解析、NMR解析などによるタンパク質間相互作用解析には、一般に複雑な手順・高額な装置・研究者の熟練等が必要である。一方、バイオインフォマティクスの手法を用いた現在の相互作用予測・解析システムでは、詳細な機能の解析を行うのに十分な精度を得ることが難しく、とくにドッキング予測（複合体モデリング）では、精度の高い予測構造の候補を十分に生成できておらず、その計算速度も重要な課題となっている。**図1**は、本研究開発の概要を示したものである。我々は、以下の４つの項目について予測・解析手法の開発を行った。(1)タンパク質-タンパク質間相互作用予測、(2)タンパク質-タンパク質、タンパク質-リガンド間相互作用部位予測、(3)タンパク質-タンパク質ドッキング予測（タンパク質複合体構造予測）、(4)物理的なタンパク質間相互作用の解析。また、ドッキング予測と分子動力学シミュレーションを芳香環水酸化ジオキシゲナーゼ（ROS）のコンポーネント間の相互作用解析に適用し、それらの特異性と電子伝達能との関係を解析するとともに、電子伝達経路の解明を目指した。

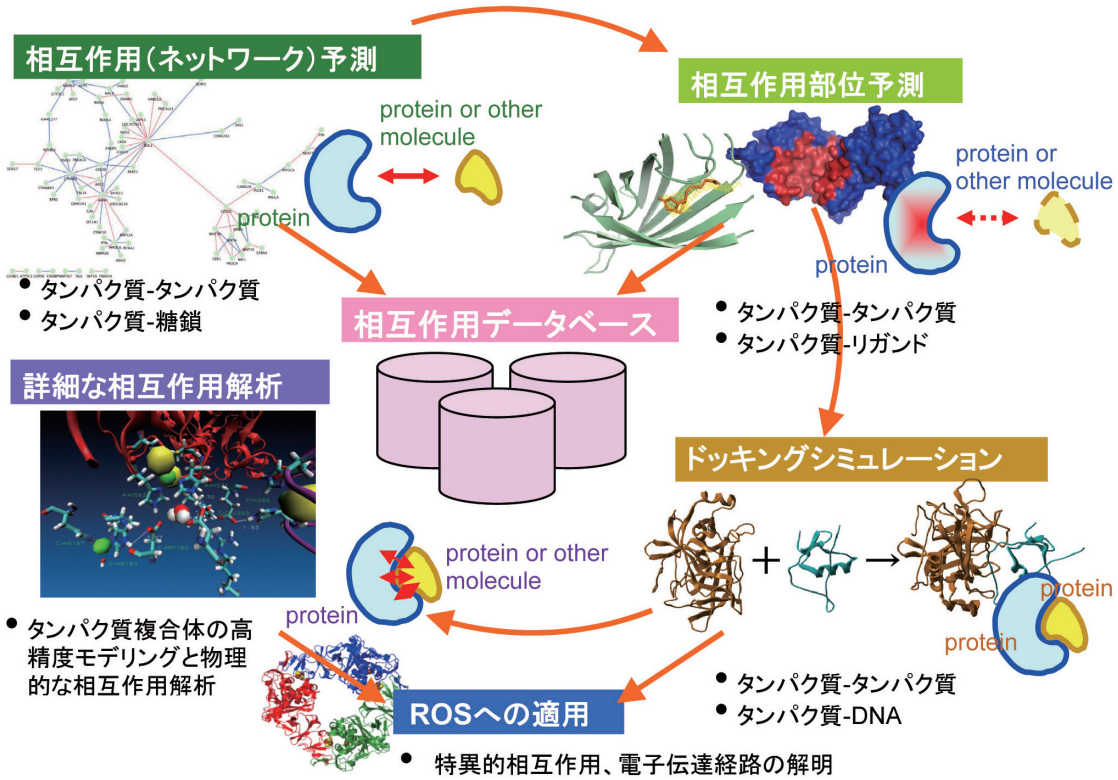

図１　研究の概要

2. 研究開発の成果

2.1 タンパク質-タンパク質相互作用予測

与えられた２つのタンパク質が相互作用するかどうかを、アミノ酸配列情報のみから機械学習法であるSupport Vector Machine（SVM）を用いて学習・予測する手法を開発した。SVMに与える配列特徴は、アミノ酸の隣接ペアの出現頻度（400×２次元）、独自のタンパク質ペアのカーネル関数を用いた。データセットとしてHPRD（Human Protein Interaction Database）を用い、5-fold cross validationで評価したときの結果は、AUC 0.885、MCC（Matthews correlation coefficient）0.625であり、Martinらの結果（AUC 0.862、MCC 0.581）、Shenらの結果（AUC 0.825、MCC 0.526）に比べて良い性能を示した。また、HPRDより取得したMAP kinaseのタンパク質間相互作用ネットワークを対象に予測を行った結果、372個の相互作用のうち347個（93.3%）を予測することができた。

2.2 タンパク質-タンパク質相互作用部位予測

タンパク質間相互作用部位予測は、アミノ酸配列の各残基が相互作用部位かどうかを予測するというものである。配列情報のみから予測する手法と、構造が既知の場合は、より高い精度で予測できるよう、配列情報と構造情報を利用して予測する手法の２つを開発した[1]。配列情報のみの予測では、PSI-BLASTにより類似の配列のマルチプルアラインメントを求め、残基ごとに配列上連続する11残基分のプロファイルを作成し、SVMの入力とする。配列情報と構造情報を利用した予測では、分子表面上、予測対象の残基と近接する15残基分のプロファイルを作成し、表面残基の極性・非極性原子の溶媒露出表面積と合わせてSVMの入力とする。また、どちらの予測手法も、SVMを２段

で適用する手法（１段目で得られた結果をさらに２段目のSVMの入力とする手法）を適用した。これは、相互作用部位が配列上連続していることを利用して、孤立して相互作用部位と予測された結果を修正することを意図したものである。

構造既知のタンパク質から抽出した563チェイン、104,331残基（31,816相互作用残基）をテストデータセットとし、5-fold cross validationで評価した結果、配列情報のみを用いた予測では、Precisionが30.0%のとき、Recall 62.9で、従来の手法（特徴量は残基の出現頻度、SVM 1段）の53.2より予測性能が高く、配列情報と構造情報を利用した予測では、Precisionが50.0%のとき、Recall 75.5で、従来の手法（特徴量は残基の出現頻度と残基単位の溶媒露出度、SVM １段）の66.2より予測性能が高いことを示した。

そのほか、我々は、SVR（Support Vector Regression）を用いて、各残基の周辺の相互作用残基数を予測する手法を新たに開発した。これは、注目している残基が相互作用部位のどの程度中心にあるかを予測することを意図したものである。予測に用いたデータセットは、配列一致度30%で冗長性を除いた168個のタンパク質からなるデータセットで、複合体の構造はPQS（Protein Quaternary Structure file sever）から取得した。予測結果を5-fold cross validationで評価した結果、全体の相関係数は0.59であった。さらに比較対象としてSVMを用いた予測も行った。その結果、SVRによる予測は、SVMによる予測と比較して、Recallは最大７%、Precisionは最大４%向上した。

2.3 タンパク質-リガンド相互作用部位予測

構造既知のタンパク質について、リガンドが結合する空間上の位置を予測する手法を開発した[2]。本手法は、タンパク質表面にメタン分子を格子状にプローブさせ、タンパク質分子とのvan der Waals相互作用エネルギーを計算するというものである。プローブの生成はDCLM（double cubic lattice method）により、力場パラメータとしてはAmber parm94を使用した。エネルギー値が小さいものをクラスタリングし、さらにそれをseedとして、より緩いエネルギー値の条件でクラスタを広げるという手法を用いた。35個のタンパク質-リガンド複合体（bound）構造と、35個の単体のタンパク質（unbound）構造からなるLaurieとJacksonのデータセットを使用し、予測を行った結果を**表１**に示す。

表１に示すように、PocketFinderやQ-siteFinderなど現在広く用いられている手法より高い精度で予測でき、とくにunbound 予測における予測精度の向上が大きいという結果を得ている。**図２**は、ストレプタビディン（PDB ID：2RTA）のリガンド結合部位予測の例を示したものである。予測順位１位の部位が黄色で、順位が下がるにつれ青色に近づく。黄色の部位が実際のリガンド（ビオチン）の位置と一致していることがわかる。

また、本研究開発に関連して、リガンド結合状態・非結合状態のタンパク質のデータベースBUDDY-systemを開発し、サービスを一般に公開している（http://www.bi.a.u-tokyo.ac.jp/services/buddy/current/index.cgi）。タンパク質、リガンドの単体からそれらの複合体、逆に複合体から単体を検索する機能をもち、リガンドに対する条件を細かく設定することが可能である。すでに16,203個の結合状態（bound）と非結合状態（unbound）の対を登録しており、現在は、結合状態と非結合状態の構造変化とダイナミクス、Biological Unitを考慮した結合残基と相互作用の詳細、リガンド周辺のmissing residueの解析結果などを登録した新しいデータベースを開発している。

表１　タンパク質-リガンド相互作用部位予測の結果

		１位の予測部位	３位以内の予測部位	平均 precision
提案手法	Bound	0.800	1.000	0.839
	Unbound	0.743	0.857	0.771
Q-SiteFinder	Bound	0.743	0.943	0.739
	Unbound	0.514	0.829	0.619
PocketFinder	Bound	0.714	0.771	0.375
	Unbound	0.514	0.657	0.354

Precisionは、予測部位と実際のリガンドの存在部位が一致している割合を示す。予測順位が１位のものと３位以内のものについては、precision ≥ 0.25であるものの割合を示す。

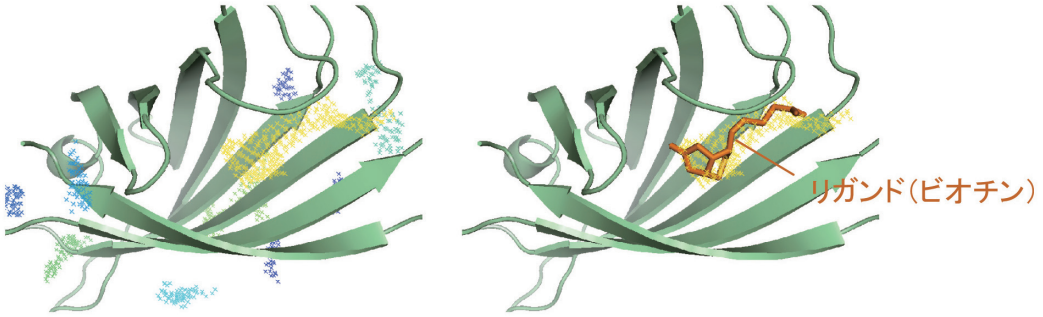

図２　ストレプタビディンのタンパク質-リガンド結合部位予測の結果

2.4 タンパク質-タンパク質ドッキング予測

ドッキングアルゴリズムとして、球面調和関数と新規に設計した正規直交基底関数での級数展開による高速内積計算を使ったアルゴリズムを開発した。ドッキングシミュレーションでは、タンパク質の相互作用を表すスコア関数を、各分子から定義されるスカラー場の関数f, gの内積の線形和

$$c(\mathrm{T}) = \sum_i w_i \int f_i(\boldsymbol{x}) g_i(\boldsymbol{x}) d\boldsymbol{x}$$

として表し、これを最小にする変換関数Tを求める。各コンフォメーションにおいてこのスコア関数の値を計算し、その値が低いものから順に、上位のものを候補コンフォメーションとする。各スカラー場は、その内積が、表現したいエネルギーもしくは性質を反映するように柔軟に定義でき、例えば、分子形状の相補性や各種ペアポテンシャル、静電相互作用などを表現することが可能である。また、本手法では、スカラー場を上記正規直交基底関数で展開することにより、スコア関数の計算に必要な内積計算を高速に行うとともに、配座空間の探索に必要な座標変換操作も高速に行えることを示した。本研究開発では、展開係数によるスカラー場の表現能力が、中心からの距離rの増加に従って劣化するという、球面調和関数に基づく基底関数を用いた方式の問題点を解決するため、とくに分子の表現空間を階層的に定義し、それぞれの階層において異なる動径基底関数を適用する手法を新たに開発した。これにより、比較的少数の係数でスカラー場を効率的に表現することができるようになった。

ドッキングによる相互作用エネルギーの評価には、原子レベルの統計ポテンシャルであるAtomic Contact Energy（ACE）をスコア関数に用いた。また、ACEには、立体障害の効果が入っていないため、原子の立体障害を表すポテンシャル関数を新たに導入した。これは、原子のvan der Waals半径内にあると正の値をとり、表面の原子に対する値は小さくするという工夫を取り入

れたもので、原子間の衝突を「ソフトに」避けるのに効果がある。現在、構造精密化の一環として、予測順位を調整する手法の開発を行っている。これは、van der Waals相互作用、クーロン相互作用、ACE経験的ポテンシャルの各項から構成されるタンパク質間相互作用ポテンシャルを導入し、デコイセットをもとに各項の最適な重み付けを行うというものであり、多くのタンパク質で予測順位を改善することができることを確認している。

表２は、本手法の予測精度と計算時間を、FFTを用いた手法で現在広く利用されているFTDockと比較して示したものである。本手法はFTDockと比較して、同程度の精度で予測するのに、16倍から160倍以上の高速化を達成した。

表２　タンパク質-タンパク質ドッキングシミュレーションの結果

複合体構造（単体構造）	予測順位（I-RMSD 値）		上位 4000 個中の予測数		計算時間 [分]	
	本手法	FTDock	本手法	FTDock	本手法	FTDock
1UGH（1AKZ+1UGI(A)）	1 (2.02)		18		0.61	101
1BRB（1BRA+6PTI）	78 (1.86)		47		1.76	29
2SIC（1SUP+3SSI）	58 (1.67)	NA	6	0	1.91	223
2PTC（3PTN+6PTI）	33 (2.30)	502	32	8	1.76	37
1CHO（5CHA(A)+2OVO）	1697 (2.24)	127	12	7	0.63	33
2KAI（2PKA(AB)+6PTI）	24 (2.04)	223	37	25	1.78	34

予測順位は、ネイティブに近い構造（ネイティブとのI-RMSD値が2.5Å以内の構造）が出現する予測順位を表し、上位4000個中の予測数は、スコア上位4000個の中のネイティブに近い構造の数を表す。

2.5 物理的なタンパク質間相互作用の解析

物理的な相互作用の解析については、分子動力学法（MD）を用いた構造精密化手法、溶媒効果の計算法、ab initio MDによる化学反応の動的解析手法などの基盤技術の開発[3][4][5][6]と、実際の系を対象にした解析研究[7][8]の２つのアプローチで研究を行っている。後者の例として、実験研究者と連携して、芳香環水酸化ジオキシゲナーゼ（Rieske non-heme iron oxygenase, ROS）のコンポーネント間相互作用・電子伝達機構の解明に関する研究を行った。３種の細菌由来の含窒素芳香族化合物カルバゾールに対する初発酸化酵素carbazole 1,9a-dioxygenase（CARDO, Pseudomonas / Janthinobacterium型［P/J型］、Sphingomonas型［S型］、Nocardioides型［N型］）は、全て末端酸化酵素、フェレドキシン、フェレドキシン還元酵素の３つのコンポーネントからなる。これらCARDOのコンポーネントの構造決定と互換性の解析を行い、特異性を明らかにした。さらに、ドッキングシミュレーションにより、各コンポーネント間の結合状態の推定を行い、正しい組み合わせでの酸化酵素-フェレドキシン間の結合には結合表面の形状の一致に加え、表面を構成するアミノ酸の電荷的な相補性が重要で、異なる組み合わせではその両方が合致しないため、結合せず電子伝達中心どうしも近づかないことが予想された。一方、正しい組み合わせのフェレドキシン－フェレドキシン還元酵素間では結合表面の形状の一致が見られるが、表面電荷の相補性はあまり認められなかった。このことは、疎水的相互作用を中心とする形状の一致がフェレドキシン－フェレドキシン還元酵素間の結合には重要で、両者の電子伝達中心が十分近づくことさえできれば電子伝達が起

こることを示唆している。

3. まとめ

本研究開発では、タンパク質間相互作用に関して、相互作用予測、相互作用部位予測、ドッキング予測、物理的相互作用解析の手法・ツールの開発を総合的に行った。今後は、各手法の予測精度のさらなる向上と、これらの有効な統合化を行っていきたいと考えている。また、開発した手法を用いたゲノムワイドな解析とタンパク質間相互作用に関するデータベースの開発も今後の重要なテーマである。CARDOの解析については、電子伝達時の相互作用の解明・体系化を推し進め、CARDOの電子伝達能の高い酵素の設計に役立てたいと考えている。CARDOは、原油成分、PCB、ダイオキシンなどの難分解性芳香族化合物の細菌による好気的分解系における重要な酵素であり、環境修復の研究にも役立つ可能性がある。

4. 研究開発実施体制

代表研究者　清水　謙多郎（東京大学大学院農学生命科学研究科）

研究開発題目

(1) 相互作用予測手法の開発

　グループリーダー　清水　謙多郎（東京大学大学院農学生命科学研究科）

(2) 芳香環水酸化ジオキシゲナーゼへの適用と実験による解析

　グループリーダー　野尻　秀昭（東京大学大学院農学生命科学研究科）

5. 参考文献

[1] M. Kakuta, S. Nakamura, K. Shimizu: Prediction of protein-protein interaction sites using only sequence information and using both sequence and structural information, *IPSJ Transactions on Bioinformatics*, **49**, 25-35 (2008).

[2] M. Morita, S. Nakamura, K. Shimizu: Highly accurate method for ligand-binding site prediction in unbound state (apo) protein structures PROTEINS: *Structure, Function, and Bioinformatics*, **73**, 2, 468-479, (2008).

[3] R. Jono, Y. Watanabe, K. Shimizu, T. Terada: Multicanonical ab inito QM/MM Molecular Dynamics Simulation of a Peptide in an Aqueous Environment, *Journal of Computational Chemistry*, accepted.

[4] T. Terada, K. Shimizu: A comparison of generalized Born methods in folding simulations. *Chemical Physics Letters*, **460**, 295-299 (2008).

[5] S. Yamasaki, T. Terada K. Shimizu, H. Kono, A. Sarai: A Generalized Conformational Energy Function of DNA Derived from Molecular Dynamics Simulations, *Nucleic Acids Research*, accepted.

[6] R. Ishitani, T. Terada, K. Shimizu: Refinement of comparative models of protein structure by using multicanonical molecular dynamics simulations. *Molecular Simulation*, **34**, 327-336 (2008).

[7] K. Inoue, Y. Ashikawa, T. Umeda, M. Abo, J. Katsuki, Y. Usami, H. Noguchi, Z. Fujimoto, T. Terada, H. Yamane, H. Nojiri: Specific interactions between the ferredoxin and terminal oxygenase components of a class IIB Rieske nonheme iron oxygenase, carbazole 1,9a-dioxyge-

nase, *Journal of Molecular Biology*, **392**, 436-451 (2009).

[8] H. Uchimura, T. Horisaki, T. Umeda, H. Noguchi, Y. Usami, L. Li, T. Terada, S. Nakamura, K. Shimizu, T. Takemura, H. Habe, K. Furihata, T. Omori, H. Yamane, H. Nojiri: Alteration of the substrate specificity of the angular dioxygenase Carbazole 1,9a-dioxygenase, *Bioscience, Biotechnology, and Biochemistry*, **72**, 3237-3248 (2008).