高精度のドッキング機能を有するタンパク質間相互作用予測システムの開発

●清水 謙多郎 [1)2)]　◆角越 和也 [1)]　◆寺田 透 [1)]　◆中村 周吾 [1)]
1) 東京大学大学院農学生命科学研究科 2) 東京大学大学院情報理工学系研究科

＜研究の目的と進め方＞

タンパク質間相互作用は様々な生命現象の鍵を握る重要な過程であるが、生化学的解析やX線結晶構造解析には複雑な手順・高額な装置・研究者の熟練等が必要であり、膨大な数の相互作用の全てをそれらの手法で解析するのは不可能である。

本研究では、これまで、生命システムの理解に、原子レベルの詳細なタンパク質間相互作用の解析からアプローチするため、高精度のタンパク質間相互作用部位予測およびタンパク質複合体ドッキングシミュレーション手法の開発を行い、さらに分子動力学法を用いて物理的な相互作用の解析を行う研究に取り組んできた。今後は、これらの手法の予測精度の向上をめざし、相互作用部位予測では、予測に有効な配列・構造特徴を明らかにするとともに、その傾向を分類・整理してデータベース化し、ドッキングシミュレーションにおいては、ドッキングのスコア関数の開発、より近似性能の高い基底関数の開発、候補構造から予測構造を絞り込むクラスタリング手法の開発、ならびに、構造変化に対応できるフレキシブルドッキングの手法の開発に取り組む。また、これらの手法を多数のタンパク質に適用し、結果を蓄積してデータベース化を図る。

本研究では、さらに、2つのタンパク質が相互作用するかどうかを配列のみから予測する相互作用予測システムを新たに開発し、相互作用ネットワークの推定を行う。現在、ハイスループットの実験手法により、系によってはタンパク質間ネットワークが網羅的に求められてきているが、インフォマティクスを用いた予測は、実験では検出が難しいものについても適用でき、ネットワーク情報からその相互作用の重要性を示すことができれば、実験研究者にとっての有用性はさらに増すと思われる。本件については、構造既知の場合は構造情報を利用し、類縁タンパク質の性質も積極的に利用して、精度の向上を図る。

＜2007年度の研究の当初計画＞

相互作用予測システムについては、アミノ酸配列のみを用いて、アミノ酸の組成、アミノ酸の双極子、体積などの値を入力とし機械学習サポートベクタマシン（SVM）で学習させる手法を開発する。とくに類縁タンパク質の性質を加味して、予測精度の向上を目指す。また、開発した手法を、実験によって明らかとなっているタンパク質間相互作用ネットワークに適用してその有効性を評価し、その上で多数の系に適用してネットワークの推定を試みる。

一方、昨年度より開発を行っている、タンパク質間相互作用部位予測については、タンパク質のアミノ酸配列情報のみを用いて予測する手法と、タンパク質のアミノ酸配列情報と構造情報の両方を用いて予測する手法の2つを開発しており、すでにプロトタイプが完成している。とくに後者については、タンパク質-タンパク質間相互作用残基を正しく予測できる割合（recall）を従来の手法の44.6%から76.3%に向上させた。本年度は、予測に有効な配列・構造特徴を明らかにするとともに、その傾向を分類・整理してデータベース化することを試みる。また、本年度は、新たに、タンパク質-低分子の結合部位（タンパク質のリガンド結合部位や酵素の基質結合部位）の予測手法の開発を行う。現在すでに手法の骨格部分の開発を進めており、小規模なテストデータに対して、Pocket FinderやQ-site Finder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るという結果を得ている。今後はさらに、予測に用いる相互作用エネルギー関数の改良、結合の伴う構造変化への対応など、いっそうの精度向上を図るとともに、開発した手法をシステムとして統合する。ドッキングシミュレーションについては、昨年度に引き続き、ドッキングのスコア関数の開発、より近似性能の高い基底関数の開発、候補構造から予測構造を絞り込むクラスタリング手法の開発を行い、また、情報科学的解析による側鎖モデリングと分子動力学計算を用いたさらなる側鎖モデリングの精密化および複合体形成による立体構造変化の予測を行う。

＜2007年度の成果＞

1. 相互作用予測システム

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サポートベクタマシン（SVM）を用いて学習・予測する手法を開発した。配列特徴としては、(1) アミノ酸の隣接ペアの出現頻度（400 × 2次元）、(2) (1) の3つ組の出現頻度（8000 × 2次元）、(3) アミノ酸の物理化学特性に基づく分類の隣接ペアの出現頻度（49 × 2次元）（具体的には、側鎖の双極子と体積をもとに7つに分類、側鎖の双極子は、密度汎関数理論B3LYP/6-31G*、GAUSSIAN03に組み込みの機能を使用して計算、側鎖の体積はSybyl6.8を使用）、(4) (3) の3つ組の出現頻度（343 × 2次元）の4通りを試したところ、(2) が最も良い結果を示した。タンパク質ペアのカーネル関数Kは、K'を個々のタンパク質に対して適用されるカーネル関数とすると、

$$K((i,j),(k,l)) = K'(i,k)K'(j,l) + K'(i,l)K'(j,k)$$

であり、RBFカーネルを適用した。データセットしてHPRD（Human Protein Interaction Database）を用い、5-fold cross validationで評価したときの結果は、sensitivity 0.767、precision 0.841、accuracy 0.811、MCC（Matthews correlation coefficient） 0.625であり、Shenらの方法（2007）の結果sensitivity 0.674、precision 0.684、accuracy 0.681、MCC 0.363より高い予測性能を示した。

ネットワークの推定については、HPRDより取得したMAP kinase（Ras-Raf1-MEK1-ERK1-ELK1-SRF）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. 相互作用部位予測システム

タンパク質間相互作用部位予測については、昨年度に引き続き、タンパク質のアミノ酸配列情報のみを用いて予測する手法と、タンパク質のアミノ酸配列情報と構造情報の両方を用いて予測する手法の2つを開発している。本年度はとくに後者について、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あたりにあるか（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。その手順

は以下の通りである。ターゲットのタンパク質の配列に対し、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、プロファイルを作成する。それと合わせて、各表面残基の極性・非極性原子の溶媒露出表面積を計算し、残基ごとに、空間的に隣接する 14 残基を加えた 15 残基分のプロファイルと溶媒露出表面積を取り出して機械学習 SVR の入力とする。

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 で評価したところ、是全体の相関係数は 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。

本年度は、また、タンパク質 - リガンド結合部位予測の手法を開発した。これは、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ている。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 and 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 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
我々の手法	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
Pocket-Finder	Bound	0.714	0.771	0.375
	Unbound	0.514	0.657	0.354

予測順位が 1 位のものと、3 位以内のものについて、precision ≧ 0.25（25% 以上、予測部位と実際のリガンドの存在部位が重なっているものの割合）の結果を示す。

表 1 に示すように、Pocket Finder や Q-site Finder 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。今後はさらに、予測に用いる相互作用エネルギー関数の改良、結合の伴う構造変化への対応など、いっそうの精度向上を図るとともに、開発した手法をシステムとして統合する。また、本研究に関連して、タンパク質 - リガンドの複合体構造と単体構造のペアのデータベースを構築しており、すでに公開している。

3. ドッキングシステム

ドッキングシミュレーションについては、昨年度に引き続き、ドッキングのスコア関数の開発、より近似性能の高い基底関数の開発、候補構造から予測構造を絞り込むクラスタリング手法の開発を行っている。ドッキングシミュレーションにおいては、これまで、原点から離れた場所での近似性能を改善するため、実空間を複数のレイヤーに分割し、レイヤーごとに基底関数のセットを用意してスカラー場を展開してきたが、動径基底関数として用いている現在の Legendre 多項式では、とくに原点に近い部分での極値の位置の分布に偏りがあるため、本年度は、新たな直交関数系を導入し、予測精度の改善を試みた。また、高精度の相互作用解析を行うため、*ab initio* 分子動力学（MD）とマルチカノニカル MD を統合した手法を新たに開発し、相互作用部位における化学反応の動的な解析を可能にする基盤技術を開発した。

＜国内外での成果の位置づけ＞

相互作用予測については、配列情報のみを用いた手法としては、従来の主要な手法に比べて高い予測性能を達成しており、さらにドメイン情報を用いた予測手法を合わせて用いることにより、さらなる予測精度の向上が期待される。SVR 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は、これまでにない新しい手法であり、相互作用部位予測の精度向上だけでなく、その成果は、ホットスポットとの関係など、相互作用部位の性質を解析する上で重要と考えられる。タンパク質 - リガンド結合部位予測手法の開発では、Pocket Finder や Q-site Finder 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ることを実証しており、とくに unbound 予測における予測精度の向上が大きいことを示した。ドッキングによる複合体構造予測については、FTDock など、FFT を用いた手法がよく利用されているが、我々の手法は、それらに対し 160 倍から 1700 倍の性能向上を達成している。

＜達成できなかったこと、予想外の困難、その理由＞

ほぼ計画通りの成果が得られており、タンパク質相互作用予測（ネットワーク予測）、タンパク質 - リガンド結合部位予測では、当初の計画以上の成果が得られた。一方、タンパク質間相互作用部位予測では、予測精度の向上に重点を置く余り、相互作用部位の傾向を分類・整理してデータベース化するところまでは至っていない。今後、早急に取り組みたいと考えている。

＜今後の課題＞

各手法の予測精度の向上が重要な課題である。とくに、ドッキングについては、高速性を生かしたアンサンブルドッキングなど、複合体形成時の構造変化に対応できる手法の開発に取り組みたいと考えている。また、ネットワーク予測、相互作用部位のデータベース化などゲノムワイドな研究に力を入れていきたい。多数の対象に対する網羅的ドッキング、物理的な相互作用、ダイナミクスも含めた結果のデータベース化も重要な課題である。

＜成果公表リスト＞

1) 論文／プロシーディング（査読付きのものに限る）

1. 0801161647
M. Hirano, R. S. Davis, W. D. Fine, S. Nakamura, K. Shimizu, H. Yagi, K. Kato, R. P. Stephan, M. D. Cooper: IgEb immune complexes activate macrophages through Fc-gamma RIV binding, Nature Immu., 8, 762-771 (2007).
2. 0801161657
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, accepted.
3. 0702131217
S. Yamazaki, et al.: Mechanism of the difference in the binding affinity of *E.coli* $tRNA^{Gln}$ to glutaminyl-tRNA synthetase caused by non-interface nucleotides in variable loop, Biophysical Journal, 92, 192-200 (2007).
4. 0801231659
R. Ishitani, T. Terada, K. Shimizu: Refinement of comparative models of protein structure by using multicanonical molecular dynamics simulations, Molecular Simulation, accepted.

2) データベース／ソフトウェア

1. 0801171140
タンパク質 - リガンド結合状態／非結合状態ペアデータセット
http://www.bi.a.u-tokyo.ac.jp/services/buddy/current/