公募研究：生命システム情報

タンパク質相互作用のネットワーク予測から原子レベルの結合予測までの統合的研究

●清水 謙多郎 1)2)　◆角越 和也 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反応であるが、生化学的解析や立体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、タンパク質間相互作用について、以下の 4 つのアプローチから予測・解析手法の開発を行う。(1) タンパク質 - タンパク質間相互作用予測および相互作用ネットワーク予測、(2) タンパク質 - タンパク質、リガンド、糖鎖、DNA 間相互作用部位予測、(3) タンパク質 - タンパク質間ドッキング予測（複合体構造予測）、(4) タンパク質 - タンパク質、リガンド、糖鎖、DNA 間の物理的な相互作用解析。(1) については、タンパク質の配列、進化解析情報、GO（Genome Ontology）情報を総合的に適用し、さらにネットワーク構造の既知部分や他の類似のネットワークを鋳型として予測を行う手法を合わせて開発する。(2) については、配列・構造特徴の抽出法、カーネル関数の改良、SVR（Support Vector Regression）の導入などにより、精度の向上を図る。(3) については、結合時の構造変化を加味したフレキシブルドッキングと構造精密化のため手法を開発する。(4) については、溶媒効果を採り入れたマルチカノニカル分子動力学シミュレーションにより、原子レベルの詳細な解析を行えるようにする。以上、(1) ～ (4) を実現する統合化されたシステムを構築する。本研究では、タンパク質間相互作用ネットワーク予測や相互作用部位予測をゲノムワイドに適用し、また、従来個別に行われてきた物理化学的な相互作用の解析とデータの蓄積を通して、生命システムの理解に原子レベルからアプローチすることをめざす。

＜ 2008 年度の研究の当初計画＞

タンパク質 - タンパク質間相互作用予測および相互作用ネットワーク予測については、アミノ酸配列のみを用いて、機械学習 SVM（Support Vector Machine）で学習させる手法をすでに開発しているが、本年度はさらなる精度の向上をめざし、以下のことを行う。(a) 学習させる配列特徴の工夫、カーネル関数の設計を行う。(b) 相互作用する互いのタンパク質の配列類似性、進化情報を用いた予測手法を開発する。(c) GO 情報を用いた予測手法を開発する。(d) (a) ～ (c) を総合的に適用した予測判定の手法を検討する。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、その上で多くの系に適用してネットワークの推定を試みる。

タンパク質 - タンパク質間相互作用部位予測については、アミノ酸配列情報のみから予測する手法と構造情報も利用して予測する手法の 2 つを開発し、従来よりも高い精度を実現している。本年度は、とくに、SVR を導入し、各相互作用残基の周辺残基数を予測する手法を開発する。これにより、相互作用部位の詳細な構成（hot spot など）や特性を予測・解析する。進化情報を利用した手法をあわせて実現し、両者を併用した場合の効果を実際のタンパク質に適用して評価する。これと並行して、タンパク質 - リガンド、糖の結合部位予測手法を開発する。

タンパク質 - タンパク質間ドッキング予測については、現在開発中のシステムに効果的な相互作用スコア関数を導入する。現在、Atomic Contact Energy という統計的な関数を主に用いているが、その他にも疎水性相互作用、クーロン相互作用などの項も個別に求めて計算実験することにより、それらをスコア関数に取り込むことの効果を検討し、スコア関数の最小化によってよりネイティブに近い複合体構造が得られるようにする。また、類似の複合体を形成するタンパク質から、それらの構造特徴を表す経験的ポテンシャルを抽出し、ドッキングのスコア関数の精密化を試みる。

タンパク質 - タンパク質、DNA 間の物理的な相互作用解析については、マルチカノニカル分子動力学法を用い、溶媒分子の存在下での精密なモデリングおよび詳細な相互作用の解析を行う。

＜ 2008 年度の成果＞

1. タンパク質 - タンパク質相互作用予測システム

昨年度に引き続き、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。

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、Martin（2005）、Shen（2007）らの方法に比べて高い正答率を示した。また、ネットワーク予測については、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. 糖鎖結合タンパク質予測

アミノ酸配列情報のみを用いて、ゲノムワイドな解析にも適用できるレクチンを予測するシステムを開発した。

まず、本研究で扱うレクチンを明確に再定義した。これは、従来のレクチンの定義にはあいまいなところがあるからである。そこで我々は、糖鎖を直接修飾しない糖鎖結合性タンパク質を一律にレクチンとして扱うことにし、これらのタンパク質をデータベースから抽出する際の検索条件の定式化を図った。

具体的には、UniProt Knowledgebase からアノテーション情報をもとに、検索ツール SRS（Sequence Retrieval System）を用いて、検索条件に合致するレクチンのアミノ酸配列を収集した。さらに、レクチンの配列特徴を効果的に学習させるため、これらのアミノ酸配列に対し、BLAST によるクラスタリングを行い、配列冗長性を排除したデータセット（正例データセット）を作成した。一方、非レクチンのデータセット（負例データセット）としては、実際に発現が確認されているタンパク質の中から、レクチンの検索条件に合致しないものをランダムに収集し、上と同様にして冗長性を排除したものを用いた。

次に、これらのデータセットを構成するアミノ酸配列から特徴ベクトルを作成し、正例／負例の情報とともにSVMに学習させた。配列情報を特徴空間上に写像させるカーネル関数としては、3-spectrum kernelを用いた。その結果、AUCの値は0.80を超え、比較的高い予測精度が得られることがわかった。また、次元削減のため、アミノ酸のグルーピングも試みた。グルーピングは、タンパク質の2次構造と極性に基づくLevittらの6分類、さらにCysteineを区別した筆者らによる7分類、またBlosum50スコアの相関に基づくLauneyらの6分類の3種類を適用し、これに、20種類のアミノ酸を直接符号化した手法（グルーピングを行わない）を合わせて、性能評価を行った。予測精度については、グルーピングしない場合の性能を上回る結果は得られなかったものの、筆者らの7分類が最も高い予測精度を実現し、その値はグルーピングしない場合の性能とほぼ同等であった（計算速度は10倍以上高速である）。

また、ゲノムワイドの予測として、ヒト遺伝子統合データベースH-invDBに開発した手法を、BLASTと組み合わせて適用したところ、適切なしきい値を設定すれば、両主法の組み合わせにより、BLAST単独よりも高精度の予測が行えることを確認した。BLASTとSVMで両方レクチンと判定された配列のグループは、他のものに比べ、糖鎖結合と関係あるキーワードが、アノテーション情報中に記述されている配列がより高い頻度で存在した。

3. タンパク質 - リガンド結合部位予測システム

筆者らが開発したタンパク質 - リガンド結合部位予測システムは、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し、エネルギー値の小さい部位をリガンド結合部位と予測するというものである。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、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た。現在、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について検討を行い、とくに後者については、構造変化を分類するデータベースを作成中である。その他、本研究に関連して、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を作成し、一般に公開している。

4. ドッキング予測と相互作用解析

タンパク質 - タンパク質のドッキングについては、動径基底関数を改良と、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する枠組みを利用したパラメータの調整により、約7割のケースで予測精度を改善することができた。また、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最適な重み付けを行い、候補構造を選択する手法を開発した。

そのほか、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。糖鎖結合タンパク質予測については、筆者の知る限り利用できるものがない。糖鎖結合部位を予測する方法はいくつか発表されており、また多くの研究者がレクチンと糖鎖の3次元構造情報に基づきドッキングシミュレーションを行なってきたが、これらの結合部位予測では結合しないことを予測する能力が明らかに十分でない。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、糖鎖結合タンパク質予測（レクチン予測）では、当初の計画以上の成果が得られた。ドッキング予測については、候補構造の選択を行う手法を開発し、一定の効果を挙げたが、今後はさらに広範な精度向上をめざし、早急にポテンシャル関数の改良を行い、またフレキシブルドッキングと合わせて実現するなどの工夫を試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

1. 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, 49, 25-35 (2008)
2. 0805081419
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, accepted
3. 0801231659
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).
4. 0901071711
W. Cao, K. Sumikoshi, T. Terada, S. Nakamura, K. Kitamoto, K. Shimizu: Computational Protocol for Screening GPI-anchored Proteins, Proceedings of the First International Conference on Bioinformatics and Computational Biology (BICoB), Springer Lecture Notes in Bioinformatics Series, 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/

九州工業大学の皿井明倫教授と、タンパク質 -DNAの複合体形成のメカニズムの解明に関する共同研究を行っている。