ゲノムスケールでの転写制御ネットワークの解析

●皿井 明倫
九州工業大学・情報工学部

＜研究の目的と進め方＞

遺伝子発現制御は最も重要な生物機能のひとつでありそれは膨大な転写因子とそのターゲット遺伝子の複雑なネットワークで実現される。本研究では、これまでに開発してきた転写因子とターゲット予測法を組み合わせることにより、転写制御ネットワークをゲノムスケールで解析するストラテジーを確立したい。まず、アミノ酸配列情報、進化情報、および構造情報を組み合わせてDNA結合およびRNA結合蛋白質を精度よく予測する。また、配列情報、構造情報および計算機シミュレーションを組み合わせて、転写因子のターゲットをゲノムスケールで精度よく予測する。これらの予測結果は、実験による既知データとともにすべてデータベースに統合し、これらの転写因子とターゲットをゲノムスケールで網羅的に解析することにより、転写制御ネットワークの階層性やモジュール構造などを明らかにしたい。

＜2008年度の研究の当初計画＞

本研究ではまず、これまでに開発してきた転写因子とターゲットの予測法を組み合わせることにより予測精度の向上を計る。これまでの解析から、複数の方法の組み合わせで予測精度を上げられることがわかっており、これらの統合した予測を自動的に行えるようにする。転写因子の予測では、これまでに開発した配列組成やコンテクスト情報、アラインメントを利用した進化情報、および構造を利用した予測法を組み合わせる。そして、最も精度を上げるような組み合わせを探求する。転写因子のターゲット予測については、既存の配列情報を利用した方法や、構造情報および計算機シミュレーションを用いて蛋白質・DNAの直接認識と間接認識の方法を組み合わせる。これらの統合した予測をゲノムスケールで自動的に行うためのツールを作成しWeb上で公開する。一方、これらの開発にとって重要となる蛋白質・核酸相互作用に関する構造データや実験データを集めたデータベースの開発をすすめる。

ゲノムスケールで転写制御ネットワークを解析するため、まず酵母ゲノムの転写系について、網羅的に予測された転写因子やターゲット部位などの結果を既知の実験データと統合した統合データベースを構築する。次に、この統合データベースを網羅的に解析することにより、転写因子とターゲット、その産物のネットワークを自動的に構築する。この結果を検証するため、酵母の細胞周期にかかわる転写制御系などについて、遺伝子発現データなど既知の情報から得られたネットワークと比較する。

＜2008年度の成果＞

転写因子の予測では、これまでの方法では結合するDNA配列の情報は考慮してこなかったが、蛋白質アミノ酸が結合する単独塩基および隣り合う2塩基のコンテクストを考慮してアミノ酸と塩基の相互作用を統計的に解析し、その結果を予測に取り入れた。DNA結合そのものの予測精度はほとんど向上しないが、結合DNAの配列コンテクストについてより詳しい情報を得ることができた。

蛋白質とDNAの特異的認識のメカニズムを明らかにするには、結合インターフェイスの特徴を明らかにすることが重要である。そこで、蛋白質とDNAの相互作用インターフェイスの配列、構造と熱力学的性質の関係を詳しく解析したところ、最も相互作用の安定化に寄与する残基は保存されたパッキング密度の高いクラスター領域に多くあることなどの知見が得られた。

転写因子のターゲット予測については、DNAの構造や物性をとおして配列を認識する間接認識について詳しい解析を行った。これまでに、DNAの配列に依存したコンフォメーションエネルギーを評価するためにDNAの計算機シミュレーションを行い、すべての4塩基の組み合わせ配列を含むDNA（136種類）の10nsの分子動力学計算を行った。これまでは、そのトラジェクトリーからすべての塩基ステップのパラメータの分布を調べ、共分散行列から調和近似を用いて平均場ポテンシャルという量を計算した。しかし、パラメータの分布を調べてみると、極値が複数ある場合など必ずしもガウス分布しているとは限らず、調和近似には限界がある。そこで、パラメータの分布そのものの情報を用いて直接に平均場ポテンシャルを計算する方法を開発した。これを用いて、まずDNAの構造について配列と構造の適合性をスレッディングという方法で評価したところ、従来の方法よりも高い特異性（Z-score）が得られた。この方法により、予測の精度を上げることができると期待される。

間接認識は、ヌクレオソームのポジショニングにも重要と思われる。ヌクレオソーム形成はゲノムスケールでの転写の制御に深く関わっていると考えられている。そこで、我々の平均場ポテンシャルを用いてヌクレオソーム複合体構造の配列と構造のスレッディングを行ったところ、ヌクレオソーム配列は複合体構造に特異性を示した。このことは、ヌクレオソーム形成において、DNAの配列に依存したコンフォメーションや曲がりやすさが重要な役割を果たしていることを示唆する。さらに特異性のメカニズムを詳しく解析するため、すべての塩基ステップ（2塩基配列）について配列・構造スレッディングを行ったところ、ポテンシャルエネルギーはDNAのピッチで周期的に変動することがわかった。フーリエ解析により周期性の程度を計算すると、特定の塩基ステップが周期性に寄与していることがわかった。現在、これらの結果を用いて、ヌクレオソームのポジショニングの予測を行っている。

一方、アミノ酸と塩基に直接相互作用による直接認識について、蛋白質・DNA複合体の構造情報を用いた予測法をChIP-chipデータと組み合わせて評価を行った。ChIP-chipデータそのものにはまだ実験的な不確定性があるが、直接認識の統計ポテンシャルを組み合わせると特異性が上がる場合のあることがわかった。これは、ChIP-chipデータ単独よりも精度よくターゲットを同定できることを示唆している。現在、酵母ゲノムを用いて、細胞周期にかかわる転写因子などについてパイロット研究をすすめている。

また、我々はこれまでに蛋白質とDNAの相互作用の実験データを用いて結合サイトを予測する方法を開発したが、今回この方法をシアノバクテリアのゲノムに応用し、転写因子SYCRP1についてゲノムスケールでターゲット部位と遺伝子を予測した。予

測結果を検証するため、予測された結合配列と転写因子との結合実験を行ったところ、ほとんどの予測結合サイトにSYCRP1が実際に結合することが確かめられた。

次に、転写制御ネットワークを解析するため、実験的によく調べられている酵母ゲノムの転写系について、ターゲット遺伝子のプロモータ上に結合する転写因子のコンテクストを解析し、その組成、転写因子間の順序や距離などの情報をもとにクラスタリングを行い機能との相関を調べている。また、酵母ゲノムの転写系について統合データベースを作成し、これらの情報を統合している。そして、転写因子とターゲットについてネットワークを自動的に表示するツールを作成した。

一方、これらの研究の基盤となるデータベースや解析ツールの開発も行った。転写因子やターゲット予測の研究を支援するため、蛋白質・核酸相互作用熱力学データベース、蛋白質・核酸複合体構造データベースなどを開発し公開しているが、これらのデータ更新や機能強化を引き続き行った（データベース／ソフトウェアリスト参照)。一方、構造情報に基づいて蛋白質とDNAの直接認識と間接認識の特異性を計算するWebサーバ、ReadOUT、を公開しているが、蛋白質・DNA複合体データの更新にともない、統計ポテンシャルの更新を行った。また、構造空間（ストラクチュローム）における分子相互作用ネットワークの情報を俯瞰し、検索・可視化するためのツール、PDBnet、を開発し公開しているが、データの更新と機能強化を行った。これを用いて、転写因子によるターゲット認識の協同性の解析もすすめている。一方、転写制御研究者を支援するため、転写制御に関する各種データベースや解析ツールなどを集めた転写制御ポータルサイトを作成し公開しているが、この更新も行った（データベース／ソフトウェアリスト参照)。

＜国内外での成果の位置づけ＞

これまでに、配列情報に基づいて転写因子のターゲット予測が行われているが、精度の点で大きな問題がある。また、最近は遺伝子発現データやChIP-Chipデータなどが増加し、これらのデータから転写制御ネットワークが推定されている。しかし、これらの実験データの精度や推定方法にも問題がある。したがって、今後は予測や実験の精度をさらに向上させる必要がある。我々はこれまでに、転写因子とそのターゲットを予測する方法をいくつか開発してきた。そして、複数の方法を組み合わせることにより、単独の情報を用いる方法よりも精度を上げることができることを示してきた。そこで本研究では、これをさらに発展させ、精度を上げた方法により、ゲノムスケールで転写因子とそのターゲットを予測し、このデータを既知の実験データなどと組み合わせて解析することにより、転写制御ネットワークを構築するというストラテジーを確立しようとしている。本研究で開発された新しい情報解析技術や統合データベースは、ゲノム機能解析に大きく貢献するであろう。

＜達成できなかったこと、予想外の困難、その理由＞

これまでのところ､研究はほぼ計画どおりすすんでいる。ただ、ゲノムスケールでの膨大な情報の網羅的な解析や、それを支援するデータベースや解析ツールの開発には、多くの資金や人的資源が必要であり、現在の体制ではなかなか望むようなスピードではすすまない。

＜今後の課題＞

これまでの研究をもとに得られたさまざまなノウハウを予測法のさらなる改良にフィードバックする。予測結果が既知データと明らかに異なる場合は、その原因を調べて予測法の改良にフィードバックする必要がある。そして、精度を上げた予測結果をさらに比較するというサイクルを繰り返す。いまだに完成されていない予測法に関しては必要に応じて更なる改良を続ける。特に、予測の基礎となる蛋白質・DNA認識のメカニズムなどについては、さらに詳しい研究を行う必要がある。予測された転写制御ネットワークについては、既知の知見と詳しく比較検討し、そこから得られる遺伝子発現のメカニズムの妥当性を細胞周期などの具体的な例について詳細に検討する。また、プロモータ・遺伝子レベルでの転写因子の協同性やコンテクスト、ネットワークレベルでの転写制御の階層構造や因果関係などについて解析をすすめたい。

＜成果公表リスト＞

1) 論文／プロシーディング

1. 0801312048
S. Ahmad, Y. H. Singh, M. J. Araúzo-Bravo, A. Sarai "Sequence-based prediction of residue-level properties in proteins" in *Machine Learning in Bioinformatics,* eds, Y.-Q. Zhang and J.C. Rajapakse, John Wiley & Sons, (2007).
2. 0901151706
K. Omagari, H. Yoshimura, T. Suzuki, M. Takano, M. Ohmori, and A. Sarai "ΔG-based prediction and experimental confirmation of SYCRP1-binding sites on *Synechocystis* genome" **FEBS J.** 275, 4786-4795 (2008).
3. 0901151709
Alex V. Kochetov, Shandar Ahmad, Vladimir I. Ivanisenko, Nikolay A. Kolchanovand Akinori Sarai "uORFs, reinitiation and alternative translation start sites in human mRNAs" **FEBS Lett.** 582, 1293-1297 (2008).
4. 0901151720
S. Ahmad, Y. H. Singh, M. J. Araúzo-Bravo, A. Sarai "Sequence- based prediction of residue-level properties in proteins" in Machine Learning in Bioinformatics, eds, Y.-Q. Zhang and J.C. Rajapakse, pp. 157-187 John Wiley & Sons, (2008).

2) データベース／ソフトウェア

1. 0507070127
蛋白質・核酸相互作用データベース、ProNIT
http://dna01.bse.kyutech.ac.jp/jouhou/pronit/pronit.html
2. 0702132136
蛋白質・DNA認識特異性の予測サーバReadOut:
http://gibk26.bse.kyutech.ac.jp/jouhou/readout/
3. 0702132139
ストラクチュロームでの分子ネットワーク解析ツールPDBnet:
http://gibk21.bse.kyutech.ac.jp/pdbnet/
4. 0702132142
転写制御ポータル
http://gibk21.bse.kyutech.ac.jp/tfportal/TRP/TRP-j.html

本研究では､蛋白質とDNAの相互作用の計算機シミュレーションに関して、東京大学農学生命の清水謙多郎研究室と共同研究を行っている。