ゲノムスケールの転写因子とターゲット予測

●皿井 明倫
九州工業大学・情報工学部

＜研究の目的と進め方＞

近年多くの生物種の全ゲノム配列が次々と明らかにされ、ゲノムスケールでの生物機能の体系的解析が可能となりつつある。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用いてゲノムスケールで予測を行うことにより、遺伝子発現制御メカニズムを解明する新たなストラテジーを確立することをめざす。まず、酵母をパイロット研究の対象として、機能未知の遺伝子から転写因子を予測し、さらにそのターゲット遺伝子をゲノムスケールで予測する。そして、既知データとの比較のサイクルを繰り返して予測方法の改良を行う。これらの結果はライブラリー化し他の情報とデータベースに統合し予測サーバとともに公開する。本研究で開発した新しい情報解析技術や統合データベースを用いて、ゲノムの機能解析に貢献したい。

＜ 2007 年度の研究の当初計画＞

本年度は、以下の項目について解析を行う。

（1）転写因子予測：これまでに、DNA に結合する蛋白質を予測するために、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、および構造情報を利用する方法を開発してきたが、これらの予測精度をあげるための改良を行う。構造情報を利用する方法では、転写因子と DNA の複合体の構造データを更新する。また、蛋白質自体の構造がわかっていない場合でも、配列からホモロジーに基づき構造を予測してから DNA 結合活性を予測するという方法を確立する。一方、蛋白質の全配列を用いた場合とドメインごとの配列と構造を用いた場合で予測精度に差がでるかどうか検証する。これらの予測を酵母のすべての ORF について網羅的に行い、方法論の検証を行う。

（2）転写因子のターゲット予測：蛋白質・DNA 複合体の構造データを更新し、アミノ酸と塩基の直接相互作用による直接認識と DNA の構造や物性をとおした間接認識について、それぞれの統計ポテンシャルを更新する。この統計ポテンシャルを用いて、蛋白質・DNA 認識の特異性や相互作用エネルギーを計算する。間接認識に関しては、既知の複合体構造データに基づく経験的な方法を補完するため、計算機シミュレーションにより DNA の配列に依存したコンフォメーションエネルギーを計算する。これらの方法を組み合わせて、転写因子のターゲット予測を酵母ゲノムについて網羅的に行い、ChIP-chip などの実験データと比較する。

（3）データベース・ツールの開発と公開：これまでに開発してきた、蛋白質・核酸相互作用熱力学データベース、ProNIT、酵母の転写制御データベース、ポータルサイト、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、などのデータベースや解析ツールの更新や機能強化をさらに推進する。

＜ 2007 年度の成果＞

本年度は、次の課題に関して以下のような成果を得た。

（1）転写因子の予測

これまでに、蛋白質が DNA に結合するかどうかを予測するため、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、および構造情報を利用する方法を開発してきた。これらの方法ではトレーニングのためのデータが必要であるが、本年度は、DNA 結合蛋白質のデータセットの追加を行った。これらのデータは、Swiss-Prot (UniProt) および Gene Ontology のアノテーションなどを用いて収集した。DNA 結合ドメインの情報については、Pfam および CATH を用いた。上記の方法を用いて、DNA 結合蛋白質全体、DNA 結合ドメイン、および、非 DNA 結合蛋白質について計算を行った。配列情報を用いる方法では、アミノ酸のコンポジションベクトルを用いて、DNA 結合蛋白質全体と DNA 結合ドメインとの比較を行った。その結果、蛋白質の全領域の配列よりも、DNA 結合ドメインのみを用いたほうが予測確率が高いという傾向が得られた。また、構造情報がある場合あるいは配列のホモロジーから構造モデルが構築できるものについては、構造モデルに基づく DNA 結合蛋白質の予測も行い、予測精度が上がることが確かめられた。一方、DNA 結合蛋白質をファミリーごとに分けて予測を行う解析や、RNA 結合蛋白質の予測は現在解析中である。

現在、パイロット研究として酵母ゲノムを用いてゲノムスケールでの予測の検証を行っている。ORF の中にはまだ機能未知の遺伝子が多くあるが、上記の方法で DNA 結合蛋白質と予測された機能未知の遺伝子についてさらに詳しい機能解析をすすめている。

（2）転写因子のターゲット予測

蛋白質・DNA 複合体の構造情報に基づくターゲット予測法については、予測精度を向上させるため、蛋白質・DNA 複合体の構造データベースの更新を引き続き行った。このデータを用いて、アミノ酸と塩基に直接相互作用による直接認識と DNA の構造や物性をとおして配列を認識する間接認識のそれぞれの統計ポテンシャルを更新した。これらの統計ポテンシャルを用いて、蛋白質・DNA 認識の特異性の計算などを行うとともに、ターゲットの予測を行った。いくつかの蛋白質・DNA 複合体については、直接認識と間接認識を組み合わせることで特異性が上昇し予測精度が上がることが示された。また、従来の配列に基づく予測法に構造情報に基づく方法を組み合わせることで、配列に基づく予測では見逃された結合配列が予測できることがわかった。

間接認識については、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も行った。この計算では、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い、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について、統計ポテンシャルに相当する平均場ポテンシャルという量を計算した。これまでの結果から、計算機シミュレーションによって

DNA コンフォメーションエネルギーの配列依存性と特異性を予測できることがわかった。また､間接認識においては､DNA のバックボーン構造が塩基間のステップパラメータと相関して変化することを示唆する結果が得られた。このことは､塩基のコンフォメーションだけでなく、それに伴って変化するバックボーンのコンフォメーションも、蛋白質による DNA 配列の認識にとって重要な役割を果たすことを示唆する。

一方、パイロット研究として酵母ゲノムを用いてゲノムスケールでのターゲット予測を行っている。これまでに、すべてのプロモータ領域について予測された結合サイトを後述する酵母の転写制御データベースに統合した。また、予測された結合部位と結合強度をグラフとして視覚化できるようにした。これらの情報を、遺伝子の制御領域の機能情報、実験的にすでに知られている結合配列、ターゲット遺伝子の産物などの情報などと合わせて解析することにより、転写制御のネットワークを自動的に生成することを試みている。また､このネットワークを可視化するインターフェイスも作成している。また現在、予測された結果の精度を検証するため、細胞周期にかかわる転写因子を中心に、CHIP-Chip データなどと系統的に比較している。これまでの比較ではおおむねよく一致している。

（3）データベース・ツールの開発と公開

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 (http://gibk26.bse. kyutech.ac.jp/ jouhou/jouhoubank.html) 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。これにより、予測の精度が向上した。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＞

現在、転写因子とターゲット予測の方法はいろいろあるが、まだゲノムスケールで精度よく予測を行うことはできていない。現在、マイクロアレイデータや CHIP-Chip などの相互作用データが急速に増加しており、遺伝子発現ネットワークの解析がさかんになってきた。ただ、これまでの研究の多くは、特別の情報や方法に頼った解析や予測を行っており、精度の点で限界がある。本研究では、できるだけ多くの情報や方法を用いてゲノムスケールでの解析に応用しようとしている。

＜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思うようなスピードですすまない。

＜今後の課題＞

今後の課題としては、以下のようなことがあげられる。（1）まず転写因子やターゲットを予測する個別の方法を改良し、予測精度を向上させる。（2）複数の方法や情報を組み合わせて予測精度を向上させる。（3）酵母ゲノムなどについて、予測結果と実験データを比較し、それを方法論の改良にフィードバックする。（4）細胞周期など、特定の機能に対応する制御ネットワークを解析する。（5）これらの研究に必要な転写制御に関する各種データベースや解析ツールの開発をさらにすすめる。

＜成果公表リスト＞

1) 論文／プロシーディング

1. 0801312056
   M. J. Araúzo-Bravo and A. Sarai "Role of DNA Conformational Change in the Specificity of Drug-DNA Recognition" *Nucleic Acids Res.* doi:10.1093/nar/gkm892 (2007).
2. 0801312053
   A. V. Kochetov, A. Palyanov, I. I. Titov, D. Grigorovich, A. Sarai, N.A. Kolchanov "AUG_hairpin: prediction of a downstream secondary structure influencing the recognition of a translation start site" *BMC Bioinformatics* 8:318 (doi:10.1186/1471-2105-8-318) (2007).
3. 0801312051
   S. Fujii, H. Kono, S. Takenaka, N. Go and A. Sarai "Sequence-Dependent DNA Deformability Studied using Molecular Dynamics Simulations" *Nucleic Acids Res.* (doi:10.1093) (2007).
4. 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).
5. 0801312037
   Y. Kaku, Y. Murakami, A. Sarai, Y. Wang, S. Ohashi and K. Sakamoto "Antigenic properties of porcine teschovirus 1 (PTV-1) Talfan strain and molecular strategy for serotyping of PTVs" *Archives of Virology* 152, 929-940 (2007).
6. 0801312034
   Y. Yonetani, H. Kono, S. Fujii, A. Sarai, and N. Go "DNA deformability and hydration studied by molecular dynamics simulation" *Mol. Simulation* 33, 103-107 (2007).

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