審査の結果の要旨

氏名　　劉　瑩

本論文では、既に公開されている複数の超並列レポーターアッセイのデータを用いて遺伝子の上流 DNA 配列から転写活性予測モデルを構築する手法を提案し、構築された転写活性予測モデルが従来法のものより高精度であることを示した。

遺伝子の転写はヒストン修飾やクロマチン構造、制御配列などさまざまな配列によって制御されている。遺伝子の発現を制御している様々な仕組みを理解し、遺伝子の転写ネットワークを解明することは生命の理解のために極めて重要である。本論文では遺伝子配列の上流領域における DNA 配列から遺伝子の転写活性を予測するシンプルなモデルを構築する手法を提案している。

本論文では過去に８つの論文で発表された超並列レポーターアッセイのデータを用いて再解析を行っている。超並列レポーターアッセイとは、プラスミドにターゲット配列（遺伝子上流の制御 DNA 配列相当）・レポーター遺伝子・ランダムバーコード配列を組み込み、作製したプラスミドライブラリから実際に遺伝子の発現量を観察する。具体的には cDNA のシークエンシングとバーコード配列のシークエンシングを組み合わせることでターゲット配列とそれに対するレポーター遺伝子の発現量をペアとして大量に観測する。

ある分化状態の細胞では、細胞の状態に応じて特定の種類のエンハンサーやプロモーターだけが活性を持っていると考えられ、構造が単純なプラスミドの上ではターゲット配列から遺伝子の活性が予測できると考えられる。そこで、ターゲット配列から遺伝子発現の活性が予測できることを仮定して発現量を推定するモデルを作ることを考えた。

このような予測モデルを構築するための従来手法では、制御配列の位置効果を学習モデルに取り込むために数多くのパラメータが必要となり、結果として可変長の DNA 配列に対する学習は多くの学習データを必要としていた。また、学習されたモデルの解釈は難しく、モデルから生物学的な転写ネットワークのメカニズムを直接推定・理解することは難しかった。そこで、本論文では DNA 配列を学習の特徴ベクトルとして直接用いることをせず、TRANSFAC データベースに収載されている DNA 結合モチーフ毎に Position Weight Matrix スコアを集計し、集計したスコアを特徴ベクトルの要素として用いている。この前処理により特徴ベクトルを大きく単純化した。また、発現量の予測を行うために決定木を用い大まかなクラスタリングを行った。決定木から得られた各ク

ラスター内では多変量適応回帰スプラインを用いることで最終的な発現活性の予測値を計算した。この際、過学習を避けるために自由度を制限している。

この問題に用いることのできる機械学習の手法には様々なものが考えられるが、提案手法では比較的シンプルな手法を組み合わせて転写活性を予測するモデルを作成している。より複雑な手法を用いることで単純な予測精度は向上させることが可能かもしれないが、生物学者にとって使いやすい転写活性予測モデルを提供するという観点からは提案手法が優れていると考えられる。具体的には、各遺伝子について予測に関与するパラメータ数が小さくなり、活性予測モデルから転写ネットワーク構造の考察がより容易になると考えられる。

また、本手法について交差検定により転写活性の精度を推定し、従来法より提案手法により学習された予測モデルの方が高精度に転写活性を予測できていることを示した。また、学習モデルに含まれるパラメータ数（予測子の数）も従来法より小さくなっており、人間による解釈がより容易になっていることを示唆した。

なお、本論文は、入江 拓磨、矢田 哲士、鈴木 穣との共同研究であるが、論文提出者が主体となって手法の提案・実装・分析・検証を行ったものであり、論文提出者の寄与が十分であると判断した。

よって、博士（科学）の学位を授与できると認める。

以上１，６００字