論文内容の要旨

Integrative bioinformatics analysis of transcriptional regulatory program in breast cancer cells
(乳癌細胞における転写制御プログラムの統合的バイオインフォマティクスによる解析)

氏名　新井田　厚司

転写プログラムの異常が癌の発生・進行に重要というのは周知の事実であり、これまでの癌研究の歴史の中で、様々な転写因子が原癌遺伝子及び癌抑制遺伝子として同定されてきた。また近年、マイクロアレイ技術による網羅的発現解析によりスクリプトームが細胞の癌化の過程で劇的に変化し、異なる癌のタイプで大きく異なることが明らかにされている。

様々な種類の癌の中でも、特に乳癌についてはマイクロアレイ技術による解析が広汎に行われている。乳癌は組織学的にも、予後、治療に対する反応性においても多様で、マイクロアレイを用いた発現解析は、このような表現型の多様性の下に潜む、トランスクリプトームの多様性を明らかにしてきた。しかしながらマイクロアレイにより得られた膨大なトランスクリプトームに関する知識と比べると、トランスクリプトームの多様性、更には表現型の多様性を生み出す遺伝子制御の機構については、得られている知識はごくわずかである。またこれまで、悪性度に関連する転写プログラムについてはまだほとんど解析はされてはおらず未知のままである。

転写制御プログラムを解明するためには、制御配列情報と網羅的発現プロファイルを統合する計算的アプローチが必須である。これまでに多くのアプローチが開発され、酵母のような下等生物の系に対しての適用はかなりの成功を収めている。しかしヒトのような高等生物のもつ複雑な遺伝子制御システムに対する適用はまだ萌芽的段階にある。この学位論文は乳癌に統合的バイオインフォマティクスによる解析を適用し乳癌の悪性度に関連するシス制御モチーフの存在を示すものである。

癌細胞における転写制御プログラムを解明するため、本研究ではBayesian networkをもちいて制御配列情報と網羅的発現プロファイルデータを統合し、発現プロファイルデータに付随する表現型情報に相関するシス制御モチーフを探索するための方法を確立した。まず始めに個々の遺伝子の発現量と表現型の相関を計算しさらにその値を"メタ発現値 (meta expression value)"として相関するようなシス制御モチーフをBayesian networkにより探索した。

まず始めに制御配列、制御モチーフ、発現値データの三つの統合すべきデータを用意し

た。制御配列として、は転写開始点の上流 500bp から下流 100bp のコアプロモーター配列を用いた。シス制御モチーフとしては二種類の PWM のデータセットを用意した。既知の転写因子結合モチーフは TRANSFAC 及び JASPAR データベースから手に入れた。更に、未知の制御モチーフを探索するために ab initio モチーフ発見プログラム(ab initio motif discovery program)の一つ DME を用いて"メタ発現値"の両端に位置する遺伝子の制御配列から頻出モチーフを抽出した。これらのモチーフ群の冗長性をクラスタリングにより取り除いたあと各遺伝子の制御配列に対しそれぞれの PWM スコアを計算し、そのスコアを複数の閾値により二値化し、sequence feature table を作成した。つまりここで sequence feature はある PWM スコアの閾値でのモチーフの有無を意味し、sequence feature table は行に各遺伝子、列に各 sequence feature を割り当てられた、二値行列となる。

発現値のデータとしては、マイクロアレイ実験により得られた網羅的発現プロファイルデータを用意した。それぞれのデータセットは数千の遺伝子の複数のサンプルでの発現値の情報をよびサンプルに付随する表現型の情報を含む（ 例えば histological grade や予後等の癌の悪性度に関する情報)。本研究においては生の発現値ではなく発現量と表現型の相関を計算し"メタ発現値（meta expression value）"とした。つまり発現値のデータはそれぞれの遺伝子の"メタ発現値"を各要素としてもつ 1 次元のベクトルとなる。発現値のデータは training data と test data に 3 対 1 の比で分け、training data に含まれる情報のみを DME を用いた新規モチーフサーチを含める一連のモチーフ探索解析に用い、test data を用いてその結果の評価を行った。

発現値に関連するモチーフの探索は Bayesian Network の構造学習を利用して行った。本研究の方法では sequence feature が遺伝子発現を制御する一層のネットワーク構造を仮定する。　この方法は酵母の系においてシス制御モチーフの組み合わせから遺伝子発現パターンの予測に成功した先行研究に基づいたものである。このアプローチのメリットは PMW スコアの閾値等の sequence feature に関する柔軟な条件や、高等真核生物において更に重要な働きをしているものと思われる sequence feature 間の相互作用を取り入れられることである。酵母の系の先行研究においては本質的に連続値であるはずの発現値を発現クラスターの帰属を表す二値のデータに離散化し解析を行った。しかしながらこのような離散化は情報の欠損につながり、また、解析結果は離散化の際の閾値の選択に依存する可能性が報告されている。　この問題を解決すべく本研究では新しい評価関数を導入して連続値のデータをそのまま扱えるようにした。この評価関数はあるデータが与えられた条件下で連続値の値がある組み合わせの二値のデータに依存するモデルの確率を表す。sequence feature table と発現値のデータに対して、この評価関数を最大化するような sequence feature の組み合わせを greedy search により探索した。まず発現値を制御する sequence feature のないモデルからスタートして、評価関数をもっとも増やすような sequence feature を一つずつ加えることを繰り返した。

はじめに、この方法の実用性を確かめるためにいくつかのヒト細胞の発現プロファイルデータの解析を行った。肝細胞特異的発現においてはHNF1及びHNF4の結合モチーフ、骨格筋細胞特異的発現においてはMEF2の結合モチーフ、HUVECにおけるTNFαによる発現誘導にはNFκBの結合モチーフが相関していることが示され、既知の報告との一致が見られることからこの方法の実用性が示された。

次に乳癌の組織学的多様性を生み出す転写プログラムに注目し、histological gradeに関連付けられるシス制御モチーフの探索を行った。Histological gradeは細胞の分化や増殖能に関する指標を統合したスコアで乳癌の悪性度を測る際に使われる。　発現プロファイルデータ中の全ての遺伝子についてG1（高分化型、67サンプル）とG3（低分化型, 高増殖性、54 サンプル）の間の発現量の違いをt統計量をもちいて計算し、Bayesian Networkを用いてその値に相関するシス制御モチーフを解析した。30個のbootstrap sampleを用いた再現性の確認と単一のsequence feature及びそのペアに対する順位和検定の結果より、ELK1、 E2F1、 NRF1 およびNFYの結合モチーフが有意なsequence featureだとわかった。これら4つの組み合わせに対するP値は1.33×10^{-15}と低く有意な結果であった。また実際の依存関係の解析によりこれらのsequence featureの存在がG3サンプル群での遺伝子のupregulationに正に相関していることがわかった。

最後に、より直接的に癌の悪性度を反映している指標としての予後に注目して解析を行った。それぞれの遺伝子に対して生存時間との相関をCox回帰モデルを用いて計算しその値に相関しているようなシス制御モチーフを探索した。解析の結果histological grade と同様、ELK1、 E2F1、 NRF1 およびNFYの結合モチーフが7.17×10^{-12}という有意なP値で予後と相関していることが示された。これらの結果を全て考慮に入れるとELK1、E2F1、NRF1 およびNFYの結合モチーフが主要な乳癌の悪性度と関連するシス制御モチーフであると考えられる。

ELK1はETS転写因子ファミリーの一員として知られている。ETSファミリーの転写因子は中央のコア配列がGGA[A/T]からなる似た様なモチーフに結合するので、ELK1結合モチーフも他のETSファミリーのメンバーにも結合すると思われる。ETSファミリー遺伝子の多くがRAS-MAPKシグナル伝達経路により制御される転写因子であり、ETS遺伝子の制御異常は細胞の悪性化、腫瘍化を惹起する。いくつかのETS遺伝子は白血病及びEwing腫瘍において染色体転座によりchimeric oncoproteinを形成していることが報告されており、ETS遺伝子の異常発現はその他様々な悪性腫瘍について観察されている。E2Fファミリーは DPタンパク質とヘテロ二量体を形成し、ファミリー間で共通の結合配列を認識するものと考えられている。E2Fファミリーは細胞周期のmaster regulatorとして知られ、G3腫瘍における遺伝子のupregulationとの関連は、histological gradeの基準に分裂指数が含まれ G3腫瘍が高増殖性と定義されている事実と一致する。NRF1 は核のゲノムにコードされるミトコンドリア遺伝子を発現誘導しミトコンドリアの呼吸能力（respiratory capacity）を上昇させることが知られている。癌細胞における NRF1 の機能は報告されて

いないが、NRF 結合モチーフとの相関は悪性腫瘍における代謝活性化を反映している可能性がある。

以上、本研究により、閾値パラメーターに依存しない Bayesian Network によるシス制御配列と発現値データの統合解析方法を確立し、この方法を用いて乳癌細胞の悪性化に関連するシス制御モチーフの存在を示した。今後、本研究で得られた知見、方法を基礎として、上記で述べたような様々なタイプのデータを統合・解析する癌細胞における遺伝子制御システムの包括的な理解に向けて研究が進むことを期待する。