

論文内容の要旨

論文題目

Identifying Combinatorial Regulation of Transcription Factors and Binding Motifs
(転写因子と結合モチーフによる組合せ的転写制御の推測)

氏名 加藤 護

生物にとって必要不可欠な細胞内プロセスの一つが転写である。転写は遺伝子発現の最初の段階であり、遺伝子の発現パターンは発生制御、形態形成、細胞分化、組織特異性、ホルモン伝達、細胞応答といった生命現象と直接関係する。転写は真核生物において高度に複雑であり、複数の遺伝子特異的転写因子が関与すると考えられている。遺伝子特異的転写因子は、プロモーターに存在する短い DNA 配列（転写因子結合モチーフ）に結合し、クロマチン構造の変化、あるいは転写前駆複合体形成のあらゆる段階で影響を及ぼす。この様に、真核生物の転写は（遺伝子特異的）転写因子とその結合モチーフとの組み合わせによって制御され、これらの様々な組み合わせが複雑な生命現象を生み出す第一歩であると考えられている。これを転写因子と結合モチーフの組合せ的転写制御と言う。

ゲノム研究において、遺伝子（mRNA）発現を全遺伝子規模で調べるために開発された技術が、DNA マイクロアレイである。DNA マイクロアレイは遺伝子の mRNA とハイブリダイズする cDNA が全遺伝子規模で貼り付けられたスライドガラスであり、これを使って全遺伝子の（通常、相対的な）mRNA 存在量を測定する事ができる。その規模は、例えば出芽酵母で測定されたケースを例に挙げれば、約 6000 遺伝子で 17 時刻点である。この単純な単細胞真核生物においてさえ約 100,000 程度の実数データが得られる事になり、これを手作業で処理する事は不可能である。大規模なデータを適切に処理し、生物学的に意味のある情報を抽出する計算機的方法の開発が必要となる。

マイクロアレイデータを処理する一般的な計算機的方法は次のようなものである。最初に、DNA マイクロアレイデータから得られた遺伝子発現パターンをクラスタリングする。次に、クラスタ化された遺伝子群に対してその上流プロモーター配列を調べ、統計的に有意に高頻出している、指定された長さの一種類のモチーフを抽出する。しかしながらこの方法には2つの限界点がある。一つ目は、真核生物の転写において基本となる組み合わせ的制御を解明できないというものである。単一モチーフだけを抽出するので、組み合わせ的制御を解明する事は出来ない。二つ目は、たとえモチーフが抽出できたとしても、それに結合する転写因子を特定する事が出来ないというものである。

近年 DNA マイクロアレイとは異なるタイプのマイクロアレイ技術が開発された。ChIP (クロマチン免疫沈降) マイクロアレイである。これには、コーディング領域にハイブリダイズする cDNA の代わりに、ノン・コーディング領域にハイブリダイズする cDNA が全遺伝子規模で貼り付けられている。対象とするタンパク質が結合しているプロモーターを抗体を用いて選別し、それを ChIP アレイにかける事によって、全遺伝子規模で (一) タンパク質 - (全) プロモーター間の相互作用の有無を調べることが出来る。最近、Yeast Proteome Database によって注釈付けされているほとんど全ての酵母転写因子に関する ChIP マイクロアレイデータが公表された。ChIP マイクロアレイは、転写因子がプロモーターに結合するか否かの強力な証拠を提供するので、転写因子が結合するターゲット遺伝子の同定だけでなく、転写因子が結合する制御モチーフの同定までゲノムレベルで行える可能性がある。しかしながら現在のところ、ChIP データを処理して、DNA マイクロアレイデータやプロモーター配列データと統合化し、従来方法が持つ2つの限界点を克服するコンピューティショナルな方法は開発されていない。

本研究において、私はこれを克服する新しい統合的方法を開発した。従来方法では DNA マイクロアレイデータだけを使用し、一種類の高頻出モチーフだけを探索していた。本方法では ChIP マイクロアレイから得られるデータを加え、これを使ってモチーフに結合する転写因子を推定する。さらに、一種類のモチーフではなく、複数種類のモチーフを探索するコンビナトリアル・モチーフ解析を採用する。本研究ではこれらを適切に統合化し、ある遺伝子セットに特異的なモチーフ・コンビネーションを予測するのみならず、モチーフ・コンビネーションを構成する要素モチーフに結合する複数の転写因子を予測する事も可能にした。

本方法は4つのステップからなる。1) 第一ステップにおいて、まず ChIP マイクロアレイデータから転写因子のターゲット遺伝子を決定し、そのプロモーターにおいて有意に高頻出している単一のモチーフを同定する。2) 第二ステップにおいて、それらモチーフのあらゆる組み合わせを列挙し、ある特定の遺伝子セット (本研究においては、G1、S、S/G2、G2/M、または M/G1 phase で特異的に発現する遺伝子セット) において有意に高頻出しているモチーフ・コンビネーションを選別する。3) 第三ステップで、プロモーター上にそれらモチーフ・コンビネーションを持っている遺伝子に対して、DNA マイクロアレイから得られる発現パターンをチェックし、発現パターンが時刻にわたって凝集しているモチーフ・コンビネーションをスクリーニングする。4) 最後に、ChIP データをもう一度使い、それら選ばれたモチーフ・コンビネーションを持つプロモーター・セットに結合する転写因子群を同定し、さらに (モチーフ・コンビネーションを構成する) 要素モチーフに結合する転写因子群を第一ステップの結果から列挙し、これら2つの転写因子群をマッチングして、モチーフ・コンビネーションを構成するそれぞれの要素モチーフに結合する転写因子を推測する。この手続きによって、ある遺伝子セットに高頻出するモチーフ・コンビネーションと、その要素モチーフに結合する複数の転写因子を予測することが出来

る。これはまた、それら転写因子とモチーフの組合せ (TF・motif コンビネーション) がそのセットの遺伝子を制御していることを直接的に示唆する。

私はこの方法を、生物学的現象としての重要性と利用可能なデータの存在から、出芽酵母の細胞周期に適用した。細胞周期はあらゆる生物に普遍的な細胞過程であり、出芽酵母の細胞周期に関する知識、データは比較的蓄積されている。しかしながら、多くの研究はその転写後のメカニズムに集中し、転写制御やそのメカニズムについては依然不明な点が多い。特に標準的なモデルでは幾つかの phase を制御する 5 つの制御因子 (あるいはそのコンビネーション) しか同定されておらず、制御因子が同定されていない、あるいは良く特徴付けられていない phase が複数存在する。

まず初めは、ChIP データからのモチーフ抽出 (ステップ 1) についての結果である。本方法は、従来知られていたモチーフに加え、新しいモチーフやバリエーションを予測した。例えば Mcm1 に対する ATAATTA (後に実験的に証明)、Mth1 に対する CAGCAG、Ash1 に対する GCGGCA 等である。さらに (個別的) タンパク - タンパク相互作用データと組み合わせることで、新しい複合体 Stb1 + Swi6 + Swi4/Mbp1 の存在を予測した (Mbp1 に関しては、後に実験的に証明)。さらに、ChIP データからの有意な高頻出モチーフが 3 つのタイプに分類される事を発見した。モチーフ 1 が転写因子 1 の ChIP データから抽出されたとする。一つ目は、転写因子 1 がモチーフ 1 に直接結合する direct-binding motif、二つ目は、転写因子 1 が他の DNA 結合転写因子と接触し、その DNA 結合転写因子がモチーフ 1 に結合する piggyback-binding motif、三つ目は、転写因子 1 がその抽出モチーフ 1 と同じプロモーター上にある別のモチーフに結合し、その抽出モチーフ 1 には別の転写因子が結合する cross-binding motif である。最後の二つは転写因子 1 がモチーフ 1 に直接結合しない、indirect-binding motif である。

次は、細胞周期の各 phase に特異的な TF・motif コンビネーションについての結果である。本方法は、従来知られていた制御因子 (のコンビネーション) に加え、新しいコンビネーションを発見した。これらをまとめ、私は酵母細胞周期における新しい転写制御モデルを提案した。従来のモデルは以下であった。G1 phase に {Mbp1, Swi6} (Mbp1 と Swi6 から成るコンビネーション)、{Swi4, Swi6} が ACGCGT, CRCGAAA にそれぞれ結合し、遺伝子を制御する。G2/M phase に Fkh1/2 と Mcm1 が、GTAAACAA と ECB モチーフに結合し、Ndd1 と協同して遺伝子を制御する。M/G1 phase に Mcm1 と Swi5/Ace2 が ECB モチーフと RRCCAGCR に結合して、遺伝子を制御する。新たなモデルでは、これらに、今回予測されたコンビネーションによる制御が加わった。例えば、G1 phase に {Stb1, Swi6, Mbp1/Swi4} コンビネーションが ACGCGA に結合して遺伝子を制御する。Ste12, {Swi6, Mbp1/Swi4} が TGAAAC, CGCGTC に結合して遺伝子を制御する。S phase に Met4, {Swi4, Swi6}, Hir1/2 が MET モチーフ, CGCGAAA, CCR 配列に結合してヒストン遺伝子を制御する、など。この新しいモデルにおいては、ほとんどの転写因子がコンビネーションの要素となっていた。これは酵母細胞周期における転写制御の描像を、従来の

単一転写因子による制御から、複数の転写因子による制御へと一変した。

この結果からさらに考察を進め、酵母細胞周期における TF-motif コンビネーションが3つの制御モードに分類できる事を発見した。一つ目は waiting-activating system であり、これは転写因子（の組合せ）が何らかのシグナルを受け取るまでプロモーター上に結合して待機し、受け取ると転写を活性化するというシステムである。二つ目は joint-process combination であり、2つ、あるいは複数の細胞過程（例えば、M/G1 phase 制御とフェロモン応答など）をつなぐ組合せである。3つ目は joint-phase combination であり、ある phase で主要な役割を果たす転写因子と、次の phase で主要な役割を果たす転写因子の組合せである。最後のモードは特に興味深い。というのは、まるで細胞周期における遺伝子制御を回し続ける為に、ある phase の転写因子が、中間の phase で、次の phase の転写因子に遺伝子制御をリレーしているかの様に見えるからである。

本研究において、今回開発した新しい方法が、従来方法が抱えていた二つの欠点を克服できることを実証した。そのみならず、今後実験によって詳しく調査されるべき、生物学的に興味深い幾つかの仮説が提案された。しかしながら本方法で解明できるのは転写因子と結合モチーフの組み合わせ、それらのターゲット遺伝子だけであり、これは完全な転写制御ネットワークの一部である。シグナル伝達経路、リガンド-レセプター相互作用まで含めた完全な転写制御ネットワークを解明するためには、今後、本研究を土台に、大規模タンパク-タンパク相互作用データ、ゲノム・スケールのノックアウトデータといった様々なゲノムデータを適切に統合化して行く必要があるだろう。