線虫のゲノムワイドな転写制御ネットワーク解析

A genome-wide analysis of transcriptional regulatory networks in *C. elegans*

東京大学医科学研究所　大里直樹

The Institute of Medical Science, The University of Tokyo　Naoki Osato

派遣期間　2009 年 4 月 1 日～2010 年 9 月 30 日

April 1, 2009 – September 30, 2010

研究機関　Program in Gene Function and Expression, University of Massachusetts Medical School, 55 Lake Avenue North, Worcester, Massachusetts, USA

研究指導者　Prof. Marian Walhout

Summary

To understand *C. elegans* transcriptional regulatory networks, we planned to develop a computational method to predict transcription factor binding sites (TFBS) from the results of yeast one-hybrid experiments (Y1H) in the lab. First, we predicted TFBS based on the results of Y1H, but we found that many TFBS candidates were predicted for each TF. To reduce the number of TFBS candidates, we examined positional biases of TFBS in promoter sequences and observed that some TFBS candidates were located near translational or transcriptional start sites. Interestingly, we predicted a TFBS for a dimer of a nuclear hormone receptor and it was validated experimentally. This implied that Y1H might be able to detect the binding of a dimer TF. To promote the analysis of a large-scale dataset, we need to improve computational and experimental methods to do these analyses automatically.

As a collaborative research, we predicted TFBS of the *C. elegans* DRM transcription factor complex from ChIP-chip experimental data produced in the lab, and found that the predicted TFBS was underrepresented on X chromosome and tended to be located near genes involved in early development. These results are consistent with the results of other experimental analyses.

線虫は多細胞生物のモデル生物として、様々な特徴をもっている。例えば、1) 受精卵から成虫に至る全細胞の発生、分化の過程が細胞系譜として明らかになっている。2) 体が透明で、発生段階を逐一観察できる。3) 3〜4 日で成虫になるため、子や孫の状況も容易に追跡できる。4) パーキンソン病、アルツハイマー病、遺伝性高血圧症などをはじめとする様々な疾患遺伝子の相同遺伝子が存在する。5) 全ゲノム配列が解明されている。さらに、ゲノム配列から遺伝子が予測された結果、線虫とヒトでは遺伝子数に大きな違いがないことがわかった。複数の遺伝子が連携して機能することにより、発生過程などのさまざまな生命現象を制御している。遺伝子ネットワークを調べることにより、生命現象の仕組みや進化、多様性について、より詳しく知ることができると期待される。

派遣先の研究室では、線虫の遺伝子ネットワークについて調べるために、線虫の約 900 種類の転写制御因子と線虫の遺伝子のプロモーター配列の相互作用を、Yeast one-hybrid assay (Y1H) を用いて調べている。Y1H の実験結果から、プロモーター配列上に存在する、転写制御因子の DNA 結合配列をバイオインフォマティクスの手法により予測し、実験による検証を行う。線虫の転写制御因子のうち、その結合配列が明らかになっているものはまだ少なく (〜100 種類)、線虫とヒトの転写制御因子結合配列に共通あるいは相違な部分や、線虫とヒトの転写制御の違いについてはまだよくわかっていない。派遣先の研究室において、私は初めに Y1H の実験結果から、ある転写制御因子が結合するプロモーター配列のセットとその転写制御因子が結合しなかったプロモーター配列のセットを作製した。次に既知のツールを組み合わせて、転写制御因子結合配列の予測を行った。しかし、転写制御因子の結合配列の予測は、結合配列にあいまいな配列(degenerate sites)があることを許して予測を行うと、非常に多くの候補が見つかってくる。そのため、転写制御因子が結合しなかったプロモーター配列の情報を用いて候補を絞ったが、まだ多数の候補が残る場合が多いことがわかった。次に予測された DNA 結合配列のプロモーター配列上の位置の偏りをもとに、さらに候補を絞ることを検討した。いくつかの既知の転写制御因子の DNA 結合配列について、プロモーター配列上の位置を調べると、遺伝子の翻訳開始（または転写開始）点に近い位置に偏って存在する場合があることがわかった。一方、Y1H の実験では、線虫の約 900 の転写制御因子と約 20,000 の遺伝子のプロモーター配列の組み合わせを調べるため、現在の実験解析のペースで進むとすると、すべて調べるために数年間はかかると考えられる（実際には徐々にペースを上げている)。転写制御因子によっては、結合したプロモーター配列の数がまだ少ない（5 個以下）場合もあり、このような場合は DNA 結合配列の位置の偏りを判断するのは難しくなる。酵母とヒトの転写制御因子の結合配列を予測した論文には、プロモーター配列上で位置の偏りを示すものは、それぞれ 74%と 30%程度見つかったことが報告されている。線虫においても、既知の転写制御因子の結合配列と同じあるいは似ている配列が、プロモーター配列上で位置の偏りを示していたので、プロモーター配列上の位置の偏りをもとに、ある程度、候補の配列を絞ることができると考えられた。

このような試行錯誤の後、研究室の Y1H の実験データから、実験による検証をするための DNA 結合配列の候補を選び出した。

実験の状況として、研究室における人の入れ替わりや研究者の興味や経験などの理由により、予測された転写制御因子結合配列の実験検証に関して、積極的に実験をする人を見つけることが難しいという問題があった。研究室に配属される前のローテーションの期間に来た学生が実験に興味を示したので、実験をしてもらうことになったが、Y1H の実験の経験はなく、一回の実験に通常 2〜3 ヶ月かかるため、複数の条件で実験してもらうことはできなかった。しかし、候補として選んだ DNA 結合配列のうち、核内受容体の結合配列についてよい結果が得られ、興味深いことに、ホモダイマーが結合すると思われる、二つ同じ配列が反復した、DNA 結合配列に結合することがわかった。候補配列を選ぶときに、反復した DNA 結合配列がプロモーター配列に多く見られることに気づき、この配列を試すことができた。Y1H は、通常、単量体の転写制御因子の結合を検出する方法と考えられているが、二量体を形成する転写制御因子の結合も見ることができる可能性が示唆された。また理由がよくわからない現象として、初めのプロモーター配列全体を用いた Y1H アッセイのときには、一つの DNA 結合配列のみを含むプロモーター配列への転写制御因子の結合が検出されたが、予測された DNA 結合配列の部分のみを用いたアッセイのときには、一つの DNA 結合配列のみでは転写制御因子の結合が検出されず、二つ反復した DNA 結合配列への結合は検出することができた。プロモーター配列全体を用いたときと、予測された DNA 結合配列のみを用いたときとでは、結果が異なった。理由として、単量体と二量体の結合の強さの違いの影響や、プロモーター配列の中の予測された DNA 結合配列以外の配列の影響が考えられる。既知の論文には、予測された結合配列を一個用いるだけでは十分ではなく、複数個を反復した配列でアッセイをすることがあり、他の予測された DNA 結合配列に対しても、このような条件を試すことができればよかったかもしれない。

DNA 結合配列の予測に関して、特に Y1H の実験により転写制御因子の結合が確認されたプロモーター配列の数が少ない場合は、複数の DNA 結合配列の候補から一つの候補に絞ることが難しいときがあった。実験で複数個の候補を試すことができればよいが、実験の負担がその分増えることになる。あるいは、そのような候補配列の実験による検証を後回しにして、Y1H により結合が確認されたプロモーター配列が増えるのを待つということも考えられる。このような予測の一連の作業を、個々の転写制御因子や実験データごとに行うとなると時間や手間がかかるため、少ない人数で大規模にこなすには、自動化する必要があると思う。今回は長期間に渡って滞在できる状況ではなかったことと、実験にも困難な部分があるため、自動化はすぐには難しいと感じた。

Y1H の実験データから転写制御因子の結合配列を予測する研究とは別に、他の研究室と共同で、線虫の転写制御因子複合体(DRM complex)の DNA 結合配列の予測を行った。この転写制御因子は線虫の vulval development など複数の機能に関わることが知られている。以前、研究室において、ChIP-chip 法を用いて、複合体のゲノム上の結合位置が調べられた。

その結果、X 染色体において、他の染色体と比較して有意に、複合体の結合が少ないことがわかった。そこで、複合体の DNA 結合についてより詳しく調べるために、ChIP-chip 法で明らかになったゲノム上の結合位置から、複合体の DNA 結合配列を予測することになった。ChIP-chip 法による実験データから、複合体のゲノム上の結合領域を、既存のツールを用いて調べた。その結果、使用するツールにより、結合領域がある程度、変わることがわかった。そのため、DNA 結合配列予測には、複合体が最も強く結合する領域のゲノム配列を、各染色体から 50 箇所、1kbp の長さで取り出して用いることにした。既知の転写因子結合配列予測のツールを複数使用して、配列を予測した。複数の候補配列から、そのゲノム上の位置と ChIP-chip 法による複合体の位置を比較することにより、候補配列を三種類に絞ることができた。現在、予測された DNA 結合配列に、実際に複合体が結合するのかどうかを、共同研究者に調べてもらっている。この候補配列のうちの一種類は、複合体が結合する、ゲノム上で近接した二箇所の結合部位を含むと考えられる。また三種類のうちの一種類は、複合体を構成する転写制御因子の一つが単量体でゲノム上に結合するときに予測された配列によく似ていた。ChIP-chip 法により、複合体と単量体の両方の結合部位が得られたのではないかと考えている。複合体の結合配列の候補は、単量体の結合配列と少し異なる配列で、複合体を構成することにより結合配列が少し変化する可能性が示唆された。この候補配列をゲノム上で探索してみると、ChIP-chip 法の結果と同様に、X 染色体上に有意に少ないことがわかった。実験解析を行っている学生が他の実験解析の結果と合わせて、この結果を学会で発表したところ、専門家から多くの注目を集めたようであった。現在、論文として投稿中である。

さらに、ChIP-chip 法のデータから、X 染色体上で有意に少ない転写制御因子結合配列が予測されたことから、さらに染色体上で分布に偏りのある配列が存在するかどうかを、計算機を用いた解析によって調べた。ChIP-chip のデータを用いないで、線虫のプロモーター配列から、5～11bp までの長さのすべての組み合わせの DNA 配列が、ある染色体に特異的に過剰または過少に存在するかどうかを統計検定を用いて調べた。その結果、X 染色体上で有意に少ない DNA 配列が他に見つかった。その配列の下流に位置する遺伝子の機能を調べてみたところ、初期発生に関連する機能が多くみられた。また、DNA 配列のプロモーター配列上の位置を調べたところ、トランスポゾンに由来する反復配列中に多く存在することがわかり、このトランスポゾンは、他の染色体と比較して、X 染色体上で少なかった。見つかった DNA 配列は、トランスポゾンに由来する可能性が示唆された。

計算機によってターゲットの DNA 結合配列を正確に予測することの難しさや、実験の難しい部分や研究体制の問題があり、短い期間で多くの解析を進めるのは厳しい状況であったが、実験データやゲノム配列からの DNA 結合配列予測に関して、新たな経験を得ることができ、転写制御に関する新たな知見も得ることができた。今後も今までの研究の経験をもとに、遺伝子ネットワークとさまざまな生命現象や疾患との関係について研究を進めていきたいと考えている。遺伝子ネットワークをゲノムワイドに調べようとする試みは、近

年、急速に進んでいて、大規模な実験解析も計画されている。これらの大量の実験データを解析するための技術開発やデータベースの構築など計算機による解析が必要とされる場面も多い。計算機による解析と実験を組み合わせて、効果的に研究を進めていくことができればと考えている。