非コード領域の配列モチーフ検出システムの構築とそれに基づくゲノムアノテーション

●須山 幹太
京都大学大学院医学研究科

＜研究の目的と進め方＞

ゲノム中には遺伝子コード領域の他に、遺伝子発現制御のための配列モチーフ（転写因子結合部位など）が存在する。これらのモチーフは比較的短い配列であるため、ゲノム中でのアノテーション（同定・分類）が困難な領域である。現在、多種多様な動物ゲノム配列決定が急速に進んでいることから、比較ゲノム解析による局所的な配列保存度合の評価から、非コード領域に存在する配列モチーフを解析することが可能な段階になってきた。これを踏まえ、本研究は、動物ゲノムを対象に、これらの配列モチーフ同定のための解析システムを構築し、得られたモチーフのゲノム中での高精度なアノテーションを網羅的に行なうことで、遺伝子発現制御についての理解を深めることを目的としている。

＜2008年度の研究の当初計画＞

・計算機環境の構築とプログラムの整備

非コード領域の配列モチーフの同定にはゲノム配列アライメントを活用するが、それは膨大なデータであるため、その解析には比較的記憶容量とメモリーの大きな計算機が必要である。まずそのための計算機環境を構築する。また、このゲノム配列アライメントを解析するには専用のプログラムを開発する必要がある。既に基本的な部分のプログラム開発をおこなっており、本年度はそのプログラムを完成させ、さらに、より汎用性のあるものへの拡張を計画している。

・モチーフの検出およびアノテーション

既にモチーフ検出のためのプログラムは複数開発されており、本研究では、それらを使用する。モチーフ検出の対象とするゲノムは、昆虫や脊椎動物など、すでに複数の近縁種ゲノムが決まっているものが考えられるが、本年度はまず脊椎動物を対象とした解析を行なう。実際に新規モチーフを検出するには、モチーフ検出プログラムのパラメータを検討する必要がある。そこで、まず既知モチーフの検出能力を指標に、パラメータの最適化を行なう。次に、得られたモチーフについて、ゲノムアライメント中での保存度合を評価することで、ゲノム配列中での確度の高いアノテーション（同定・分類）を行なう。

＜2008年度の成果＞

これまでにゲノムアライメントから任意の部位での部分アライメントを高速に抽出するためのプログラム開発を行なってきた。その核となる部分は既に完成している。具体的には、アライメントとしてUniversity of California, Santa Cruz（UCSC）のGenome BrowserよりMAF形式で提供されているゲノムアライメントを使用した。このアライメントを自作プログラムによりインデックス化することで、任意の部位および生物種のアライメントが瞬時に取得できるようになった。これによりゲノムアライメントをもとにした、非コード配列モチーフの大規模な解析が可能となった。さらに、Adam Siepelらによって開発されたアライメント中での保存部位評価法である*phastCons* [A. Siepel *et al.*（2005）*Genome Res.* 15:1034-1050.]の結果を大量に一括して視覚化するプログラムを開発した。視覚化することで、各モチーフの保存の程度とゲノム上近傍に位置する遺伝子や他の保存部位との関係が容易に把握できるようになった。

また、マイクロアレイのデータをもとにした解析から、コアプロモータ領域に存在する新規モチーフの同定に成功した。これは本研究遂行のためのパイロットスタディーと捉えることができる。そのモチーフをヒトゲノム中に検索し、1000以上の遺伝子の上流に存在することを明らかにした。さらに、そのモチーフに相当する領域のゲノムアライメントの解析から、そのモチーフが、哺乳類だけでなくニワトリやゼブラフィッシュを含めた脊椎動物にも高頻度に存在することを見出した。

さらに、ある遺伝子に的を絞った詳細な解析として、peripheral myelin protein 22（*pmp22*）の発現制御に関与するシス因子の解析を行なった。この遺伝子は正常なミエリンの形態形成において重要な役割を担っており、その発現レベルの異常が、たとえばCharcot-Marie-Tooth disease type 1A（CTM1A）といった神経疾患を惹き起こしていることが知られている。我々はメダカをモデルとし、その遺伝子の構造が哺乳類のそれと同じことを実験的に決定した。またメダカにおいて*pmp22*を過剰発現させると、哺乳類で見られるのと同じような、神経伝達速度の遅延が観測された。さらに、脊椎動物間での比較ゲノム解析から、メダカと哺乳類の間で保存されている非コード領域モチーフを新たに2つ見出した。これらのモチーフが*pmp22*の発現制御に関与していることが示唆される。

＜国内外での成果の位置づけ＞

ゲノムアライメントをもとにした局所保存配列についてのデータ解析として、Xieらによるものがある[Xie *et al.*（2007）P*roc. Natl. Acad. Sci. USA* 104:7145-7150.]。彼らの研究は、特に高頻度で見られるモチーフに的を絞った解析である。

本研究はモチーフの検出だけでなく、それがゲノム中にどのように存在するか、またそれぞれについてどれくらい確からしいかを評価することでゲノムワイドな非コード領域のアノテーションを試みる点が独創的である。また、ゲノム中には頻度は低いが何らかの機能を担っている考えられる配列モチーフも多数存在する。それらも考慮に入れ、ゲノム配列中での非コード領域の配列モチーフについての高精度アノテーションを目指している。このようなデータはこれまでになく、本研究で得られた結果をデータベースとして公開することは、分子生物学的や分子医学においても有用な情報資源となると考えられる。

＜達成できなかったこと、予想外の困難、その理由＞

ゲノムアライメントには配列決定が未完な生物種も含まれており、解析対象としている領域で、ある生物種の配列が未完であると、配列保存度合の評価が困難になる。例えば、ある遺伝子の上流部分を解析していて、ゲノムアライメント中に含まれるほとんどの哺乳類において保存された局所配列があっても、ただ一種類の生物種でそこの配列が未完であった場合、実際にはその部位がその生物種でだけ異なるのか、それとも完全に保存されているかを知ることは実験的に配列決定をしない限り不可能である。そのような未完配列の混入は予想以上に多く、遺伝子のプロモーター領域の半数程度でそのような未完配列がみられた。

この問題への対処法としてアライメント中での保存部位評価法である *phastCons* [A. Siepel *et al.*（2005）*Genome Res.* 15:1034-1050.] の使用を考えている。このプログラムを用いれば、任意の数の生物種からなるアライメントに対して、その進化系統的な距離に従った保存度合の評価が可能になる。そこで、解析しようとする領域のアライメントにゲノム配列が未完の生物種が含まれている場合、その生物種を除外したアライメントを作成し、*phastCons* によって保存度合を評価し直す、という手段を踏めば未完ゲノム配列の混入に対処できると考えられる。

＜今後の課題＞

新規ゲノム配列は今後も増え続けることが予想される。一般にゲノム配列が増えるほど、モチーフなどの機能性部分配列の検出感度が上がると考えられる。また、ゲノムアライメントのためのプログラムも、複数のグループにより改良が試みられており、より精度の高いゲノムアライメントが得られるようになることが期待できる。そのため、常に最新のデータ（利用可能なゲノム配列とアライメント法）を用いた解析に留意することが必要である。

モチーフを抽出する具体的な方法としては、２つの方法が考えられる。一つ目の方法は、ゲノムワイドに保存部位を集め、類似した環境（3'UTR など）に共通して現れるモチーフを探す、といったものである。この方法は Xie らによるもの [Xie *et al.*（2007）*Proc. Natl. Acad. Sci. USA* 104:7145-7150.] に多少類似してくるが、頻度が高いモチーフに重点を置かないことで、より多くの新規モチーフが得られる可能性がある。二つ目の方法は、マイクロアレイの結果から、ある条件で共に発現してくる遺伝子のプロモーター領域に焦点をあて、そこに共通している保存配列からモチーフを同定する方法である。使用するマイクロアレイの結果としては、NCBI で作成しているマイクロアレイデータのレポジトリである、Gene Expression Omnibus（GEO）が有用であると考えている。

また、新規の転写部位結合部位などのモチーフが得られた場合、データ解析の信憑性を確認する意味でも、実験による検証が重要であると考えられる。具体的な方法としては、たとえばゲル・シフト・アッセイが考えられる。現在、そのための最低限の実験設備を整えている。まず、既に得られているいくつかの新規モチーフの候補について、何らかのトランス因子の結合部位であるかどうかの確認を行なう予定である。系が確立したら、新規モチーフの候補について、順次、実験的な検証を行なう。

得られた非コード領域モチーフを用いた応用的研究として次の２つの研究を考えている。ひとつは、ヒトに固有の転写因子結合部位の変異の探索である。ヒトと他の霊長類との表現型の違いをもたらすものとしては、遺伝子コード領域の差異だけでなく、遺伝子の発現制御を担う領域、たとえば転写因子結合部位での変異による遺伝子発現の差が考えられる。これまで、そのようなゲノム中でのシス因子のアノテーションは、遺伝子コード領域のそれに比べてあまり進んでいなかったが、本研究で得られるゲノムワイドな非コード領域モチーフのアノテーションをもとにして、この方向での解析が容易になると考えられる。もうひとつの研究は、転写因子結合部位などのシス因子の進化に関するものである。具体的には、遺伝子コード領域の変異速度とシス因子の変異率の比較を行なう。それにより、種分化の要因として、遺伝子産物本体の変異が重要なのか、発現制御を行なうようなシス因子が重要な役割を果たしているのか、についての知見が得られるものと期待される。

また、最近、ゲノム中での非コード領域モチーフ探索の派生的研究として、特異的抗体を用いたゲノムの修飾部位の検出にも着手した。断片化したゲノム配列を修飾に特異的な抗体を用いて免疫沈降し、次世代シーケンサーを用いた配列決定する、ChIP-seq 法を行なう予定である。得られた配列断片に共通して存在するパターンを調べることで、ゲノムの修飾についての配列モチーフの検出が可能になることが期待できる。

これらの解析から得られたモチーフについて、ゲノムスケールでのアノテーションを行ない、データベースとして公開する。その際、ゲノム中での各モチーフについて、それが真の機能部位であるかどうかの確度を与えることは大変重要である。なぜなら、短いモチーフの場合、確率的に偶然現れることも多く、擬陽性である場合が多いからである。この判別には、ゲノムアライメント中での保存の強さや、近傍の遺伝子までの距離やその遺伝子の転写の方向などが重要な指針となると考えられる。また、モチーフとアノテーションされた部位にみられる SNPs との対応を取ることで、機能に影響するような SNPs の候補も得ることできるので、そのデータも合わせて公開する。

＜成果公表リスト＞

1）論文／プロシーディング（査読付きのものに限る）
なし

2）データベース／ソフトウェア
なし