大規模ゲノム情報の比較技術と知識発見

●矢田 哲士
京都大学大学院情報学研究科

＜研究の目的と進め方＞

本研究課題では、ポストシークエンス時代に切望されるさまざまな配列比較技術を確立するとともに、その適用による配列情報の解読に取り組む。具体的な研究項目として、(1) miRNA のターゲット推定、(2) ゲノム比較によるアノテーション（(株) 三菱総合研究所・野口英樹研究員との共同研究)、(3) ゲノム配列間距離の計測（京都大学・北村隆行研究員との共同研究)、(4) プロモーター配列の大域的なモデル化に挑戦する。2008 年度は、主として (1) と (3) の研究項目について考察した。

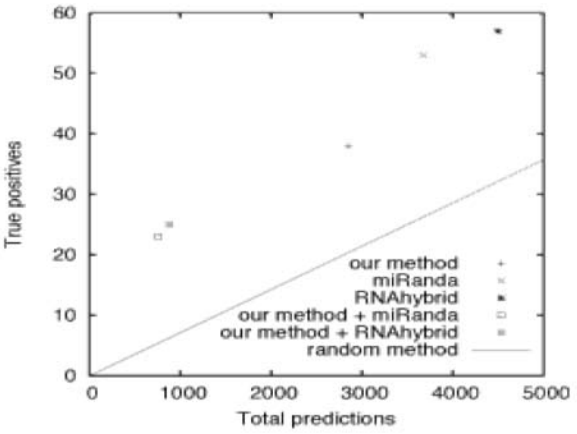
＜ 2008 年度の研究の当初計画＞

(1)miRNAのターゲット推定では、miRNAとターゲット遺伝子の間では転写情報の一部が共有されているとの仮説を立て、この仮説に立脚したターゲット遺伝子推定法の確立を試みる。miRNAが適切な遺伝子をターゲットするためには、両者が同じ時期に存在しなければならず、その制御の一部は、両者の転写によってなされている可能性がある。昨年度は、実験的に検証されたヒトmiRNAのターゲットデータを網羅的に収集し、miRNA（あるいはそのホスト遺伝子）のプロモーターとターゲット遺伝子のプロモーターに共通のシス因子が存在するかどうかを調べたところ、およそ40%のデータについて、統計的に有意なシス因子の共通性が観察された。本年度は、このシス因子の共通性に基づいたmiRNAのターゲット遺伝子の推定法を開発し、その有効性を評価する。

(3)ゲノム配列間距離の計測では、全長数百万塩基以上に渡るゲノム配列間の距離及びシンテニー領域抽出に対する計算コストを抑え、且つ高精度に計測する新たな算出法の確立を目指す。その方法として、コルモゴロフ複雑性を基づいてデータ配列の類似性を算出する正規化情報距離（normalized information distance、以下NCDに略）をゲノム配列比較に適応し、配列間距離の算出を試みる。この方法は近年mtDNAやタンパク質などで比較的短い配列において検証が始められているが、全長数百万塩基以上の塩基配列に対する有効性は未だ検証されていない。我々は今回、真正細菌の完全ゲノムや真核生物の染色体全領域など全長数百万塩基以上の塩基配列に対して適用し、その有効性を検証する。

＜ 2008 年度の成果＞

(1)miRNAのターゲット推定では、miRNAとターゲット遺伝子のプロモーターに観察される統計的に有意なシス因子の共通性に立脚したmiRNAのターゲット遺伝子の推定法を開発し、その有効性を評価した。この推定法では、まず、与えられたヒトmiRNAについて、そのプロモーターを同定し、その多重アラインメントから候補シス因子を抽出する。ここでは、ヒト、マウス、ラット、イヌで完全に保存されている6塩基以上のオリゴマーを候補シス因子とした。次に、あるヒト遺伝子について、miRNAの場合と同じ手順で候補シス因子を抽出し、miRNAの候補シス因子との間の共通性を検出する。そして、5%の有意水準を満足するシス因子の共通性が観察されると、その遺伝子は与えられたmiRNAのターゲットであると判定する。この推定法の精度は、ランダム法の精度を統計的に有意に上回り、miRNA-mRNAの結合部位の保存性に頼らない従来法の精度とほぼ同等であることが確かめられた（下図を参照）。このことは、ヒトmiRNAのターゲッティングが転写のレベルで調節されている場合があることを示している。



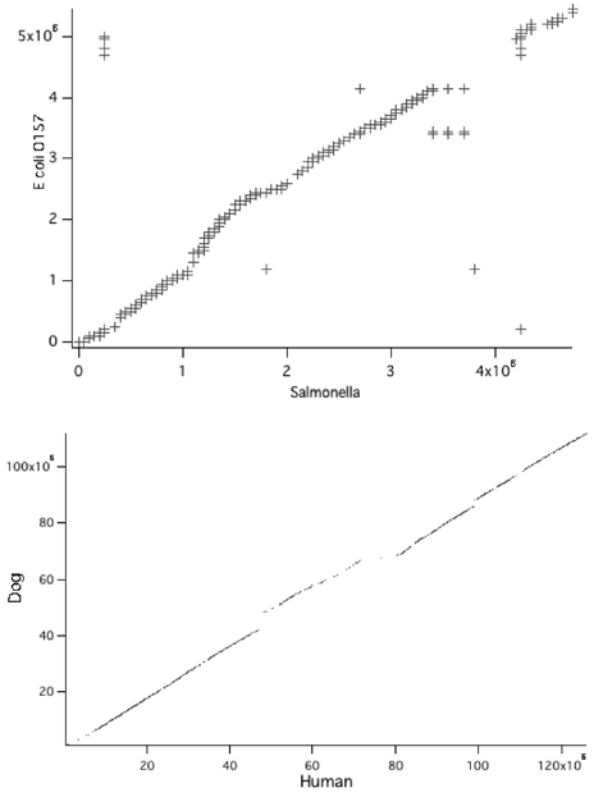
(3)ゲノム配列間距離の計測では、全長数百万塩基以上で構成されているゲノム配列に対し、計算コストの高いアラインメントを行うことなく、配列に内在する巨視的な特徴に基づいて配列を分類する新しい手法の確立を目指した。まず手始めとして、Evolsimulatorなど人工データを生成するソフトウェアを用いて百万塩基前後の配列を生成し、塩基配列の長さに対するNCDの有効性、適用限界を検証した。また、実在する二つの比較対象の配列に対して新たな手法として、まずNCDが適応可能と考えられる一定の固定長に配列をブロックに分割、ブロック単位での配列のペアに対して、全てのブロックの組み合わせのNCDを算出する。次に全てのブロックを使用し且つ重複が無い条件の下、ブロック単位でのNCDの合計が最小になる組み合わせとなるブロックペアを求め、そのペアの組み合わせを基に、配列全体におけるNCDを再計算することによって配列全体の情報距離を算出する手法を構築した。この方法により、バクテリアの完全ゲノムに対するアライメントなど他の手法で得られるシンテニー領域、配列間距離と同等の結果を算出することが確かめられた。

＜国内外での成果の位置づけ＞

(1)miRNAのターゲット推定では、miRNAとターゲット遺伝子の間では転写情報の一部が共有されているとの仮説を統計的に検証し、この仮説に立脚したターゲット遺伝子の推定法を確立した。従来の推定法は、miRNAとターゲット遺伝子の結合部位に観察される統計的な特徴や近縁種におけるこの部位の保存性に着目してターゲット遺伝子を推定してきた。つまり、本研究課題でのアプローチは、miRNAのターゲット遺伝子の推定に全く

新しい視点を導入する独創的なものである。また、miRNAのターゲットデータには、近縁種における結合部位の保存性が認められないデータが30%ほど存在し、その比率はデータの蓄積に従って増加する傾向にある。近縁種における結合部位の保存性に着目している従来法は、これらのデータに対してほとんど無力であるが、ここでのアプローチは、これらのデータに対して堅牢であることが計算機実験で示されている。

⑶NCDにおけるバクテリアにおける計測では、*E.coil* CFT073-*S. dysenteriae*など比較的遠い種などの場合においても、ゲノムアライメントが示すシンテニー領域と同じシンテニー領域を示し（下図上を参照）、またその配列のブロック化による手法で得られた配列間距離から作成した系統樹は、他の研究で既に妥当性があると考えられる系統樹と同じもしくは類似した結果を示した。染色体レベルにおいては、ヒト、マウス、イヌなどの哺乳類のX染色体を比較した場合、シンテニー領域を抽出する他のソフトウェアと同じ領域を示した（下図下を参照）。また、ヒトの17番染色体とマウスの11番染色体などの別の染色体同士の比較においては、他のソフトウェアでは見られないシンテニー領域を示す場合も見られた。



＜達成できなかったこと、予想外の困難、その理由＞

⑴miRNAのターゲット推定では、以下のことを明らかにした。まず、miRNAのプロモーターに存在するシス因子とターゲット遺伝子のプロモーターに存在するシス因子について、およそ40%のデータで統計的に有意な共通性が存在する。そして、この共通性に立脚したターゲット遺伝子の推定が可能である。この共通性は、当該プロモーターが、他のプロモーターに比べ、保存度が高いことに起因するようである。このことは、miRNAやターゲットの遺伝子が転写のレベルで複雑に制御されていることを示している。また、この推定法は、学習データを必要としないポータブルな推定法である。一方で、この共通性を裏付ける独立な証拠を集めることができなかった。Barikは、イントロンに存在するあるmiRNAが、そのホスト遺伝子と拮抗する機能を持つ遺伝子ファミリーのmRNAをターゲッティングする例を報告している（NAR 2008）。Martinezらは、イントロンに存在する線虫miRNAの発現が、そのホスト遺伝子よりも、それ自身によって制御されていることを報告している（Genome Res. 2008）。

⑶配列をブロック化しNCDを計算する手法は、シンテニー領域が局在しているとする仮定を前提としているが、比較的近縁種の場合においてもシンテニー領域が細かく且つ幅広く分散している場合が存在する。このような場合、我々の手法は他のソフトウェアに比べてシンテニー領域を検索する感度が落ち、ブロック化し計測されたNCDが過大評価される問題が生じた。特に比較する二つの配列の全長が大きく異なる場合においてこの傾向が多く見られた。またこのような状況に対してブロックサイズを小さくし、精度を高めることができるがその場合、他のソフトウェアに比べて計算速度が著しく低下してしまう問題が生じた。

＜今後の課題＞

⑴miRNAのターゲット推定では、以下の項目を考察したのちに研究成果を論文にまとめる。予測できた、あるいはできなかったターゲット遺伝子に何らかの傾向はないか? miRNAとそのターゲット遺伝子のプロモーターに観察される共通のシス因子には、どのような既知のシス因子が含まれているか? 共通のシス因子として数多く観察されるものはないか? 一方、シス因子の共通性を裏付ける独立な証拠を集める努力を継続する。以上が順調に進捗した場合は、ヒト以外の生物種について、我々の推定法の有効性を検証する。さらに、その結果をデータベース化するとともに、インターネットを介したターゲット推定を提供するシステムを開発する。

⑶ゲノム配列間距離の計測研究では、NCDによる計測の精度及び計算速度の向上を目指す。計測精度の向上についてだが、現時点では一般的に普及している汎用の可逆圧縮アルゴリズムを用いているため、ゲノム配列の特徴を捉えたデータ圧縮を行っていない。近年では特定のデータ配列に特化し、その配列の特徴のみを抽出、圧縮する不可逆圧縮を用いた手法が注目されているが、我々はこの手法を応用し、ゲノム配列専用の不可逆圧縮の開発を試みる。計算速度においては、これまでは配列を分割するブロック長を固定し、局所的なNCDを計測していたが、計算速度の向上のため、まず手始めに比較的大きなブロック長でNCDを計測し、次にその結果からシンテニー領域である可能性が高いと思われる領域のみ、もしくは優先的に計測するという段階的な手法を採用することにより計算時間の削減を試みる。

＜成果公表リスト＞

1）論文／プロシーディング

1. 0901141128
Sung-Joon Park, Natsuhiro Ichinose, and Tetsushi Yada: Probabilistic Graphical Modeling for Large-scale Combinatorial
Regulation of Transcription Factors, In the Proc.of Workshop on Knowledge, Language, and Learning in Bioinformatics (KLLBI-08) ,72-86 (2008) .