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・知能情報学専攻　生命情報学講座

＜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。

＜ 2007 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。そこで、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、その統計的な有意性を評価する。

(2)ゲノム比較によるアノテーションでは、特に non-CpGプロモーターにおいて、引き続き予測に重要なエレメントの抽出と効果的な予測手法の検討を行う。一方で、下流のエキソンの情報なども取り入れた最終的なプロモーター予測モデルのプロトタイプ構築を行う。現状の予測モデルでもすでに既存モデルの予測精度(non-CpGプロモーターに対する)を大きく上回るが、さらに下流に存在する遺伝子の構造情報 (コード領域やスプライスサイトなどの情報)を統合することで、特異性の高い実用的なプロモーターのアノテーションを目指す。我々は、過去の研究ですでに高精度の遺伝子予測モデルを構築しており、これをプロモーター予測モデルに統合することで、迅速に高精度な統合モデルの構築が可能である。

(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を計算コストを抑えて計測する新たな算出法の確立を目指す。その方法として実空間繰り込み群により配列情報の縮約を行うことによって、各配列の固有のマクロなパターン情報を抽出し、その配列同士の情報を比較をすることによって配列間距離を計測する方法を試みる。また、別の方法として、各配列のコルモゴロフ複雑性を計測し、その類似性を用いたゲノム配列間距離の算出も試みる。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、ここでは、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。

＜ 2007 年度の成果＞

(1)miRNAのターゲット推定では、以下に示す解析により、miRNAとターゲット遺伝子の間では転写情報の一部が有意に共有されていることが示された。ここでは、実験的に検証されたヒトmiRNAのターゲットデータ 80例について、 miRNAに関する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。まず、各々のプロモーターについて、ヒト、マウス、ラット、イヌで保存されている 6塩基以上のオリゴマーを抽出し、次に、相互作用する miRNAとターゲット遺伝子のプロモーターの間で抽出されたオリゴマーを比較し、両者に共通するオリゴマーを同定した。すると、 40%のデータについて、統計的に有意なオリゴマーの共通性が検出された。

(2)ゲノム比較によるアノテーションでは、転写開始点 (TSS)周辺に CpGアイランドを持たず遺伝子発現の組織特異性が高い non-CpGプロモーターを中心に、ヒトーマウス間の比較ゲノム解析とコアプロモーターの予測モデル構築を行っている。 non-CpGプロモーターでは、 TSS近辺からその上流 200 bpまでの領域が、CpGアイランドを持つプロモーターと比較して有意に保存されていた。そこで、この領域を中心に位置特異性を考慮した 2次のマルコフモデル (重み行列 (=0次のマルコフモデル)の拡張版)を構築し、予測精度が最も高くなる領域を調べたところ、 [-50,+50](TSSを+1とする)の範囲のとき AUC値 (ROCカーブの下側面積)が最大 (0.88)となった。さらに、マウスの相同領域でも同じモデルを構築してヒトモデルのスコアと統合することで、 AUC=0.91という高い精度を実現できている。 [-50,+50]の領域には、上流部の TATAボックスや TSSのイニシエーターのほかにも、下流にDPEなどの重要なコアプロモーターエレメントが存在する。一方で、これらのエレメントを持たないプロモーターも本手法で予測できていることから、当該領域にはこれら以外にも未知のエレメントもしくは、何らかの配列的特徴が潜んでいるものと考え、解析を進めている。

(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。ここでは、フラクタル現象から特徴を抽出する際に用いられる実空間繰り込み群をゲノム配列に適用することで、あるいは、画像データや音楽データの分類に用いられるコルモゴロフ複雑性をゲノム配列について計算することで、配列間の距離を計測した。古細菌、真正細菌のゲノム配列に関する計算機実験では、これらのアプローチによって計算された配列間の距離は、生の配列から計算された距離とよく一致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の確立に道筋をつけた。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。

(2)ゲノム比較によるアノテーションでは、 CpG/non-CpGプロモーターの両方で高精度なコアプロモーター予測モデルを構築した。特に、 CpGアイランドという特異な配列を持たない non-CpGプロモーターを高精度に判別した意義は大きい。本研究を通じて、基礎転写活性にかかわるコアプロモーターが、組織特異性の高い non-CpGプロモーター間でも、また種間 (ヒトーマウス間)でも共通性の高い配列であることが示された。組織特異性の高い発現パターンを示す遺伝子の多くが non-CpGプロモーターで制御されていることが、マイクロアレイ情報を用いたこれまでの解析で明らかになっており、発現制御機構を知る上で、 non-CpGプロモーターの構造について一定の知見を与える本研究の成果は意義深い。

(3)ゲノム配列間距離の計測では、実空間繰り込み群をゲノム配列に適用することで、あるいは、コルモゴロフ複雑性をゲノム配列について計算することで、配列間距離の計測を実現した。古細菌、真正細菌のゲノム配列に関する計算機実験では、これらのアプローチで計算された配列間距離は、生の配列から計算された距離とよく一致することが確かめられた。一方で、次のような課題も明らかになった。前者では、塩基組成が大きく偏るゲノム配列では、繰り込み操作によってその偏りが過大評価され、その結果、配列の急激な単純化が起こり、配列の情報が大きく消失する。後者では、配列間距離の差が思いのほか小さく、その堅牢性に議論の余地が残った。

(4)プロモーター配列の大域的なモデル化では、シス因子が同定されたプロモーター配列のセットから、シス因子の形成する規則性をモデル化するアルゴリズムの開発に成功した。ここでは、モデル化の枠組みに HMMを導入しているため、導出されたモデルをそのままプロモーター発見に適用することが可能である。シス因子の並びとそれらの間の距離を HMMによってモデル化したものには McPromoter[Ohler 2006]がある。ここでは、 left-to-right型の HMMが幾つも並列に連結された HMMを用意することで、プロモーター配列の多様性をモデル化することに成功している。McPromoterが (準)教師あり学習によってモデルを構築するのに対し、ここでの手法は教師なし学習によってモデルを構築する。また、 McPromoterは left-to-right型の HMMでプロモーターをモデル化するのに対し、ここでの手法は汎用性の高い一般的な構造を持つ。計算機実験では、配列ファミリーを特徴付ける主要なシス因子は left-to-right型の構造で詳細にモデル化し、副次的なシス因子は状態をマージして汎化していることが確かめられている。

＜達成できなかったこと、予想外の困難、その理由＞

(1)miRNAのターゲット推定では、予備的ではあるが、 miRNAのプロモーター情報に基づくターゲット遺伝子の推定を試みた。まず、ヒトの miRNAに関するプロモーターについて、ヒト、マウス、ラット、イヌで保存されている 6塩基以上のオリゴマーを抽出し、次に、このオリゴマーの組を問い合わせ配列として、ヒトの全タンパク質遺伝子のプロモーターを検索する。各プロモーターの検索領域は、ヒト、マウス、ラット、イヌで保存されている 6塩基以上のオリゴマー領域である。そして、有意な数のオリゴマーの共有が認められたものをターゲット遺伝子と推定する。その結果、この試みでは、数多くの擬陽性を検出してしまうことが明らかになった。その数は、従来法が miRNAのターゲット結合部位の保存性に着目しない時とほぼ同程度である。この原因として、頻出するシス因子をプロモーターに多く含むものやオーソロガスな関係にあるものが擬陽性の中には多いと考えている。

(2)ゲノム比較によるアノテーションでは、コアプロモーターが、組織特異性の高い non-CpGプロモーター間でも、また種間 (ヒトーマウス間)でも共通性の高い配列であることが示せた一方、組織特異的な発現制御を担う領域の特定には至っていない。また、TSS上流 200 bpまでの領域が non-CpGプロモーターにおいて有意に (種間で)保存されていることの意味も、未解明のままである。これらの、プロモーター間での共通性が低い (または持たない)領域から機能部位を抽出するためには、多種間でのマルチプルアライメントを注意深く行う必要があると考えられる。

(3)ゲノム配列間距離の計測では、実空間繰り込み群による試みにおいては、真核生物に繰り込み変換を用いた場合、本来比較的遠い種同士である 2つの配列に対して、謝って類似した配列同士であると誤判定する結果が多く生じた。これは繰り返し配列などの配列に多く含まれる特定のパターンを過大評価するためだと考えられ、この問題を解決するために、繰り込み関数が行うパターンの変換過程を精査していく必要性がある。コルモゴロフ複雑性を用いたゲノム間距離の測定については、計測する各配列の全長が一千万塩基以上の場合、各配列間距離同士の差が生じにくくなる傾向が見られた。これはコルモゴロフ複雑性を算出するために用いられる一般的な圧縮アルゴリズムは、長い配列の情報を効率的に圧縮するのが困難なためだと考えている。

＜今後の課題＞

(1)miRNAのターゲット推定では、 miRNAとターゲット遺伝子の転写情報の共通性に立脚した推定法の確立に力を注ぐ。これまでに、このアプローチの有効性を検証することができたが、同時に、擬陽性の検出を改善する必要性も明らかになった。ここでは、統計学的な視点と生物学的な視点の両面からこの課題に取り組む。さらに、哺乳動物に限らず、植物や昆虫などについて、このアプローチの有効性を検証する。

(2)ゲノム比較によるアノテーションでは、継続して TSS周辺の保存領域からの情報抽出を行う。また、 CpG/non-CpGプロモーターのモデルおよび下流の転写領域の情報を統合した最終的なコアプロモーター予測モデルの構築と、構築したモデルによるゲノムの解析およびアノテーションを行う。上記の試みが順調に進捗した場合は、マルチプルアライメントを利用したシスエレメント予測手法の開発と、同定したエレメントによるプロモーターのクラスタリング、発現プロファイルとの関係解析を試みる。

(3)ゲノム配列間距離の計測研究では、 Tsallisエントロピーを用いたゲノム配列間距離の算出法を考察する。これまで、コルモゴロフ複雑性をゲノム配列について計算することで配列間距離の計測してきたが、現象の複雑性をより明快に定義できるエントロピーの導入を試みる。 Tsallis統計はべき則的分布を持つ現象をうまく記述できることで知られているが、我々は、さまざまな生物種のゲノム配列において、 k-merの頻度分布がべき則的分布に従うことを確認している。

(4)プロモーター配列の大域的なモデル化では、まず、シス因子が現れる鎖の情報をモデル化する処理と Ward法による分類数を機械的に決める処理をモデル化アルゴリズムに実装する。また、アルゴリズムの性能や特性を客観的に評価するために、ベンチマークデータを精力的に収集する。そこで、 ENCODEプロジェクトから産出されたデータや関連研究が利用したベンチマークデータを吟味する予定である。また、人工的なデータの利用についても検討を行なう。さらに、構築されたモデルによるプロモーターの発見精度を通して、アルゴリズムの評価を試みる。上記の試みが順調に進捗した場合は、ゲノム配列の高次構造のモデル化に挑戦する。

＜成果公表リスト＞

本年度はなし。