多重ゲノム配列アラインメントに基づく機能情報の抽出

●後藤 修
京都大学大学院情報学研究科

＜研究の目的と進め方＞

多重配列アラインメントは、遺伝子あるいはその転写産物であるタンパク質や機能性 RNA の配列、構造、機能、進化の関連を知るための有力な手段であり、生命情報学の中心課題のひとつとして長年の研究が蓄積している。一方、塩基配列決定技術の急速な進歩に伴って、100 に近い真核生物を含む数多くの生物種について、その全ゲノム DNA 塩基配列がすでに解読され、さらに多くの配列が現在決定されつつある。これらのゲノム配列を相互に比較し、さらにはゲノム配列レベルでの多重アラインメントを作成することによって、それぞれのゲノム配列に内在する様々な機能情報を抽出することを本研究は目指している。

機能情報の抽出として最も重要なものは、ゲノム配列上の遺伝子領域を同定し、その内部構造（エキソン・イントロンの配置）を精度よく予測する「遺伝子発見」であろう。これとならんで、遺伝子の発現調節を担う領域や、mRNA のスプライシングに関わるシグナルを同定することもまた非常に重要な課題である。配列比較に基づく遺伝子発見と調節シグナルの同定は比較ゲノム学の中枢をなすものといえる。

しかし、ゲノム配列の比較を行うにはいくつかの本質的な困難が伴う。（1）ゲノム配列は通常のアミノ酸配列に比べ 10^4 ～ 10^6 倍も長い。（2）内部に転移、逆位、重複などを含み、アラインメントすべき領域の同定が一意的に決定できるとは限らない。（3）様々な物理化学的性質をもつ 20 種類のアミノ酸に比べ、4 種類の塩基間には際だった特徴差がなく、互いの性質の類似性をアラインメントに反映させることが難しい。（4）シスエレメント、エキソンなどある程度の幅をもった領域が機能の単位となる。これらの困難を克服し、高い精度と実用性を備えた多重ゲノム配列アラインメント法を開発するためには、従来法の改良に加え、これまでにない発想に基づく新規手法を開発することが必要であると考えられる。

本研究課題では、これらの困難を乗り越えるための方法を考案し、原核生物から哺乳動物に至るさまざまな大きさのゲノム配列の比較解析を現実的な計算機資源の下で可能とするソフトウェアの開発を具体的な目標とする。われわれはまた、選択的転写開始、選択的スプライシングのゲノム縦断的な検出法も開発しており、これらの制御シグナルに特に着目してその分子機構の解明に貢献したい。

＜2007 年度の研究の当初計画＞

本研究は、いくつかの段階的な過程を経て実現される。まず、2 つの長大なゲノム DNA 配列どうしのアラインメントから、両ゲノム配列上の対応関係（シンテニー）を明らかにする。複数のゲノム配列のすべての組合せについてペアワイズアラインメントを作成し、それら相互の整合性に基づいて多重ゲノム配列アラインメントの基点となるアンカー領域を同定する。そのようにして求めたアンカー領域に挟まれた領域に対してより精密な多重配列アラインメント法を適用する。そのようにして得られたアラインメントから特によく保存された領域を同定し、機能モチーフの候補とする。一方、cDNA, EST, 翻訳アミノ酸配列など転写産物の情報が得られる場合には、これらの配列をゲノム配列上にマップし、スプライスアラインメント法によってエキソン領域を正確に同定する。この「証拠に基づく（evidence based）」遺伝子発見法に加え、ゲノム配列間の保存性と *ab initio* 遺伝子発見法の原理とを組み合わせることによって、新規の（*de novo*）遺伝子領域の同定も試みる。

＜2007 年度の成果＞

本年度は次の 3 つの研究課題について、かなりの進展が見られた。

（1）転写産物のゲノム配列へのマップとスプライスアラインメントの高効率化。大量の cDNA や EST 配列それぞれに対応するゲノム領域を見出し、正確なエキソン－イントロン構造を推定するには従来大規模な計算機を必要とし、また精度の面で必ずしも満足できるものでなかった。今回我々は計算速度、必要記憶容量、アラインメント精度の多くの点で従来法を有意にしのぐ新規アルゴリズムを考案し、それを実装したプログラム *Spaln* (space efficient spliced alignment) を開発した。ヒトなどの長大なゲノム配列を対象とした場合でも *Spaln* は 1GB 以下の主記憶容量しか必要とせず、通常のパーソナル計算機で十分実行可能である。また、計算速度に関しても従来法とほぼ同等であった。特筆すべきは、問い合わせ cDNA/EST 配列が塩基置換や小規模の欠失・挿入などのノイズを含む場合でも、*Spaln* は正確にエキソン境界を見出すことができる点である。これにより、異生物種間のゲノム－転写産物間比較が容易になるものと期待される。なお、*Spaln* の実行サービスおよびソースコードの公開を、次ページに掲載するウェイブサイトで行っている [1]。

（2）多重配列アラインメントプログラム *Prime* の高速化。*Prime*[2] は我々が初めて提唱した二重反復改善法を実装した多重配列アラインメントプログラム *Prrn* を後継するプログラムで、長い欠失・挿入が存在する場合にも対応できる点に特徴がある。ベンチマークテストの結果、*Prime* は世界的に見て最も高精度な他の多重配列アラインメントプログラムに匹敵する精度を示すものの、実行速度に難があった。今回、基点設定（anchoring）とグループ化とを導入することにより実行速度の改善を試みた。その結果、平均 2% 程度アラインメント精度の低下が見られるものの、約 60% 計算効率を高めることができた [3]。現在、*Prime* は主にアミノ酸配列を対象としているが、プロモータ領域などゲノム上の非コード領域のアラインメントを高精度に行うためにも利用可能であると思われる。

（3）ブロック分割によるゲノム配列間アラインメントの効率化。すでに我々は 2 つのゲノム配列間のアラインメントを行うプログラムである *Alngg* を開発している 4)。しかし、*Alngg* の実行には大型計算機を必要とし、計算にかなりの時間を要するため手軽に利用することが困難であった。*Alngg* をはじめとして、これまで開発されてきたほとんどすべてのゲノム配列アラインメント用プログラムは、ある程度の長さの局所的な完全一致をまず求め、それを基点としてアラインメントの範囲を広げる「ボトムアップ」の方針をとっていた。今回我々が開発した CGAT (Coarse Grained AlignmenT) アルゴリズムはこれらとは逆の「トップダウン」戦略を採用している。すなわち、比較すべきゲノム配列をそれぞれ一定長のブロックに分割し、ブロック対ブロックの全組合せにつき類似性スコアを高速に計算する。このスコアを基に、ブロックレベルでの粗いアラインメントをまず求める。そのようにして求まったブロック間アラインメントに含まれる領域に限って

塩基レベルでの細密なアラインメントを通常の方法で求める。この２段階の（場合によってはさらに段階的な）手続きの有効性をふたつのバクテリア全ゲノム配列比較により検証した。その結果、ゲノム配列比較に現在最も広く利用されているプログラムである *Blastz*[5] と比較して、CGAT 法はアラインメントの感度を損なうことなく必要な記憶容量と計算量を大幅に削減できることが確認できた[6]。

＜国内外での成果の位置づけ＞

いくつかの検証の結果、転写産物のゲノム配列へのマッピングおよびアラインメントツールである *Spaln* は、この目的のために現在最も高性能であると考えられる *Gmap* や、これまで広く世界中で利用されてきた *Blat, Megablast, Sim4* などの有名なプログラムに比べて性能上明らかな優位性を示した。したがって、現時点では世界最高性能をもつといえる。すでにソースコードを公開しており、また近々論文発表も予定されているので、国内外からの広い利用が期待される。

多重配列アラインメントプログラムである *Prime* についても、現在もっとも高性能であるとされる *Mafft* や *Probcons* と精度の点では肩を並べる。ただし、今回の改良により約 60% の高速化が達成できたものの、他のプログラムに比べてまだ実行速度が劣っており今後いっそうの改良が必要である。また、ゲノム配列を対象とした場合のパラメータ値の最適化を行う必要がある。

最後に、ゲノム配列間アラインメントのための CGAT 法は我々独自の方法であり、その有効性についてこれから詳しく検討する必要がある。しかし、現状においても *Blastz* に優る計算効率とアラインメント精度を持つことを確認しており、さらなる改良を加えることによって世界有数のソフトウェアに育つ可能性を秘めている。

＜達成できなかったこと、予想外の困難、その理由＞

2006 年度に初めて本課題が採択された当初は多重配列アラインメントから機能モチーフを抽出する方法の開発に重点を置いていた。しかし、その前提となる多重配列アラインメントの作成法に時間を要し、現在まだ開発段階である。一方、配列アラインメントに基づく遺伝子発見法である *Alngg* の改良を本年度も継続した。しかし、昨年度の報告書にも記述したように、隣り合う遺伝子間境界の推定が困難であるという問題をまだ解決できていない。これについては抜本的な方法の見直しを考える必要がある。

＜今後の課題＞

ブロック分割法によるゲノム配列間アラインメントが予期通りの性能を示すことを確認できたため、この手法を多重ゲノム配列アラインメントに拡張することが今後の最も重要な課題である。遺伝子境界の推定問題は、DBTSS の内容や *Spaln* の実行結果、あるいはヒストン修飾などの外部情報を組み入れることで現実的な解決を目指す。一方、アミノ酸配列を問い合わせとした場合に、対応する翻訳領域をゲノム上から高速に見いだせるように *Spaln* の拡張を試みる。その結果とアミノ酸配列の保存性に基づくゲノム横断的な遺伝子予測法である *FamilyWise* とを組み合わせることにより、選択的転写開始・選択的スプライシングを含むゲノム横断的でより包括的な遺伝子予測法を確立したい。

参考文献

1) 0801222000.
2) Yamada, S, Gotoh, O, Yamana, Y: Improvement in accuracy of multiple sequence alignment using novel group-to-group sequence alignment algorithm with piecewise linear gap cost. *BMC Bioinformatics*, 7, 524 (2006).
3) 0801231213
4) Gotoh, O. *et al.*: Discovery of protein coding genes through chromosome-to-chromosome sequence comparison. *Genome Informatics* 2005, 12.19-21 (2005).
5) Schwartz, S. *et al.*: Human-mouse alignments with BLASTZ. *Genome Res.*, 13, 103-107 (2003).
6) 0801181609
7) Wu, T.D., Watanabe, C.K.: GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21, 1859-1875 (2005).

＜成果公表リスト＞

1) 論文／プロシーディング

1. 0801181609
   Nakato, R. and Gotoh, O. A novel method for reducing computational complexity of whole genome sequence alignment, in Proceedings of the 6th Asia-Pacific Bioinformatics Conference, Kyoto, p.101-110 (2008).
2. 0801231213
   Yamada, Y., Gotoh, O., and Yamana, H. Improvement in speed and accuracy of multiple sequence alignment program PRIME, IPSJ Trans. Bioinf. in press

2) データベース／ソフトウェア

1. 0801222000
   Spaln: A space-efficient and accurate tool for mapping and aligning a set of cDNA sequences onto a genomic sequence
   URL: http://www.genome.ist.i.kyoto-u.ac.jp/~aln_user/