# 論文の内容の要旨

論文題目　Research on Fast Algorithms for Comparison and Indexing of Biological Sequence Information

生物配列情報の比較と検索のための高速なアルゴリズムの研究

氏名　渋谷哲朗

近年、分子生物学において配列解析技術などの急速な進展にともない、DNA、RNA、蛋白質といった生物配列情報が爆発的に増大しており、そのため、そういった大量のデータを効率的に処理するための様々な高速なアルゴリズムが必要となってきている。これらの問題において最も重要な基盤技術の一つは、配列のアラインメント技術やモチーフ発見技術、DNA や蛋白質などの生物データベースの検索技術などの配列比較及び検索の技術である。これらの技術は、パタンマッチング技術とも呼ばれる。本論文では、そういった配列比較や検索などのパタンマッチングの問題に対し高速なアルゴリズムを提供することで、分子生物学上非常に重要ないくつかの問題を効率的に解決する。さらに、それらのアルゴリズムの多くについて、実際に大規模な生物学的データを用いた実験を通じて、性質及び性能を検証する。

本論文では、まず、分子生物学全般で極めて多用されている最も基本的かつ重要な配列比較技術の一つであるマルチプル・アラインメント問題をとりあげる。この問題には、計算論的に得られる最適解が生物学的には最適ではない、という問題がある。これに対し本論文では２つの手法を提案する。まず一つ目の手法として、多くの準最適解を出力することで代替の解を提供する、という解決法を考えられるが、そのような準最適解はわずかに異なるだけの解が極めて多くあり、欲しい解を探すのが困難である。そこで、本論文ではいかなるアラインメントが列挙する必要がないかを論じ、必要なアラインメントだけを高

速に列挙するアルゴリズムを提案する。また、このアラインメント問題においては、スコア行列などのパラメータを固定して求めた従来の最適解は生物学的に最適であるとは限らないともいわれている。そこで本論文ではさらに二つ目の手法として、パラメータを変化させた時の解をすべて得るための効率的なパラメータ空間の探索技法について議論する。

次に本論文では、先に述べたアラインメントの一変形であるスプライスト・アラインメントという手法を用いて、cDNA ライブラリに含まれる配列集合を選択的スプライス集合と呼ぶ集合にクラスタリングするという問題を扱う。最近ヒトゲノムの遺伝子数と実際の蛋白質数の違いがわかってきたことで非常に脚光を浴びている選択的スプライシングの研究において、このクラスタリング結果は極めて有用である。また、本論文でも次に扱う遺伝子発見のツール等の学習セットとしても有用である。本論文では、この問題を解くための高速かつ正確なアルゴリズムを提案する。さらに、マウスの大規模な cDNA ライブラリである FANTOM を用いた実験を通して、このアルゴリズムの性能を検証する。

さらに、分子生物学研究上最も重要な問題の一つである、遺伝子発見あるいは遺伝子同定と呼ばれる問題を扱う。この問題に対しては、従来から様々な手法が提案されてきているが、それらは主に 2 種類に分類することができる。一つは隠れマルコフモデルなどの統計学的な手法を用いる方法であり、もう一つは異なる種の間の配列比較やデータベースからの類似配列検索に基づくパタンマッチング的手法である。前者は統計的な傾向が種によって異なるため、自身の種かそれに近い種の十分な学習セットがない場合うまく推定できない、という問題がある。一方、後者は類似しているものがないものについては全く見つけることができない、などの問題がある。それに対し、本論文では、両者の長所をあわせ持つような従来にはない全く新しい手法を提案する。この手法は、きわめて大規模なパタンデータベースのパタンを検索し、それらのパタンの統計的振舞いに基づいて遺伝子発見を行なう。この方法は、全く新しい学習セットのないような種のゲノムに対しても遺伝子を従来手法以上に正確に推定することができることが可能である。また、一般的にはパタンマッチング手法による遺伝子発見は統計的な手法より計算時間が大きいことが多いという問題点があるが、この手法では極めて大きなパタンデータベースからの高速検索を行なうための新しいパタン検索アルゴリズムを提案することで、一般的な統計的手法と比べても十分高速な遺伝子同定を実現している。本論文では、さらにこの手法を実際に様々な原核生物ゲノムに適用し、検証を行なう。

RNA などの生物配列は、それが形作る立体構造がその性質に極めて影響を及ぼすことが知られている。本論文では、最後に、RNA などの生物配列や 2 次構造データベースにおいて、接尾辞木というデータ構造を用いて、頻出する構造を高速に捜し出す手法を提案する。接尾辞木は、テキストなどから頻出文字列を発見したり、キーワードを検索する際に非常に重要かつ有用な検索のためのデータ構造である。本論文では、まず、同様の構造を持ち得るような RNA の部分配列とは何かを考察し、それを考慮するように接尾辞木を一般化したデータ構造とそれを構築する高速なオンライン・アルゴリズムを提案する。また、RNA

の２次構造は木構造で表現できることが知られている。さらに、そのような木構造から様々なパタンを検索あるいは発見することができる木に対する接尾辞木というものも知られているが、本論文ではそれに対する既存の最良の計算量より良いアルゴリズムを提案する。

本論文では、これらのアルゴリズムを通じて、配列比較や検索といった効率的なパタンマッチング的技法を用いることで、いかに分子生物学における様々な重要な問題を解決することができるかを示す。