論文審査の結果の要旨

氏名　木村　宏一

本論文は、近年その驚異的な進歩によって生命科学に大きなインパクトを与えている次世代型 DNA シークエンサーから産出された塩基配列断片が、ゲノム中で対応する位置を高速で検索する「塩基配列マッピング問題」に関する研究結果をまとめたものである。理想的には断片配列と一致する部分配列がゲノム上のどこかに存在するはずであるから、高速な部分文字列マッチングアルゴリズムを適用すれば済むはずであるが、実際には断片配列は無視できないほど高い実験エラーを含み、参照されるゲノム配列も断片配列をうみだしたゲノムと同一ではないので、文字間の不一致などを考慮する必要があるし、断片配列が転写産物からとられた場合は、イントロンの存在も無視できない。このような事情を考慮したアルゴリズムは、従来でも BLAST を始めとして、数多く開発されてきたものの、次世代シークエンサーの生み出すデータは、その量が桁違いに多く、実験エラー率の高さや断片長の短さをも考慮にいれて最適化したアルゴリズムを開発する必要がある。すなわち、計算速度と記憶容量の両面での最適化が望まれる。実際ここ数年、この分野は非常に活況を呈しており、次々と新しいアルゴリズムやツールが発表されているが、我が国からの貢献は少ないという状況下で、論文申請者の以下に述べる貢献は、世界的に先駆的なものを含む、貴重なものであると言える。

論文の主な内容は、第一章の序論に続く、３つの章に分かれている。

第二章では、ゲノム塩基配列への高速マッピングに適した手法として、近年サフィックス・アレイなどの従来法に代わって注目されている Burrows-Wheeler 変換に基づくマッピング用の効率的なデータ構造として、階層的バイナリ文字列を提案している。一般にゲノムのマッピング問題は Burrows-Wheeler 変換されたゲノム配列上でランク関数やセレクト関数を計算する問題に帰着させることができるが、このデータ構造を用いれば、これらの関数を比較的少ないメモリ使用量で高速に計算できる。その性能は計算機実験でも確認している。

第三章では、ペアエンド方式の配列データを効率的に扱うためのデータ構造として、局所化サフィックス・アレイを提案している。ペアエンド法は、次世代型シークエンサー等から得られた配列がしばしばユニークにマッピングできない状況、特にゲノムがリピート配列を多く含んでいる場合の困難、を克服するためによく用いられる方法で、ほぼ一定の長さの塩基配列の両端の配列を決定する。このため、二つの読み取られた配列（リード）は、ある距離の範囲にマップされなければならないという制約条件がつくが、この条件をサフィックス・アレイや、Burrows-Wheeler 変換などにおいて、効率的に取り扱うことは困難であった。申請者らが提案したデータ構造を用いれば、この問題をうまく解決でき、そのことは実データを用いた計算機実験でも実証している。

第四章では、マップされる配列のエラー率が高く、ゲノム配列とのミスマッチが多い場合に適したマッピングアルゴリズムを提唱している。このような状況は、たとえばライフテクノロジー社の SOLiD シークエンサーのデータを扱うときに問題になる。SOLiD では、隣接２塩基の配列を４色で表現する２塩基符号化法を用いているため、データの復号化の過程でエラーを訂正できるものの、元データのリード配列のエラー率は比較的高くなってしまう。一方、BLAST をはじめとする、ミスマッチを含む配列のアラインメントを高速に求めるアルゴリズムの基本戦略は、まず調べたい配列の短い部分配列（シード配列）がターゲット配列中に完全一致する場所を高速に検出し、その後、その周囲をじっくり調べることである。検索配列が短くてエラーを多く含んでいる場合、このシード配列をどう設定するかが問題になる。その解決策として、シードの中にマッチを調べない場所を設ける穴開きシードを用いる方法が広く用いられているが、メモリ使用量が多くなるという欠点をもつ。そこで申請者は、本章で等エントロピー分割法という方法を用いて、シード配列を柔軟に求めることを提唱している。この方法の原理は、検索配列からシードをとるときに、ターゲット配列中に候補位置が何個ぐらいかが大体同じになるように、その長さを調節するというものである。申請者は、この方法を、穴開きシードを用いた BFAST プログラムと実データを用いて比較したところ、速度と感度を犠牲にせずに、大幅なメモリの節約を実現できることを確認した。

以上、記してきたように、本論文に記された申請者の研究内容は、いずれも生命科学の最新の課題の解決に向けた、オリジナルな貢献を示すものであり、学術的に評価できる。従って、博士（生命科学）の学位を授与できると認める。