論文審査の結果の要旨

氏名　曲　薇

近年の技術革新により、高度に並列化された超高速 DNA 配列解読装置が開発され、ゲノム DNA 配列の読み取り速度が 100 倍以上高速化された。これにより、ゲノム DNA 配列、発現情報を反映した mRNA 配列などが網羅的に取得できるようになり、生命科学に大きな変革がもたらされつつある。

現在の超高速 DNA 配列解読装置は、その並列性、高速性を利点として持つ一方、1 本のリードが短いという問題点を抱えている。リード長が短く、かつ読み取り誤りを含むため、そのクラスタリングや参照ゲノム配列への貼り付けが困難で、貴重な大量配列データを十分に活用することが難しいことが大きな問題となっている。

参照ゲノム配列がある場合、従来の手法では、個々のリードを直接参照配列と比較していたため、一意に張り付かなかったり、張り付く場所がなかったりするリードが大量に捨てられる結果となり、超高速 DNA 配列解読装置の性能を十分に引き出すことができなかった。

本論文では、ゲノムへアラインメントする前に配列の頻度情報を利用してクラスタリングする手法を提案した。頻度がより高いリードと類似の頻度の低いリードは、その多くが読み取り誤りである可能性があるため、配列情報と頻度情報から読み取り誤りを反映した配列の親子関係をモデル化し、クラスタリングを行っている。クラスタリングと代表配列の検出においては、ハッシュ法を活用し、高速な線形時間アルゴリズムの構成に成功した。この結果、転写開始点や small RNA 配列中のミスを補正し、全配列の約 5%を新たに正確な位置へアラインメントできることを示した。

本論文が提案した手法は、現在の超高速 DNA 配列解読装置を活用するために必要な技術であると認められる。将来の DNA 配列解読装置の進歩によってリード長の問題は解決されるが、短い非コード RNA の解析においては短いリードの扱いは今後も重要な課題であるから、本論文の成果の重要性は一時的なものではないと考えることができる。

以上の点から、本論文は生命情報科学に貢献する重要な貢献であると認めることができる。