審 査 の 結 果 の 要 旨

氏　　名　小野　幸輝

次世代シークエンサーと呼ばれる超高速シークエンサーは、初期には１００塩基以下の短いリードしか読むことができなかったが、ロングリードシークエンサー近年は数千塩基を読むことができるが広く使われるようになり、数十万塩基に達するロングリードも得られるようになっている。ショートリードよりも多くのエラーが含まれることが知られているロングリードを対象とする様々なツール・アルゴリズムが開発されているが、正確な塩基配列が不明のソースから得られる実リードから、それらのツール・アルゴリズムの性能を評価することは困難である。本論文では、ロングリードの統計的特徴を理解するための解析行うとともに、ロングリードを用いたツール・アルゴリズムの評価に用いることのできる２つのロングリードシミュレーターを開発した。

１つ目のシミュレータである PBSIM においては、PacBio のロングリードのデータセットからリード長およびシークエンシングエラーの統計解析を行ない、それらの確率分布を推定した。シミュレーションでは、確率分布推定の結果に基づいてリードを生成する。対数正規分布に基づいてリード長を決めて参照配列からランダムに部分配列を取り出し、各配列位置の quality score を正規分布（CLR）および指数分布（CCS）に基づいて指定し、その quality score に基づいてエラーがランダムに導入される。この方式により、PacBio のロングリードが持つ長さおよびシークエンシングエラーの確率分布を正確に模倣したロングリードを生成することができた。また、サンプルベースのシミュレーションでは、PacBio の実リードライブラリからサンプルされたリード長、quality score を用いたリード生成を行うことに対応している。PBSIM を用いて作成した PacBio のシミュレーションデータを用いたハイブリッドアセンブリのテストにより、ハイブリッドゲノムアセンブリを精度高く行うためには、CLR では少なくとも 15、CCS では少なくとも 30 のカバレッジが必要であることが示された。PBSIM は 2012 年にリリースされ、PacBio のロングリードを用いた様々な研究に使われ、その論文（2013）年は 200 回以上引用されている。

コンテクスト特異的なエラーは、ショートリードのシークエンサーでは良く見られるが、PacBio では比較的少ないため、PBSIM では考慮されなかった。しかし、Nanopore リードにはホモポリマー特異的なエラーがあり、また PacBio のエラーには領域的な偏りがあることが知られている。これらの点を考慮したリードをシミュレートするために開発された PBSIM2 においても、リード長およびリード全長のエラー率は実データの統計解析から得られた分布からサンプルされる。リードの位置ごとの quality score は、モデル選択基準を組み合わせた隠れマルコフモデル（FIC-HMM）によって生成される。FIC-HMM は、モデル選択基準によって HMM のパラメータだけでなく状態数も同時に最適化できる生成モデルである。PBSIM2 のリードも、リード長の確率分布に基づいて参照配列からランダムに切り取られ、配列に、FIC-HMM によって生成された quality score に基づいたエラー（挿入、欠失、置換）を導入して生成される。Pac Bio シークエンサーのリードでは、quality score は隣接する配列位置でも広い範囲で変動するため、FIC-HMM によるモデル化がよく適合するが、Nanopore シークエンサーのリードでは、

quality score は近隣の配列位置で高い相関を持つため、単純なマルコフ連鎖でも一定の精度でモデル化できることが示された。リード全長のエラー率によってグループ分けしたリード群それぞれについて、セグメントごとのエラー率の分布を詳細に解析した結果、PBSIM2 は PacBio および Nanopore シークエンサーのエラー分布をよく再現しており、従来のリードシミュレータよりもエラーの不均一性を正確にシミュレートすることが示された。

本論文で開発された PBSIM および PBSIM2 は、ロングリードを対象とする様々なツールがその論文でツール評価のために用いており、これらのツールはロングリードシークエンサーによる生命科学の研究に不可欠なツール群の開発に大きく貢献している。

よって本論文は博士（科学）の学位請求論文として合格と認められる。

以上１８５８字