大規模計算システムを用いたがん体細胞変異検出アルゴリズムの研究

審 査 の 結 果 の 要 旨

氏　　名　　上田　宏生

本論文は、がん体細胞変異を検出する新しい方法を示したものであり、大規模な計算環境を用いて、迅速に且つ正確に、がんにおけるコピー数変異、点突然変異、また、腫瘍率を検出することに成功したことを示した論文である。

がんはゲノムの異常に起因する疾患であり、それゆえ、癌種や個人により異なるゲノム上の変異を特定することは、がんを理解するうえで極めて重要である。また、次世代シーケンサとエクソームシーケンスを用いた、がんゲノム解析手法が近年急速に進歩し、多数の体細胞遺伝子変異を網羅的に解析することが可能となった為、がんゲノムの解析に活用されている。しかし、シーケンサが産出するデータ量が膨大であることと、シーケンサリード中のエラーの存在や、癌サンプル中の非腫瘍細胞の混入などの問題があり、変異の解析を難しいものにしている。本論文では、並列処理を用いて処理を高速化するとともに、低腫瘍率でエラーの多いサンプルや、複雑なコピー数変異をもつサンプルにおいても正確に体細胞変異を検出する方法を検討しており、研究対象として意義ある課題を実施している。

第２章では、次世代シーケンサデータを大規模な計算環境を用いて並列計算するための新しい方法が示された。次世代シーケンサのデータ解析は、データ量が膨大である上に、ワークフローが複雑であり、解析に時間要することが大きな課題となっている。本論文の方法では、従来のファイルフォーマットとソフトウェアとの互換性を維持しつつ、ゲノム変異の検出、遺伝子発現定量といった、次世代シーケンサデータ処理の大規模な並列化を可能とした。また、複雑なフローを外部ファイルとして定義できるようにした点や、広く普及している資源管理 API を使用することで汎用的な使用が可能となった。コア数を増加させて行ったパフォーマンステストでは、50 コア並列でおよそ 40 倍の処理の高速化を達成し、従来、数週間程度必要であった解析の処理時間を数時間程度にまで短縮したことが示され、次世代シーケンサのデータ処理を迅速化する有効な解決手段であることを示した。また、12 テラベースの実リードの解析を行うなど、実用的な運用を可能にしていることを示した。

第３章では、エクソームシーケンスデータから、染色体の絶対数でのコピー

数変異、点突然変異、腫瘍率を検出するための新手法が示された。コピー数変異の検出方法に関しては、連続ウェーブレット変換を用いて倍数体性のピークを検出し、また、アリルのインバランスを測定した値を、理論上の倍数体性と腫瘍率のマトリックスに適応することにより、整数値のコピー数と腫瘍率を検出する方法が示された。本手法は、既存の手法では用いられることのなかった方法を適応したという点で新規性を有する。また、点突然変異の検出時にノイズ分離の為に EM 法を用いる点も、従来行われていない手法であり、新規性を有する。さらに、既存の手法ではエクソームデータを用いて、アリルごとのコピー数解析を行うことはできないが、本論文の手法では、SNP アレイのデータと比較しても良好な結果が得られることが示され、絶対数のコピー数変化も、実用的な腫瘍率の範囲で高い精度で検出できることが示された。

実際のシーケンサリードを用いた単体細胞変異とコピー数のシミュレーションを行う方法を開発し、シミュレーションリードを用いた検証が行われた。点突然変異の検出において、高い検出感度（偽陰性率 2%）と検出精度（偽陽性率 2%）を達成し、従来の手法と同等の検出感度の場合は、半分以下の偽陽性率で精度と感度の高い検出が可能であることを示した。これらの結果は、特に腫瘍率が低いサンプルにおいて、従来問題となっていた検出感度と検出精度の問題を、本研究の方法が大きく改善したことを示している。

以上の研究成果は、本論文の方法ががん体細胞変異の検出において非常に有用なツールであることを示したものであり、特にコピー数変異、点突然変異、腫瘍率を正確に、一度に解析できるという点でも評価できるものであり、すでに、国内の 5 研究機関において本研究の手法が活用されている。今後、本論文の手法のさらなる活用によってがんゲノム解析の進展が期待されるものである。よって本論文は博士（工学）の学位請求論文として合格と認められる。