論 文 の 内 容 の 要 旨

論文題目　Design of Bayesian Hierarchical Models for Accurate Detection of Somatic Mutations（高精度な体細胞変異検出のための階層ベイズモデルの設計）

氏　　名 森山　卓也

癌はゲノムの変異により起きる病気である。細胞は、タバコ、アルコール、紫外線、酸化ストレス、感染症などの刺激を受け、日々ゲノムに変異を蓄積させては、DNA 修復系による修復を繰り返している。DNA 修復を免れ、後天的に蓄積したゲノムの変異はやがては細胞の機能に異常をもたらし、癌を引き起こす。癌はゲノムの変異を原因とする病気であることから、ゲノム変異の情報は、癌の研究や治療において、不可欠な情報である。癌研究においては、体細胞変異の情報を用いて癌の進化の過程を推定し、これをもとに新たな治療方針の模索が進められている。また、癌治療においては、次世代シークエンサー（NGS）技術の発展に伴い、低コストでゲノム情報を取得できるようになったため、NGS データから検出したゲノム変異の情報から患者ごとに治療方針を提案する癌ゲノム医療が現実に推し進められている。そのため、NGS データから高精度にゲノム変異を検出する手法の開発は癌ゲノム分野における重要課題の一つである。

後天的に起きたゲノムの変異は体細胞変異と呼ばれる。通常、体細胞変異を NGS データから検出する際は、腫瘍由来のシークエンスデータと正常組織由来のシークエンスデータがそれぞれ少なくとも一つ以上利用される。体細胞変異を検出する方法としては大きく二つの方法があり、一つ目は腫瘍一検体のシークエンスデータに基づく方法で、二つ目は多検体の腫瘍に基づく方法である。一検体のシークエンスデータに基づく方法では、腫瘍一検体と一つの対応する正常細胞のシークエンスデータが用いられ、多検体の腫瘍に基づく方法では、多検体の腫瘍と一つの対応する正常組織のシークエンスデータが用いられる。体細胞変異検出手法の性能改善には、NGS データ特異的な性質や生物学的な事前知識の適用が重要と報告されているが、単一検体に基づく方法、多検体に基づく方法の両方で十分に活用されていない。

単一検体に基づく方法に関しては、NGS データ特異的な性質を利用するために、階層ベイズモデルを基に検出手法が開発されてきた。しかし、これらの既存手法における階層ベイズモデルにおいては、単一の性質のモデル化に焦点を当てており、複数の性質を同時に考慮する設計は為されていない。

多検体に基づく方法に関しては、体細胞変異が共有される性質や癌の進化系統樹のもつ性質などの、生物学的な事前知識の利用に注目が置かれている。まず、体細胞変異が

共有される性質を利用する手法に関しては、少なくとも一つの検体が変異をもつ場合、検出の閾値を下げると精度が改善できるという知見をもとに統計モデルが設計されている。ここで、変異が共有される性質を利用するには、少なくとも一つの検体が変異をもつことを高い確度で判定することが重要であり、検出される候補変異数と、検出の特異度が重要であると考えられる。しかしながら、既存の統計モデルでは変異数のみを利用し、検出の特異度までは考慮されていない。さらには、既存の統計モデルの設計の問題により、NGS データ特異的な性質は利用できない。次に、系統樹の性質を利用する手法に関しては、相反する性能評価の報告が上がっており、一部の論文では性能が悪いとして報告する一方で、他方では性能が高いと報告されている。そのため、系統樹の性質が体細胞変異検出にとって有用な性質かどうかはそもそも十分に考察されていない。

以上のことから、既存手法においては、NGS データ特異的な性質や生物学的な事前知識の適用は十分になされておらず、体細胞変異の検出精度には依然として改善余地があると期待される。本学位論文においては、NGS データ特異的な性質や生物学的な事前知識の適用方法を考案し、高精度な体細胞変異検出手法を提案する。

まず、一検体腫瘍に基づく方法に関して、体細胞変異の検出を行う手法 OHVarfinDer を提案する。既存研究において、NGS データ特異的な性質を複数同時に統計モデルに加味する階層ベイズモデルの方法に関しては十分な研究がなされていなかった。この点に関し、我々の提案手法では分割に基づくモデル統合方法により、明示的に複数の階層ベイズモデルを一つの階層ベイズモデルとして統合する。この方法では、各観測変数に対し、観測を生成した統計モデルを示す新たな観測変数を導入することで、複数の階層ベイズモデルの統合が可能になる。この統合方法はベイズモデル平均化と異なり、ベイズファクターの計算において、分子と分母で重みパラメータが等しい場合では、事前に重みパラメータなどの設定が不要である。我々は、シミュレーションデータと実データに基づき、提案手法の評価を行った。シミュレーションデータによる評価では、単一の性質が利用できる場合では他の既存手法と同程度の性能を示し、複数の性質が利用可能な場合においては既存手法を上回る性能を示した。実データに関しては、TCGA のベンチマークデータに基づく評価を行い、ほとんどの場合で既存手法を上回る性能を示した。

次に、多検体腫瘍に基づく体細胞変異検出手法 MultiMuC を提案する。変異共有の性質の利用においては、既存手法では検出される変異候補の数に着目しているが、変異検出の特異度や NGS データ特異的な性質は考慮していない。変異検出の特異度を利用するために、提案手法では二種類の潜在変数を導入する。一つ目の潜在変数は少なくとも一つの変異候補が検出されているかを表し、二つ目の潜在変数は変異候補が高い特異度で検出されていて、変異候補数も十分多くある状態を表す。提案手法では、これらの潜在変数の導入によって、検出された変異候補の数と検出特異度の両方を利用する。また、NGS データ特異的な性質を利用するために、そのような性質を加味した変異検出手法の確率モデル内のデータ生成確率の利用に着目した。通常、それらの変異検出手法からは、

ベイズファクター、ないしベイズファクターに変換可能な事後確率のみが得られ、データの生成確率は直接利用することはできない。我々は、このようなベイズファクターしか得られない状況においても、事後分布の推定や最大事後確率推定には影響が無いことを示した。このアイデアから、変異検出手法の出力として得られるベイズファクターをもとに階層ベイズモデルを構築した。そのため、提案手法では、既存の変異検出手法内のデータ生成確率を通じて、NGS データ特異的な性質を利用することができる。我々は、実データに基づくシミュレーションにより、提案手法の性能評価を行った。このシミュレーションでは、複数の癌の系統樹構造とクローンの混合比率を用意することで、癌のシークエンスデータを複数生成した。この性能評価によって、我々の手法は、多数の検体において変異が共有されていることを利用し、既存手法の精度をさらに改善可能であることを示した。

最後に、がんの進化系統樹の性質が体細胞変異検出に対して有用かどうかを考察する。この考察では、変異検出の結果を生成する確率モデルに仮定をおいた元で、系統樹を用いて変異検出を行う手法と、系統樹を用いずに変異検出を行う手法の感度と特異度の期待値を評価した。また、系統樹を用いた検出手法の方が高い特異度を示すための十分条件を導出した。これらの評価から、どのような状況下でがんの進化系統樹が変異検出に有用かを明らかにし、特定条件下において変異検出の精度向上に有用たり得ることを示した。