

審査の結果の要旨

氏名 佐藤 建太

細胞ごとに RNA の塩基配列を網羅的に決定し、発現量を解析する一細胞 RNA-seq が大規模に行われるようになり、従来にない新しいデータ解析手法が求められている。本論文は、大規模化した一細胞 RNA-Seq における類似細胞検索と RNA 発現組成の推定について記述したものである。本論文は 4 つの章から構成されている。

第 1 章では、これまでの RNA 定量技術を概説しながら一細胞 RNA-Seq の位置づけを明確にし、一細胞 RNA-Seq の主要な実験技術と解析技術を紹介してから本研究の意義を説明している。

第 2 章では、一細胞 RNA-Seq で得られた発現プロファイルのリファレンスとしてデータベースを構築し、それとは別に得られた発現プロファイルをクエリとして発現パターンが類似した細胞を検索する手法を提案している。この手法は、発現プロファイルをビットベクトルとして要約することで省メモリ化しつつ、ビットベクトルに検索用の索引を構築して高速な検索を可能にした。この手法を実装したソフトウェアは `CellFishing.jl` としてパッケージ化され、オープンソースソフトウェアとして公開されている。`CellFishing.jl` の性能は、4 つの実データを用いて先行研究である `scmap` と比較されている。その結果、`CellFishing.jl` が先行研究で開発された `scmap` と比較して検索の精度で同等以上であり、データベースの構築時間や検索速度では大幅に上回ることが示されている。また、細胞間の類似性を詳細に評価するためにクエリとリファレンスの細胞間で発現変動遺伝子を検出する機能を実装し、実データを使ってそれが生物学的に妥当と考えられる発現変動遺伝子を検出できることを実証している。さらに、100 万細胞を超える規模の巨大なデータセットについても `CellFishing.jl` が省メモリかつ非常に高速に類似細胞検索を行えることを示している。

第 3 章では、一細胞 RNA-Seq において Unique Molecular Identifier (UMI) により発現量の増幅バイアスを排除したデータから RNA 発現組成を Bayes 推定する手法を提案している。この手法では組成を内積空間の元とみなす

Aitchison 幾何学を基礎としており、シンプレックス上の正規分布を導入して任意の正規分布を事前分布として設定できるようにしている。事前分布のパラメータをデータから経験的に推定する手法についても検討している。RNA 発現組成の Bayes 推定は、大規模データにも適用可能にするため、事後分布をシンプレックス上の正規分布で近似する変分推定法で行っている。開発した手法はトイモデルによるシミュレーションと実データを用いて検証され、期待した通りに Bayes 推定が行えることを実証している。

第 4 章では、第 2 章と第 3 章の研究を振り返って全体の総括をし、本研究の今後の展望を述べている。

以上、本論文は一細胞 RNA-Seq の大規模なデータセットにも適用可能な類似細胞検索と不確実性を考慮した組成推定の手法を提案したものである。その結果には学術上・応用上の貢献が認められる。よって、本論文は博士（農学）の学位請求論文として合格と認められる。