審 査 の 結 果 の 要 旨

氏 名　　張 子龍

単一細胞 RNA シーケンス（scRNA-seq）は、一度に多数の細胞の RNA 発現データを得ることができる手法であり、この手法によりこれまで新しい細胞型やサブタイプの発見などがなされてきた。しかしながら、得られる細胞あたりのリード数が相対的に少ないため、実際には発現している遺伝子に対してリードが割り当てられないことに起因するドロップアウトと呼ばれる問題や、全体的に高レベルのノイズが観察されるという問題が存在する。本論文は、これらの問題に対応するために開発されてきた多くの前処理法や次元圧縮（視覚化）法の性能評価を行い、データの性質に応じた活用法と今後の展望を述べたものであり、4 つの章から構成されている。

第 1 章では、本研究の背景や目的、そして意義を述べている。scRNA-seq データ解析の歴史とともに、本研究で入力として用いるカウントデータ行列を得るまでの一般的な手順を概説している。また、性能評価に供する方法を三つのカテゴリに分け、それぞれのカテゴリに含まれる方法を概説している。第一カテゴリには、次元圧縮（視覚化）のみを行う 2 つの方法として、主成分分析（PCA）および t 分布型確率的近傍埋め込み法（t-SNE）が含まれている。第二カテゴリには、前処理から視覚化までを一つのプログラム内で行う 4 つの方法（ZIFA、PHATE、CIDR、そして MAGIC）が含まれている。そして第三カテゴリには、主に前処理のみを行う 9 つの方法（SAVER、SAVER-X、scImpute、DCA、autoImpute、DrImpute、LSimpute、kNN-smoothing、scRMD）に加えて、新規に開発した次元圧縮まで行う方法（DAE）の計 10 個が含まれている。第三カテゴリの方法は独自の視覚化手段を持たないため、本研究ではこれら 10 個の前処理法に t-SNE を連結して視覚化まで行うパイプラインとして性能評価している。

第 2 章では、性能評価に用いた scRNA-seq データおよび計 16 パイプラインの詳細を述べている。scRNA-seq データは、シミュレーションにより生成されたデータ（シミュレーションデータ）とリアルデータの 2 種類を用いている。 シミュレーションデータは、R パッケージ Splatter を用いて生成した 8 種類のデータセット、および R パッケージ powsimR を用いて生成した 9 種類のデータセットを用いて性能を評価している。また、リアルデータについては、計 15 種類のデータセット（ヒト由来の 5 個とマウス由来の 10 個）を取得し、原著論文から得た細胞型情報を正解として評価を行っている。リアルデータは、56～3,005 個の細胞数、3～32 種類のグループ数、そして 19,020～41,480 個の遺伝子数から構成されている。方法については、新規開発したオートエンコーダに基づく方法の背景となるニューラルネットワークの原理、

DAE モデルのフレームワーク、そして学習手段について述べている。また、計 16 のパイプラインの大まかな特徴をまとめるとともに、評価基準として用いた 3 つの指標（ARI、NMI、そして HOMO）について述べている。

第 3 章では、結果と考察を述べている。Splatter のシミュレーションデータでは、kNN-smoothing、SAVER、SAVER-X の 3 つの手法が高い性能を示したと述べられている。次に、powsimR のシミュレーションデータでは、PHATE が全体として最高性能を示し、特に細胞数が多い場合（2,000 個程度）に有効であることが示されている。また、CIDR は全体としては次点ながら、細胞数が 200 個程度と比較的少ない場合に特に有効であることが示されている。また、遺伝子数次第で性能に大きな違いが出ることも明らかにされている。具体的には、30,000 遺伝子の場合は DrImpute と scImpute、20,000 遺伝子の場合は PHATE、10,000 遺伝子の場合は kNN-smoothing が高い性能を示したと述べている。Splatter のシミュレーションデータで kNN-smoothing が高い性能を示したことは、遺伝子数が少ないことによる可能性が高いという考察も述べている。

計 15 個のリアルデータ解析結果では、DrImpute が最高性能を示し、PHATE がそれに続いていることを示している。この結果は、powsimR のシミュレーション解析結果と一致しており、全体として入力データ中の遺伝子数に応じて手法を使い分けるほうがよいというガイドラインを示している。また、本論文中で新規開発した DAE の結果についても考察しており、似た枠組みである DCA よりもよい性能を示したものの最高性能ではなかった理由として、遺伝子数の少ないデータで学習が行われたためである可能性が高いことが述べられている。

第 4 章では、総合討論として、結果のまとめや今後の展望を述べている。

以上、本論文は、scRNA-seq カウントデータを視覚化するための前処理の性能を包括的に評価することにより、入力データの遺伝子数・細胞数・グループ数の違いに対応した実践的なガイドラインを提案したものである。その成果は、学術上応用上寄与するところが少なくない。よって、審査委員一同は、本論文が博士（農学）の学位論文として価値あるものと認めた。