論文の内容の要旨

応用生命工学専攻
平成 28 年度博士課程進学

氏名　　佐藤建太
指導教員名　清水謙多郎

論文題目　一細胞 RNA シーケンシングにおける大規模細胞検索と統計的推論

背景

DNA 配列シーケンシング技術の向上と実験技術の革新により、RNA-Seq で細胞のトランスクリプトームを細胞レベルで定量する技術、いわゆる一細胞 RNA-Seq は飛躍的な発展を遂げた。一細胞 RNA-Seq は、多数の細胞の平均的な RNA 量を定量する従来の RNA-Seq では困難な粒度で遺伝子発現の多様性解析を可能にし、今日の細胞生物学においては重要なツールとなった。さらに、一細胞 RNA-Seq の最初期には数十から数百個程度の細胞数に留まっていたスループットも大きく改善し、2010 年代後半には数千個を超える細胞での同時定量も行える手法が登場した。現在では、組織や臓器に留まらず、全身の細胞を一細胞 RNA-Seq で解析して、個体を構成するすべての細胞型を明らかにしようとする野心的な研究も登場している。

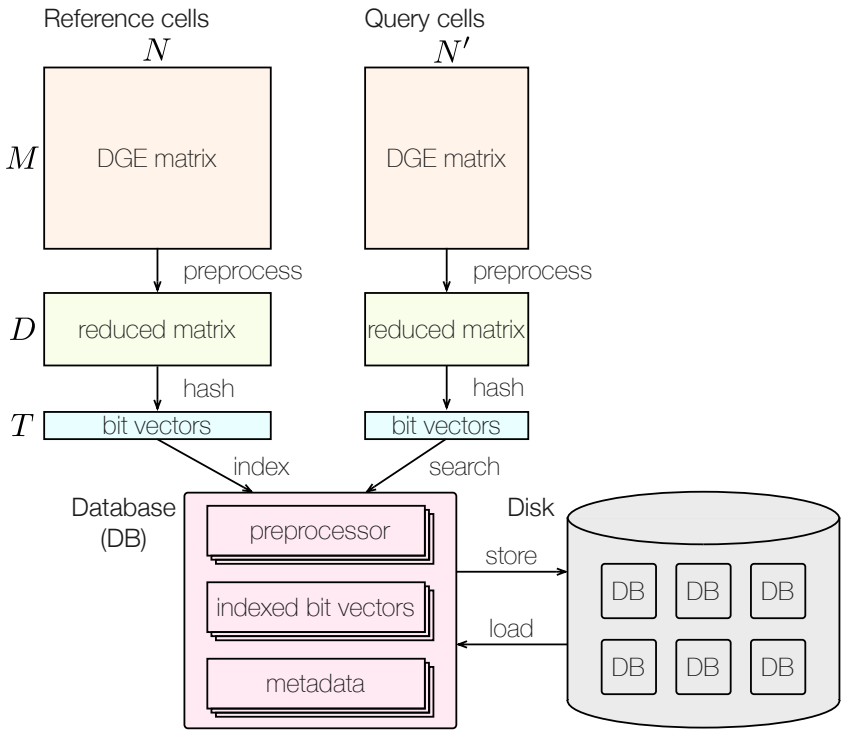
こうした一細胞 RNA-Seq の大規模化に伴い、2 つの課題を考える必要が生じた。1 つ目は、既存の一細胞 RNA-Seq から得られたデータを有効に活用する解析手法の開発である。例えば、一細胞 RNA-Seq の実験で新しく得られた細胞の発現プロファイルが既知の細胞のそれと似ているのか、もし似ているとしたら発現量の異なる遺伝子はあるか、などを比較検討するためのソフトウェアが必要になる。2 つ目は、細胞に極微量しか含まれない RNA 分子を一細胞 RNA-Seq で定量したとき、その計測につきまとう不確実性を統計的にどう扱うかという問題である。この問題は一細胞 RNA-Seq の大規模化に伴い細胞あたりにシーケンシングできるリード数が限られるようになると、より顕著になると考えられる。特に、類似した発現プロファイルを持つ細胞間の差異を評価するとき、計測に伴う不確実性の大きさは重要な問題となりうる。

大規模な細胞検索技術の開発

本研究では、一細胞 RNA-Seq で定量された細胞の発現プロファイルからデータベースを構築し、別に得られた発現プロファイルをクエリとして類似した発現プロファイルを持つ細胞を検索するソフトウェアを開発した。この技術は細胞の発現プロファイルを次元削減したのちビットベクトルに変換して省メモリ化を実現し、さらに検索用の索引をつけることで大規模なデータベースでも類似細胞の高速検索が可能になる。データベースは一度構築すればディスクに保存してすぐさま使用できるので、既存の研究で得られた細胞の発現プロファ

イルをデータベース化すれば、新しいデータセットを予め準備したデータベースに対して検索するという応用が可能である。開発したソフトウェアはプログラミング言語 Julia のパッケージとして、CellFishing.jl[1] として GitHub 上で公開されている。図 1 に CellFishing.jl の概要を示す。

図 1　CellFishing.jl の概念図



CellFishing.jl の性能は 5 つの実データセット（4 つのアノテーション付き中規模データセットと 1 つの大規模データセット）を用いて評価した。比較として、同じく類似細胞検索を行うソフトウェアである scmap[2] を用いた。最初に、検索して見つけた類似細胞がクエリの細胞と同じ細胞型になるかを確認した。この評価では、データセットを細胞が重複しないようクエリ用とリファレンス用に 1 対 4 の割合でランダムに分割し、リファレンス用の細胞を使ってデータベースを構築した。実験したほとんどのパラメータセットにおいて、CellFishing.jl は scmap より高い細胞型の一致率を示した。同時に、検索にかかった時間を計測したところ、CellFishing.jl は scmap より最大で 100 倍以上高速であった。また、データセットに含まれる細胞型の割合の偏りを考慮し、割合が 1% を下回る希少な細胞型についても一致率を詳細に確認したところ、CellFishing.jl は十分高い一致率であった。同様の実験を異なるバッチ間・生物種間・実験プロトコル間で試したところ、CellFishing.jl は scmap と同等以上の一致率を示した。

一方、検索により類似した細胞が見つかっても、類似度スコアのみではクエリと同じ細胞型の細胞と考えられるかを判断するのは難しい。そこで、クエリ細胞と検索で見つかった類似細胞の間で発現が異なる遺伝子を検出する機能を実装し、その評価を行った。CellFishing.jl はデータセットにある細胞の発現プロファイルを可逆圧縮して保存する機能を持っており、そのデータとクエリ細胞の遺伝子発現を比較することでクエリと類似細胞の間で発現が異なる遺伝子を検出できる。実験では、あるデータセットから immature B cell とラベル付けされた細胞をクエリ用として取り分け、データセットの残りの細胞に対して検索を行い、CellFishing.jl の発現変動遺伝子検出機能を使って比較したところ、生物学的に合理的で有意義と考えられる遺伝子セットが

検出された。

最後に、CellFishing.jl の大規模データセットへのスケーラビリティを確認した。100 万以上の細胞を含む大規模データセットから一部をクエリ用にランダムに選び、残りの細胞から $2^{13}, 2^{14}, \ldots, 2^{20}$ とサイズを変えながらリファレンス用の細胞をランダムに選んでデータベースを構築した。この実験では、索引を用いない線形探索と比較して大幅な高速化を実現し、100 万細胞を超える最大サイズのデータベースでも 1 秒あたりの検索細胞数が約 1,600 に達するなど、CellFishing.jl が今日得られる最大規模のデータセットに対しても十分適用可能であることを示した。さらに、データベースの保持に必要なメモリサイズは数百メガバイト程度で済み、ラップトップ級の計算機でも使用可能であることが分かった。

組成の Bayes 推定手法の開発

本研究では、細胞のトランスクリプトームを組成とスケールに分け、それぞれを Bayes 推定する方法を開発した。ここで、組成とはシンプレックス（合計が 1 となる d 個の正数の組からなる集合）の元を指す。ここに Aitchison 幾何学を導入すると、組成は内積空間の元として扱えるようになり、遺伝子のフィルタリング・集約・主成分分析などの重要な操作が線形変換として組成に対して定義できる。

Bayes 推定に用いたモデルの概要は次のとおりである。ある細胞について、トランスクリプトーム組成 x の遺伝子 i に対応する成分を x_i、スケールを κ とすると、一細胞 RNA-Seq における UMI カウント n_i のノイズは平均 κx_i の Poisson 分布

$$n_i \sim \mathrm{poisson}(\kappa x_i)$$

でモデル化できる。このモデルでは、事前分布を Dirichlet 分布に取ると一定の条件のもとで事後分布も Dirichlet 分布であり、Bayes 推定は解析的に容易になるが、Dirichlet 分布は柔軟性を欠くので事前知識を十分に反映できない。そこで、シンプレックス上に正規分布を導入し、それを事前分布とすることで、任意の平均と共分散を事前分布に反映できるようになる。なお、本研究を着想するきっかけとなった Sanity[3] という手法も Bayes 推定を行うが、Sanity はトランスクリプトームを組成として一貫性のある形で扱わず各遺伝子を独立に扱うなど、本研究の手法とは重要な違いがある。

この Bayes モデルの事後分布は単純ではないが、事後分布をシンプレックス上の正規分布で近似することで、解析的に扱いやすい表現が得られる。この近似分布は、事後分布との間の Kullback-Leibler ダイバージェンス最小化という最適化問題（変分推定）の解として得られる。本研究で開発した変分推定アルゴリズムは高速であり、大規模なデータセットに対しても適用可能である。

Bayes 推定で細胞レベルのトランスクリプトーム組成がランダム組成として扱えるようになると、組成から計算される細胞間の距離や類似度などもランダム変数として扱えるようになる。これは、距離や類似度といった量に一細胞 RNA-Seq の定量に伴う不確実性を反映できるということを意味する。

総括

類似細胞検索の研究では、一細胞 RNA-Seq で得られる発現プロファイルは数百ビット程度のビットベクトルに要約しても十分に類似性の推定が可能なことを示し、その検索が既存手法と比較して非常に高速であることを示した。さらに、細胞レベルの発現変動遺伝子解析が有用なツールとなりうることを示した。また、現在までに得られる最大規模のデータセットに対しても、開発した手法が適用可能であることを確認した。

トランスクリプトーム組成の Bayes 推定の研究では、既存の Bayes 推定手法を Aitchison らの組成の幾何

学を用いた視点から捉え直し、より柔軟な事前知識の利用と広範な解析を可能にすると期待される手法を開発した。現在までに得られた結果は限定的ではあるが、応用範囲の広い解析手法のフレームワークとして発展していくことを期待している。

参考文献

1. Sato, K., Tsuyuzaki, K., Shimizu, K. & Nikaido, I. CellFishing.jl: an ultrafast and scalable cell search method for single-cell RNA sequencing. *Genome Biol.* **20,** 31 (2019).
2. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nature methods* **15,** 359–362 (2018).
3. Breda, J., Zavolan, M. & van Nimwegen, E. Bayesian inference of gene expression states from single-cell RNA-seq data. *Nat. Biotechnol.* **39,** 1008–1016 (2021).