類似細胞検索のための Web アプリケーション「Cell similarity search」の開発

石橋　遼（筑波大学生物学類）　　指導教員：尾崎　遼（筑波大学医学医療系）

【背景・目的】

1 細胞 RNA シーケンシング（scRNA-seq）は細胞集団・生体組織内における個々の細胞の遺伝子発現を計測でき、それらの遺伝子発現状態は細胞レベルの機能・表現型を反映していると期待される [1] 。しかしながら、遺伝子発現状態から細胞レベルの機能・表現型を予測するのは一般に容易ではない [2] 。

この問題に対し、公共データベースに蓄積されている大量の scRNA-seq データを参照データとし、新たに実験で得られた scRNA-seq データに含まれる細胞をクエリとして遺伝子発現状態が類似した細胞を検索することで、細胞レベルの機能・表現型を予測できることが期待される。類似細胞検索ツールとしては、scmap[3]、CellFishing.jl[4]、Cell BLAST[5]等が知られている。

そこで本研究では、実験系研究者が遺伝子発現に基づく類似細胞検索を簡便に行える Web アプリケーションである「Cell similarity search」の開発を目指した。

【材料】

検索対象のデータセット

本研究では概念実証として、米国 NCBI Gene Expression Omnibus に収載されている、マウス大脳皮質由来の Chromium 技術（10x Genomics 社）を用いて取得された scRNA-seq データ 380 件分を対象とした。所属研究室の井尻遥士氏によって Seurat object 形式に変換された scRNA-seq データおよびサンプル情報の提供を受けた。学習の前準備として、提供された Seurat object を AnnData に変換した。また、サンプル情報を、AnnData に含まれる細胞メタデータ（個々の細胞ごとの情報）に付加した。

【方法】

Cell similarity search の全体像

Cell similarity search は、Web UI、細胞検索エンジン、細胞メタデータ DB の 3 要素から構成される（図 1）。Web UI は、遺伝子発現データのアップロードと細胞検索の実行、検索結果としてメタデータの表示を行う。細胞検索エンジンは遺伝子発現データをクエリとして受けとり、指定した参照データベースにおける類似細胞を探索する。細胞検索エンジンには、類似細胞検索ツールである Cell BLAST[5]の Python パッケージ（version 0.5.1）を用いた。細胞メタデータ DB は、細胞ごとのメタデータの保管・管理を行う。

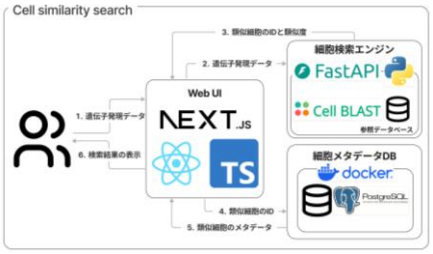

図 1. Cell similarity search の概要

Cell similarity search の性能検証

学習に使用した遺伝子発現量データを Cell similarity search の入力として類似細胞検索を行い、類似細胞検索の検索精度と、システムの応答性能を評価した。検索精度の評価には、参照データベースを検索先と指定し、対応する学習元 scRNA-seq データセットをクエリとして入力した際に、クエリに含まれる細胞と同一の細胞が類似細胞として返される割合（以下、自己検索成功率）を指標とした。また、システムの応答性能の評価には、参照データベース内で最も細胞数および遺伝子数が多い参照データベースに対して類似細胞検索を行った際の、ユーザが検索を実行してから結果が画面に表示されるまでの時間（以下、ターンアラウンドタイム）について、5 回の実行の平均を取り、指標とした。

【結果】

Cell similarity search の実装状況

ユーザは、Web UI から細胞×遺伝子の発現量行列を、CSV、TSV、AnnData といった複数のファイル形式でアップロードできる。続いて、検索対象となる参照データベースを選択し、検索・実行することで、細胞検索エンジンで類似細胞検索が行われ、類似細胞 ID と類似度のリストが返される。さらに、細胞メタデータ DB から類似細胞 ID によるメタデータの取得が行われ、Web UI に表示される。これにより、ユーザが Web UI を通じて scRNA-seq 由来の遺伝子発現データから簡便に類似細胞検索ができるシステムが実現された（図 2）。

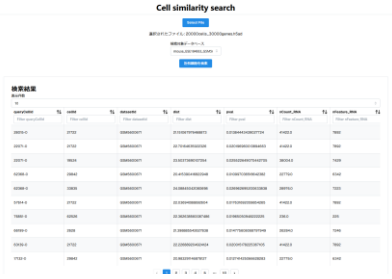

図 2. Cell similarity search の実行後画面

性能検証

全参照データベース（380 件、細胞数最大 79,485 個、最低 153 個、平均 9,350 個）において自己検索成功率は 100%となり、類似細胞を正確に検索できることが確認された。ターンアラウンドタイムの検証では、最大細胞数（79,485 細胞・32,245 遺伝子）および最大遺伝子数（60,310 細胞・55,367 遺伝子）のデータベースに対して、クエリとして 20,000 細胞・30,000 遺伝子のデータセットを入力したところ、それぞれ平均 648.06 秒、483.21 秒の応答時間を達成した。これを細胞あたりの処理時間に換算すると、約 32.403 ms/細胞および 24.161 ms/細胞となり、Cell BLAST[5]の報告値（10 ms/細胞）と比較し、ファイルのアップロードやメタデータ取得時間を含めた実運用条件下では同等の性能を実現した。

【展望】

現時点では、本システムはローカル環境でのみ使用可能であるが、今後は、一般公開に向けた調整を進める予定である。また、新しい参照データベースを容易に追加できるシステムを開発することで、継続的なデータの拡充とアプリケーションの機能向上を図れることが期待される。これにより、幅広い研究者が本アプリケーションを利用できる環境を構築することを目指す。

【参考文献】

[1] Jovic et al. (2022) Clin. Transl. Med., 12(3), e694.
[2] Zeng et al. (2022) Cell, 185(15), 2739–2755.
[3] Kiselev et al. (2018) Nat. Methods, 15, 359–362.
[4] Sato et al. (2019) Genome Biol., 20, 31.
[5] Cao et al. (2020) Nat. Commun., 11, 3458.