

機械学習を用いた

シングルセル RNA-seq データからの高精度タンパク質発現量予測

奥崎 大介（免疫学フロンティア研究センター 特任准教授（常勤）

瀬尾 茂人（大学院情報科学研究科 准教授（IDS 兼任））

石川 昌和（香川大学 バイオインフォマティクス解析センター 特命助教）

1 研究の背景

シングルセル RNA シーケンシング (scRNA-seq) は細胞ごとの転写物 (RNA) を網羅的に取得できる一方、捕捉できる RNA 分子数が限られるため発現していても検出されない遺伝子が多く、同一遺伝子から生じるタンパク質アイソフォームを配列情報だけで識別することも難しい。その結果、細胞機能や状態の差異を見落とす恐れがある。こうした弱点を補うために、RNA と細胞表面タンパク質を同時に定量する CITE-seq が開発された。タンパク質は翻訳後修飾やアイソフォームの違いを反映した最終的な機能分子であるため、CITE-seq を用いれば scRNA-seq では区別できなかった表面タンパク質アイソフォームを直接測定でき、細胞状態をより正確に捉えられる。しかし CITE-seq は測定できるタンパク質が抗体パネルに依存し、また試薬コストも高いため、Human Cell Atlas のような大規模プロジェクトでも依然として RNA 単独計測が主流である。

2 研究の目的

CITE-seq データは「RNA とタンパク質の発現量の対応関係」が対になった教師データとして利用できるという点で極めて価値が高い。このペア情報を用いて、RNA 発現プロファイルからタンパク質発現量を高精度に予測するモデルを構築できれば、CITE-seq を実施していない scRNA-seq データにも擬似的にタンパク質情報を付加できる。これにより、抗体パネルが異なる複数データセットの統合解析、タンパク質依存的な細胞型同定やシグナル経路推定、さらには治療標的候補のスクリーニングなど、幅広い応用が期待される。ゆえに、RNA だけからタンパク質量を推定する計算手法の開発は、実験コストを抑えつつ細胞機能を深く理解するための鍵となる。

本研究では、CITE-seq 由来のデータを訓練に用い、RNA 情報のみから細胞表面タンパク質量を高精度に推定する計算モデルを構築することを目的とした。具体的には、まず LASSO 回帰を用いてタンパク質ごとに情報量の高い転写物を選抜しつつ線形予測モデルを作成し、次いで複雑な非線形関係を捉える深層学習モデルを設計した。

3 研究の方法と結果

本研究では、公開されている様々な CITE-seq のデータを利用した。データベースからダウンロードしたデータについて、サンプル間・研究チーム間のバッチエフェクトを考慮し、クオリティコントロール・正規化を行い、細胞ごとに適切なタンパク質・RNA 発現量を算出した。RNA 発現量を説明変数、タンパク質発現量を目的変数として機械学習のモデルを訓練し、ク

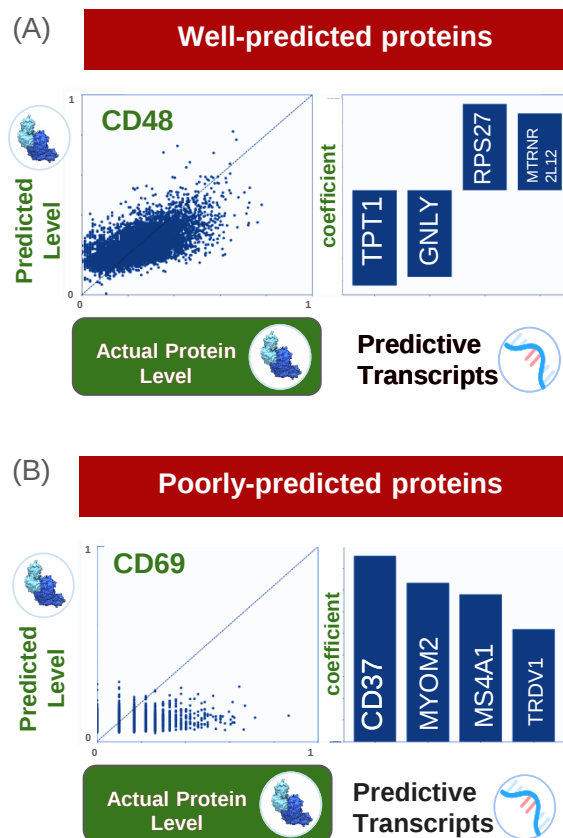


図 1: LASSO 回帰を用いた細胞表面タンパク質の発現量の予測と重要な転写産物の選択

ロスバリデーションによって評価を行う。本稿では、そのうちの COVID-19 患者由来の CITE-seq データセット（細胞数約 8 万、タンパク質 200 種）[a] についての結果を報告する。

まず、LASSO モデルでは、タンパク質の実測値と予測値の相関係数によって予測精度を評価し、その精度に基づき三つのカテゴリーに分類できた。高精度群ではわずか数十種類の転写物で実測値と高い相関を示し、なかには本来の発現遺伝子以外の転写物が主要な予測因子となる例も見られた（図 1A）。図の左の散布図はタンパク質の実測値と予測値の間に相関が見られることを示しており、図の右側は LASSO によって予測に重要と判断された特徴量（RNA 分子の種類）である。中精度群は部分的に再現性を示した一方、低精度群では RNA からの推定が困難であった（図 1B）。

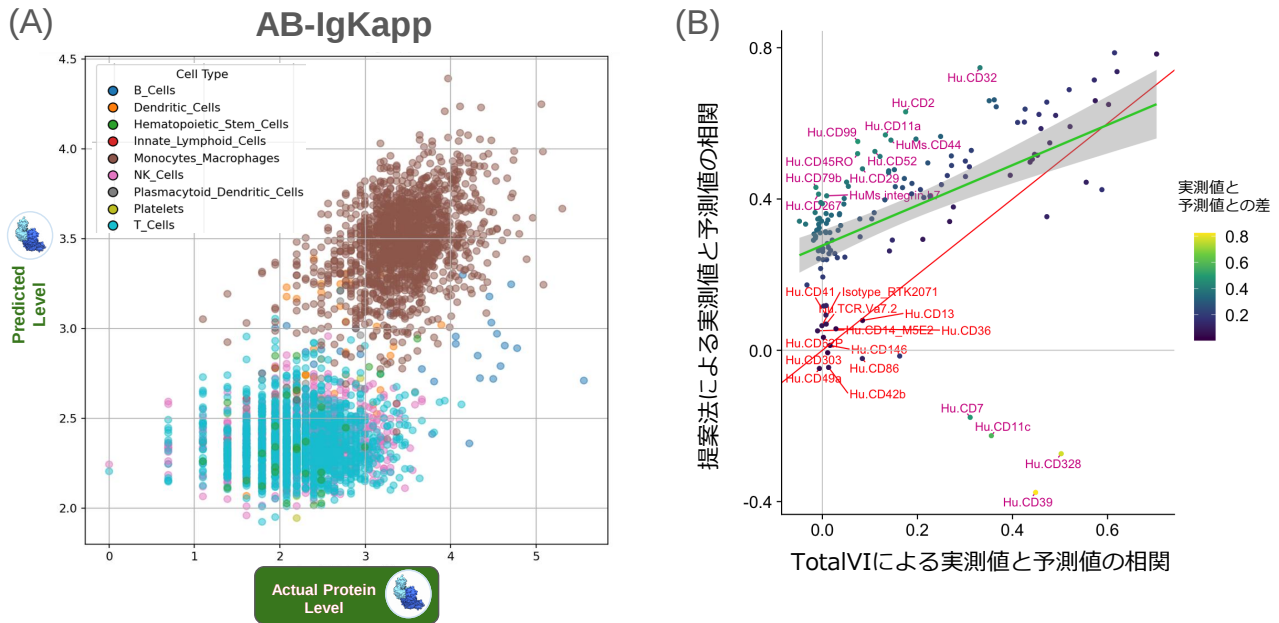


図 2: 深層学習を用いた細胞表面タンパク質の発現量予測

また、深層学習のモデルを用いることでタンパク質発現量の予測精度はさらに向上した。図 2A は、抗体パネル中の代表例として AB-IgKapp の実測タンパク質量（横軸）と本モデルによる推定値（縦軸）を単一細胞レベルで対比した散布図であり、提案モデルは細胞型に依存した発現量の違いを高精度で再現していることがわかる。図 2B は、実測値と予測値との相関、および平均二乗誤差を指標として既存手法 TotalVI[b] と比較した結果である。TotalVI では無相関（相関係数が 0 に近い）を示すようなタンパク質の種類について、提案手法では相関係数が上昇していることが確認できる。

4 今後の予定

今後は、他組織や他疾患由来データへの適用を通じて汎化性能を検証する必要がある。さらに、モデル内部で予測に寄与する遺伝子群を可視化し解釈性を高めるとともに、非表面型や分泌型タンパク質の発現推定へ対象を拡張したい。加えて、細胞の分化段階を表現学習に組み込み、発現ダイナミクスを定量的に取り込める構成を検討する。

謝辞

本プロジェクトの推進にあたり、免疫学フロンティア研究センターの Diez Diego 准教授、理学研究科の大学院生 Muhammad Neanaa さんにご助力を賜りました。

引用文献

- [a] Stephenson E., et al. "Single-cell multi-omics analysis of the immune response in COVID-19." *Nature medicine*, 27.5 (2021): 904-916.
- [b] Gayoso A., et al. "Joint probabilistic modeling of

single-cell multi-omic data with totalVI." *Nature methods* 18.3 (2021): 272-282.

発表論文等

〔雑誌論文〕

- [1] Kim SE, Noda R, Liu Y, Nakajima Y, Kameoka S, Motooka D, Mizuno S, Takahashi S, Takaya K, Murase T, Ikematsu K, Tratsiakova K, Motoyama T, Nakashima M, Kishi K, Martin P, Seno S, Okuzaki D, Mori R, "Novel integrated multiomics analysis reveals a key role for integrin beta-like 1 in wound scarring", *EMBO reports*, 26.1 (2025): 122-152.

〔外部資金〕

- [1] 2022-2024 年度, 科学研究費助成金 基盤研究 C, "時空間情報を統合した 1 細胞発現解析手法の開発", (代表) 瀬尾茂人
- [2] 2023-2025 年度, 科学研究費助成金 若手研究, "敗血症患者全白血球のシングルセルシーケンスによる重症化バイオマーカーの探索", (代表) 石川 昌和
- [3] 2023-2025 年度, 科学研究費助成金 基盤研究 B, "正常組織に影響を与えない癌幹細胞標的分子の探索", (代表) 奥崎 大介