

scMM によるシングルセルマルチオミクスデータ解析



シングルセルマルチオミクスデータから 有用な知見を抽出する人工知能技術を開発！ ～疾患の一細胞レベルでの理解を加速～

国立大学法人東海国立大学機構 名古屋大学大学院医学系研究科（研究科長・門松健治）・システム生物学分野の島村徹平（しまむらてっぺい）教授、箕浦広大（みのうらこうだい）医学部学生、分子細胞免疫学分野の西川博嘉（にしかわひろよし）教授（国立がん研究センター 研究所 腫瘍免疫研究分野、先端医療開発センター 免疫 TR 分野 分野長兼任）らの研究グループは、ディープラーニングの一種である深層生成モデルを応用し、シングルセルマルチオミクスデータから有用な知見を抽出する人工知能技術を開発することに成功しました。近年発展が著しいシングルセル解析技術の進展により、トランスクリプトーム、エピゲノム、細胞表面マーカーといったモダリティ情報を一細胞レベルで網羅的に計測することが可能となり、疾患の原因となりうる細胞集団の同定や異常細胞の機能解析が盛んに行われています。中でも、同じ細胞から複数のモダリティを同時計測することのできるシングルセルマルチオミクス解析が近年注目を集めています。シングルセルマルチオミクス解析により、単一のモダリティでは捉えきれない細胞集団の多様性や機能が明らかになることが期待されていますが、こうした複数モダリティの複雑な情報を含んだビッグデータから医学生物学的に有用な知見を発見するための手法は限られていました。

そこで、同研究グループはシングルセルマルチオミクスデータの解析に特化した人工知能技術 scMM（A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data）を開発しました。scMM は大規模データから個々のデータの潜在的な状態を推論することが可能な深層生成モデルを基盤として開発されており、複数モダリティの統合や圧縮、モダリティ間の関係性の発見などを完全に自動で行うことが可能です。

本研究は、シングルセルマルチオミクス解析を用いた研究を加速し、感染症、がん、精神疾患などの様々な疾患の一細胞レベルでの理解や新規治療法の確立に寄与するものと期待されています。この研究成果は「Cell Reports Methods」のオンライン版に掲載されました（現地時間 2021 年 9 月 15 日）。

ポイント

- ディープラーニング（※1）の一種である深層生成モデル（※2）を用いて、シングルセルマルチオミクスデータにおける複数モダリティの情報を圧縮・統合する人工知能技術 scMM を開発した。
- シングルセルマルチオミクスデータのモダリティ間の関係性を自動で学習し、モダリティをまたいだ情報の変換による欠損モダリティの補完に成功した。
- scMM を用いることでシングルセルマルチオミクスデータから有用な知見を効果的に抽出することが可能であり、感染症、がん、精神疾患などの様々な疾患の一細胞レベルでの病態解明を加速することが期待される。

1. 背景

近年、シングルセル解析と呼ばれる単一細胞レベルでトランスクリプトーム（single-cell RNA-seq; scRNA-seq）（※3）や細胞表面マーカー（※4）、クロマチンアクセシビリティ（single-cell Assay for Transposase Accessible Chromatin; scATAC-seq）（※5）などの情報を計測する一連の技術が発展しており、正常な組織内における細胞の多様性やがんの不均一性、細胞の運命追跡など様々な分野において重要な発見に寄与しています。さらに、これまでのシングルセル解析技術が単一のモダリティを計測しているのに対し、技術革新が進んだ現在では一細胞ごとにトランスクリプトームとエピゲノム、トランスクリプトームと細胞表面マーカーのような複数モダリティの同時計測が可能となっています。このようなシングルセルマルチオミクス解析によって得られるデータから、従来型の単一モダリティ解析では捉えられなかった細胞の多様性や、モダリティ間での関係性が明らかになることが期待されています。一方、シングルセルマルチオミクスデータには一細胞ごとに非常に高次元のデータが含まれ、かつ複数モダリティに渡ってデータが取得されているため複雑な構造となっています。そのため、現在シングルセルマルチオミクスデータから有用な知見を計算科学的に抽出するバイオインフォマティクス解析手法は非常に限られています。本研究では複数モダリティの高精度な圧縮・統合および、モダリティ間での関係性を学習することによる欠損モダリティの補完を主な目的としてモデルの開発を行いました。

2. 研究成果

本研究で開発した scMM は深層生成モデルを基盤技術とし、シングルセルマルチオミクスデータの解析に特化したモデルとなっています（図 1）。深層生成モデルは大量のデータから、その背後に存在する確率分布のパラメータを自動的に推論する非常に強力なモデルとして広く応用されています。scMM では各モダリティのデータ特性を反映した確率分布を選択してモデルを構築し、混合エキスパート（mixture-of-experts）（※6）と呼ばれる手法によってシングルセルマルチオミクスデータにおける複数モダリティの統合を行うことに成功しました。公開されているヒト末梢血単核細胞の scRNA-seq および細胞表面マーカーのシングルセルマルチオミクスデータを用いて scMM の性能を検証した結果、scMM により複数モダリティの情報を圧縮することで、従来手法より高い粒度で細胞集団の分離が行えることが明らかになりました（図 2 上段）。また、データを自在に生成できるという深層生成モデルの特性を活かした「擬似細胞（pseudocell）」の生成により、圧縮された情報の各次元に、どのような情報が含まれているかを可視化することが可能になってい

ます (図 2 左下)。また、モダリティ間をまたいでデータを生成することで、scRNA-seq のみのデータから細胞表面マーカーを高精度に予測することが可能であることが示されました (図 2 右下)。公開されている scRNA-seq および scATAC-seq のシングルセルマルチオミクスデータを用いた検証でも同様の結果が得られました。

scMM によるシングルセルマルチオミクスデータ解析

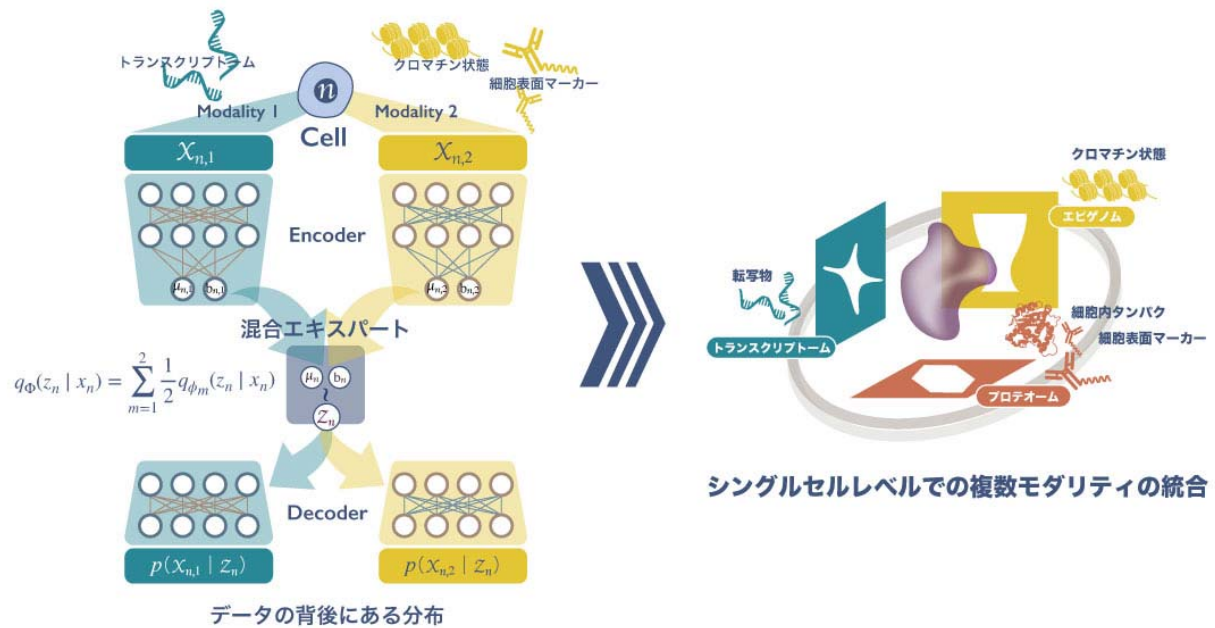
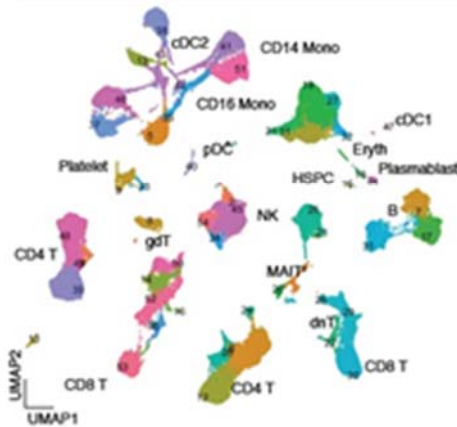
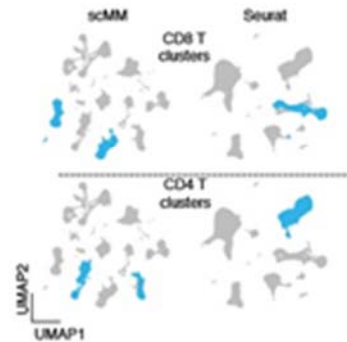


図 1 シングルセルマルチオミクスデータに特化した深層生成モデル (scMM) により、シングルセルレベルでの複数モダリティの統合が可能。

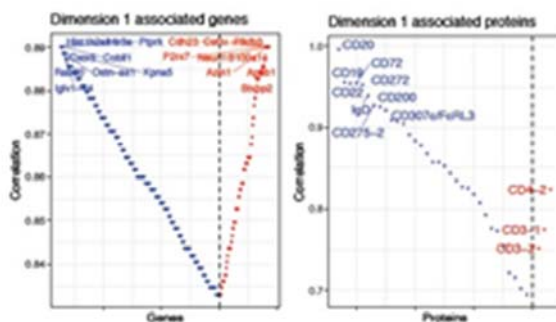
複数モダリティ情報の圧縮表現



既存手法と比較し高い分解能



圧縮表現に含まれる情報の可視化



欠損モダリティの高精度な予測

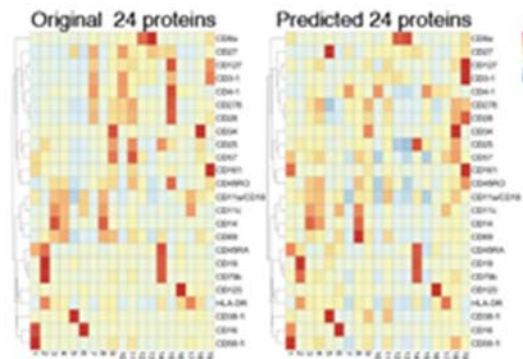


図 2（左上）トランスクリプトームおよび細胞表面マーカー情報の圧縮表現から、細胞集団を分離することが可能。（右上）既存手法と比較し、高い粒度での細胞集団解析を実現。（左下）圧縮表現に含まれる、トランスクリプトームおよび細胞表面マーカーの可視化。（右下）学習済みモデルを用いることで、トランスクリプトーム情報のみから、細胞表面マーカーの高精度予測が可能。

3. 今後の展開

scMM は各モダリティのデータ分布に応じた確率分布を用いることで、今回検証したモダリティ以外にも DNA メチル化、ヒストン修飾、空間トランスクリプトームなど様々なモダリティに対応することが可能な柔軟なモデルとなっています。これからのシングルセルマルチオミクス解析の基盤として、scMM のさらなる拡張・発展が期待されています。

4. 用語説明

※1. ディープラーニング

近年の人工知能分野の急速な発展に大きく寄与した、機械学習手法の一つ。大規模データから効率的に学習を行い、ニューラルネットワークによって非常に複雑なデータ構造を捉えることができる。

※2. 深層生成モデル (deep generative model)

統計モデルの一種である生成モデルの学習にディープラーニングを用いた一連の手法の総称。様々

な確立分布を組み合わせることで、データや解析目的に応じた柔軟なモデルを設計することが可能。強固な理論的裏付けがあることやその柔軟性から、非常に強力な手法として近年広く用いられている。

※3. トランスクリプトーム (single-cell RNA-seq; scRNA-seq)

遺伝子の発現量に関する網羅的な解析。

※4. 細胞表面マーカー

細胞の表現型と深く関わるのがわかっており標識として用いられる、細胞表面に発現するタンパク質。

※5. クロマチンアクセシビリティ (single-cell Assay for Transposase Accessible Chromatin; scATAC-seq)

DNA 配列への RNA ポリメラーゼや転写因子のアクセス性。

※6. 混合エキスパート (mixture-of-experts)

複雑な問題を分解し、個々の単純な問題を解くモデル(エキスパート)の学習に置き換える手法。

5. 研究開発プロジェクトについて

本研究開発プロジェクトは、未知のウイルスに対しても有効な検査・治療法を先制的に準備することを目標とします。具体的には、ウイルスとウイルスに対する生体応答ネットワークの中核である免疫系を対象として、包括的・網羅的かつ経時的に計測する次世代計測技術を開発し、得られたデータを用いて数理モデルを構築します。従来のウイルス分類による対処法開発ではなく、生体応答を対象にした予防・治療法の開発により、未知のウイルスに対しても有効な創薬基盤・検査手法の先制的な開発体制を整備します。

本研究成果に対する PM 松浦善治のコメント

本研究開発プロジェクトでは、ウイルスに対する宿主免疫応答について、マルチオミクス解析によって得られるビッグデータの解析とそのモデル化が重要なキーとなります。今回開発した技術により、プロジェクトにおけるビッグデータの解析が飛躍的に進み、研究のさらなる発展が期待されます。

6. 本研究への支援

本成果は、以下の事業・プログラム・プロジェクト・研究開発課題によって得られました。

JST ムーンショット型研究開発事業 (MS)

研究開発プログラム：「2050年までに、超早期に疾患の予測・予防をすることができる社会を実現」

(プログラムディレクター：祖父江 元 愛知医科大学 理事長・学長)

研究開発プロジェクト：「ウイルス-人体相互作用ネットワークの理解と制御」

(プロジェクトマネージャー (PM) : 松浦 善治 大阪大学 感染症総合教育研究拠点 拠点長／
微生物病研究所 特任教授)
研究開発課題名 : 「ネットワーク解析と数理モデルに基づく免疫応答ダイナミクスの層別化」
課題推進者 : (島村 徹平 名古屋大学 大学院医学系研究科 教授)

7. 発表雑誌

掲雑誌名 : Cell Reports Methods

論文タイトル : A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data

著者 : Kodai Minoura^{1,2}, Ko Abe³, Nam Hyunha¹, Hiroyoshi Nishikawa^{2,4}, and Teppei Shimamura^{1,*}

所属 : ¹Division of Systems Biology, Nagoya University Graduate School of Medicine, Nagoya, Japan.

²Department of Immunology, Nagoya University Graduate School of Medicine, Nagoya, Japan.

³Laboratory of Medical Statistics, Kobe Pharmaceutical University

⁴Division of Cancer Immunology, Research Institute/EPOC, National Cancer Center, Tokyo/Chiba, Japan.

DOI : 10.1016/j.crmeth.2021.100071

English ver.

https://www.med.nagoya-u.ac.jp/medical_E/research/pdf/Cell_Rep_Met_210916en.pdf