

実験の「ばらつき」だけを取り除き、病気のサインを残す ～実験ノイズと生物学的シグナルを分離する AI「Kanade」～

【本研究のポイント】

- ・ 異なる実験で得られたデータを揃える際、本来知りたい疾患や治療による違いまで消してしまうことが、細胞の解析における課題となっていた。
- ・ 「実験によるばらつき」「疾患や処置による違い」「細胞型による違い」を分けて学習し、実験由来の情報だけを選択的に取り除く機械学習法「Kanade」を開発した。
- ・ シミュレーションとマウス臓器の実データで性能を検証し、実験ごとのばらつきを補正しながら、疾患に関連する細胞状態の違いを保持できることを確認した。
- ・ 補正後の遺伝子発現量を再構成できるため、細胞の分類や可視化に加え、疾患に関連する遺伝子の探索など、その後の解析にも利用できる。

【研究概要】

名古屋大学大学院医学系研究科の坂口峻太特任助教、堤真人特任助教(現・筑波大学生存ダイナミクス研究センター助教)、名古屋大学大学院医学系研究科の本田直樹教授(兼任:広島大学大学院統合生命科学研究科特任教授)らの研究グループは、1 つ 1 つの細胞の遺伝子の働きを調べたデータから、「実験ごとのばらつき」と「比較したい生物学的な違い」を分離する、新しい機械学習法「Kanade」を開発しました。

1 つ 1 つの細胞で、どの遺伝子がどの程度働いているかを網羅的に測定する「1 細胞 RNA シーケンシング(scRNA-seq)」は、疾患や治療による細胞の変化を調べるための重要な技術です。しかし、異なる実験で取得したデータには、測定装置、試薬、実験者などの違いによる技術的なばらつきが含まれます。このような「バッチ効果」を取り除こうとすると、本来知りたかった疾患や治療による違いまで、一緒に消してしまうことがあります。特に、健常群と疾患群などが別々の実験で測定されている場合、実験ごとのばらつきと条件間の生物学的な違いを区別することが難しくなります。条件間の比較は科学の根幹であるため、両者を分離し、バッチ効果のみを補正する手法が必要でした。

Kanade は、深層生成モデルの一種である変分オートエンコーダーを基盤として、遺伝子発現データに含まれる情報を、実験上のばらつき、疾患や処置などの条件による違い、細胞型の違いに分けて学習します。そして、実験由来の情報だけを取り除き、生物学的な特徴を保った遺伝子発現データを再構成します。データを単に似た形にそろえるのではなく、「消したい違い」と「残して比較したい違い」を区別して扱うことが、この手法の特徴です。

Kanade をシミュレーションデータに適用した結果、実験ごとの技術的な違いを取り除きながら、疾患や処置に関連する細胞の特徴を保持できることが示されました。さらに、実際の糖尿病モデルマウスの膵臓データを用いた解析でも、バッチ効果を補正した後も、糖尿病に関連する細胞状態の変化を捉えることができました。

また、Kanade は補正後の遺伝子発現量そのものを出力できるため、細胞の分類や可視化だけでなく、疾患に関連する遺伝子の探索などにも利用できます。今後、異なる実験や研究施設で取得されたデータを、生物学的に重要な違いを保ちながら比較・統合するための解析基盤へと発展することが期待されます。

本研究成果は、2026 年 10 月10日国際学術雑誌『Patterns』に掲載されました。

1. 背景

近年、1つ1つの細胞における遺伝子発現を測定する「シングルセル RNA シーケンシング (scRNA-seq)*¹」技術が発展し、従来は細胞集団の平均としてしか捉えられなかった違いを、個々の細胞ごとに調べられるようになりました。発生、疾患、薬剤応答などに伴って、どのような細胞が、どのように変化するかを理解するために、重要な役割を担っています。

一方で、複数の実験データを比較・統合する際には、実験日、測定装置、試薬、実験者、研究施設などの違いによる技術的なばらつきが問題になります。この「バッチ効果」*²が残っていると、実験環境によって生じたデータのずれを、細胞の性質や疾患による変化だと誤って解釈してしまう可能性があります。その影響を取り除く処理が「バッチ補正」です。

しかし、補正によって実験間の違いを小さくすればよいわけではありません。例えば、健常群と疾患群を別々の実験で測定した場合、データに現れる違いには、疾患による変化と実験環境の影響が混在します。両者を区別せずに実験間の違いを取り除こうとすると、疾患に関連する重要な情報まで失われるおそれがあります。疾患や治療の研究で重要なのは、データを見かけ上そろえることではなく、何が変化したのかを正確に比較することです。そのため、「取り除きたい実験上のばらつき」と「残して比較したい生物学的な違い」を分離し、必要な情報を保持できる補正手法が求められていました。

2. 研究成果

本研究では、scRNA-seq データに含まれる異なる要因の情報を分離して学習し、バッチ効果だけを選択的に補正する機械学習法「Kanade」を開発しました。Kanade は、深層生成モデルの一種である変分オートエンコーダー (VAE) *³ を基盤としています。VAE は、多数の遺伝子の発現量など、複雑で高次元のデータを少数の変数に圧縮し、そこから元のデータを再構成するモデルです。通常の VAE は圧縮の際に実験上のばらつきを含めた多様な情報を「ごちゃ混ぜ」にするのに対して、Kanade では、この圧縮された情報の中で、バッチ効果、疾患や処置などの条件に関する情報、細胞型に関する情報を分離して、それぞれ別々に保持させるように設計しました。これにより、Kanade はデータ中の情報を「除去すべきバッチ効果」と「本来残すべき条件間の生物学的な違い」に分けて扱うことができます。補正の際には、バッチ効果に対応する情報だけを取り除き、条件差や細胞型の特徴を残した遺伝子発現データを再構成します。

まず、人工的に作成したシミュレーションデータを用いて Kanade の性能を検証しました。その結果、Kanade はバッチ効果を取り除きながら、健康状態と疾患状態の違いや、細胞型ごとの特徴を保持できることが確認されました。従来法では、バッチ効果の補正と同時に条件間の違いが混ざってしまう場合がありましたが、Kanade では条件に由来する情報をより明確に保持できました。

さらに、正常なマウスや糖尿病モデルマウスなど、複数の条件を含む膵臓の scRNA-seq データに適用しました。解析の結果、Kanade はデータセット間に生じる技術的な違いを補正しながら、糖尿病に関連する細胞の変化を保持できることが確認されました。これは、Kanade が単にデータを均一化するのではなく、実験上のノイズと疾患におい

て重要な違いを区別して補正できることを示しています。

Kanade ソースコードは GitHub にて公開しており、計算機環境があれば利用可能です (<https://github.com/ShuntaSakaguchi/kanade>)。

3. 今後の展開

Kanade は、疾患、薬剤応答、発生過程など、複数の条件を比較するシングルセルオミクス^{*4} 解析、特に大規模なデータを用いた解析に広く応用できると期待されます。将来的には、研究施設や設備を超え、世界各地で蓄積されるシングルセルデータをつなぐ、世界規模のデータ統合を支える解析基盤への発展を目指します。さらに、Kanade は比較的シンプルな VAE 構造をもとにしており、拡張性が高いことも特徴です。今回検証した scRNA-seq データにとどまらず、脳画像など、他の医学・生命科学データへの拡張も考えられます。「取り除きたい測定由来の違い」と「残して比較したい生物学的な違い」を分離するという考え方を、多様なデータの統合解析に展開していくことが期待されます。

また、ムーンショット目標 2「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」では、疾患の発症や進行を早期に捉え、予測・予防につなげるための研究開発が進められています。本研究は健康状態・疾患状態を跨いだ解析や疾患早期と後期のデータ統合などに有用であり、疾患の早期シグナルを捉える基盤としてムーンショット目標 2 の達成に寄与することが期待されます。

4. 支援・謝辞

本研究は、以下の研究プロジェクトの支援のもとで行われたものです。

- ・JST ムーンショット型研究開発事業目標 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 JPMJMS2024-9
- ・JST CREST JPMJCR25Q2
- ・AMED 脳神経科学統合プログラム(個別重点研究課題)JP25wm0625322, JP25wm0625210
- ・JSPS 科研費 26K21377

【用語説明】

*1 シングルセル RNA シーケンシング(scRNA-seq)

1つ1つの細胞について、どの遺伝子がどの程度発現しているかを測定する技術です。多数の細胞をまとめて測定する従来の方法とは異なり、細胞ごとの違いや、まれな細胞集団を調べることができます。

*2 バッチ効果

実験日、測定装置、試薬、実験者、研究施設などの違いによって生じる、技術的なばらつきのことです。生物学的な違いではないにもかかわらず、データ上では大きな差として

現れることがあります。この影響を取り除き、異なる実験のデータを比較・統合しやすくする処理が「バッチ補正」です。

*3 変分オートエンコーダー(VAE)

深層学習モデルの一種で、高次元のデータを少数の変数で表現し、そこから元のデータを再構成する手法です。この圧縮された表現を「潜在表現」と呼び、データの背後にある構造を学習するために用います。

*4 シングルセルオミクス

細胞 1 個ごとに、遺伝子発現、タンパク質、エピゲノムなどの分子情報を網羅的に測定・解析する技術の総称です。本研究では、主に 1 細胞ごとの遺伝子発現を扱う scRNA-seq を対象としていますが、他の手法へ拡張可能な枠組みになっています。

【論文情報】

雑誌名:Patterns

論文タイトル:Disentangling batch effects and biological signals across conditions in single-cell transcriptome

著者:坂口峻太(名古屋大学大学院 医学系研究科)、堤真人(筑波大学 生存ダイナミクス研究センター)、西健太郎(広島大学大学院 統合生命科学研究科)、本田直樹(名古屋大学大学院 医学系研究科、広島大学大学院 統合生命科学研究科、名古屋大学 One Medicine 生命-創薬共創プラットフォーム)

DOI: [10.1016/j.patter.2026.101638](https://doi.org/10.1016/j.patter.2026.101638)

English ver.

https://www.med.nagoya-u.ac.jp/medical_E/research/pdf/Pat_261013en.pdf