複数生物の遺伝子解析を行う Mix-Geneformer に関する研究

TR24010 西尾優希　　指導教授：藤吉 弘亘

1. はじめに

次世代シーケンサにより単一細胞の遺伝子発現量を計測できるようになり，細胞ごとの特性解析が可能となっている．解析の属人化を避け，効率化するために，深層学習を用いた single-cell RNA sequencing (scRNA-seq) 解析手法として，Geneformer[1] や Mouse-Geneformer[2] が提案されている．これらは，それぞれヒトとマウスの遺伝子発現情報を文章として扱い，Transformer により学習することで，汎用的な細胞の特徴表現を獲得している．この学習により獲得した特徴表現を用いることで，細胞型分類や in silico 摂動などの下流タスクで高い性能を示している．Geneformer および Mouse-Geneformer は，いずれも単一生物種を対象とした事前学習モデルであり，生物種を横断した解析は困難である．一方，生物種を横断した解析が可能となれば，マウスで得られた解析結果をヒトの解析に適用でき，創薬プロセスの短縮や研究効率の向上が期待できる．そこで本研究では，ヒトおよびマウスの scRNA-seq データを統合的に学習する Mix-Geneformer を提案する．これにより，生物種を横断した解析が可能なモデルの構築を目指す．

2. 深層学習を用いた scRNA-seq 解析

scRNA-seq 解析は，次世代シーケンサにより細胞を単一細胞レベルに分離して計測した遺伝子発現量をもとに，細胞間の多様性や状態変化を解析する手法である．scRNA-seq 解析における課題として，前処理や特徴量設計，細胞型同定などの解析工程において解析者の判断が介在する場面が多く，属人的なバイアスが生じやすい．この課題に対し，解析の自動化と汎用的な特徴表現の獲得を目的とした手法が提案されている．

深層学習を用いて細胞の特徴表現を学習する手法として，Geneformer[1] および Mouse-Geneformer[2] が提案されている．これらは Transformer を用いた scRNA-seq 解析手法であり，細胞を文章，遺伝子をトークンとして扱う点に特徴がある．具体的には，各細胞において遺伝子発現量上位 2,048 個の遺伝子を抽出し，発現量順に整列したトークン列として細胞文を構成する．両モデルはいずれも Masked Language Modeling（MLM）による事前学習を通じて，細胞型分類や in silico 摂動などの下流タスクに応用可能な特徴表現を獲得している．一方で，これらのモデルは単一生物種のデータを用いて事前学習しているため，生物種を横断した統合解析には至っていない．

3. 提案手法：Mix-Geneformer

本研究では，生物種を横断した解析が可能なモデルの実現を目的として，ヒトおよびマウスの scRNA-seq データを同一の Transformer で学習する scRNA-seq 解析モデル Mix-Geneformer を提案する．ヒトおよびマウスの scRNA-seq データを統合して学習することで，生物種に依存しない細胞表現の獲得を目指す．

3.1 Mix-Geneformer における事前学習

Mix-Geneformer の事前学習では，Masked Language Modeling（MLM）と SimCSE を組み合わせて用いる．MLM は，各細胞文内における遺伝子の関係を学習するための自己教師あり学習である．SimCSE は，ミニバッチ内の細胞文同士の関係性を捉えるための対照学習である．Mix-Geneformer の学習方法の概要を図 1 に，損失関数を式 (1)〜(3) に示す．

$$L_{\text{total}} = L_{\text{MLM}} + L_{\text{SimCSE}} \tag{1}$$

$$L_{\text{MLM}} = -\sum_{i \in M} \log P(x_i \mid x_{\setminus M}; \theta) \tag{2}$$

$$L_{\text{SimCSE}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\text{sim}(h_i^{(1)}, h_i^{(2)})/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}(h_i^{(1)}, h_j^{(2)})/\tau\right)} \tag{3}$$

式 (2)，(3) において，M は一部をマスクしたトークンの集合，$\text{sim}(\cdot,\cdot)$ はコサイン類似度，τ は温度パラメータである．式 (2) の MLM 損失 (L_{MLM}) は，マスクしたトークンを予測する過程を通じて，細胞文内における遺伝子発現順位の関係性を学習する．式 (3) の SimCSE 損失 (L_{SimCSE}) は，同一の細胞文に対してエンコーダ内の確率的な dropout により得られる 2 つの特徴表現を生成し，それらを正例対として扱う．各細胞 i に対して，同一エンコーダを 2 回通すことで得られる表現 $h_i^{(1)}$ および $h_i^{(2)}$ を正例とし，同一ミニバッチ内の他の細胞に由来する表現 $\{h_j^{(2)}\}_{j \neq i}$ を負例として対照学習を行うことで，細胞間の特徴表現の類似度を学習する．

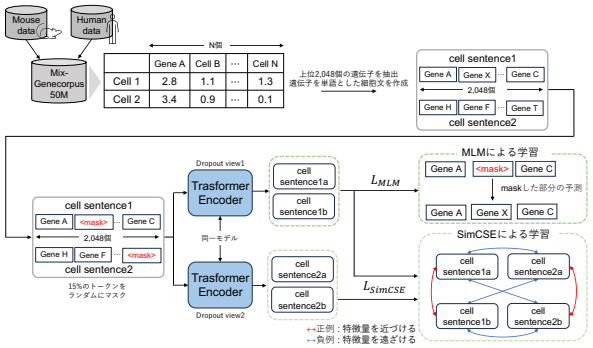


図 1: Mix-Geneformer の学習方法

3.2 学習データセット:Mix-Genecorpus-50M

Mix-Geneformer の学習データセットとして，ヒトの scRNA-seq データセットである Genecorpus-30M とマウスの scRNA-seq データセットである Mouse-Genecorpus-20M を統合し，約 5,000 万細胞から構成する Mix-Genecorpus-50M を作成した．Genecorpus-30M および Mouse-Genecorpus-20M は，複数の公開データセットを統合しており，多様な臓器や細胞型を含んでいる．

3.3 事前学習

本研究では，ヒトおよびマウスの scRNA-seq データを同一の事前学習データとして扱うことで，生物種を横断した解析が可能なモデルの作成を目的とする．事前学習では，埋め込み特徴を 256 次元とし，6 層の Transformer エンコーダを用い，MLM および SimCSE に基づく対照学習を組み合わせて学習を行った．MLM では，細胞文中の一部のトークンをランダムにマスクし，周辺のトークンから元のトークンを予測することで，細胞文内における遺伝子間の関係性の学習を促した．SimCSE による対照学習では，特徴表現間の類似度に基づく損失を導入することで，各細胞文同士の類似性を学習させた．事前学習は，バッチサイズ 8，warmup10,000 ステップを含む 10 エポックで行った．

4. 評価実験

本研究では，Mix-Geneformer の評価として，細胞型分類と in silico 摂動実験の 2 種類の下流タスクを行う．各実験において，単一生物種内の評価と，生物種の横断性の評価を行う．いずれの下流タスクにおいても，事前学習済みの Transformer に対して 10 エポックのファインチューニングを行う．

4.1 細胞型分類による評価

細胞型分類タスクでは，単一生物種内の評価として，事前学習済みモデルに対してマウスおよびヒトのデータでファインチューニングし，Geneformer および Mouse-Geneformer と細胞型分類精度を比較する．さらに，生物種の横断性を評価するため，マウスの脾臓データでファインチューニングしたモデルをヒトの脾臓データに，ヒトの脾臓データでファインチューニングしたモデルをマウスの脾臓データに適用し，細胞型ごとのモデルの特徴表現を UMAP により

可視化する．表 1 および表 2 に，マウスおよびヒトデータにおける分類精度を示す．

表 1: マウスの細胞型分類精度

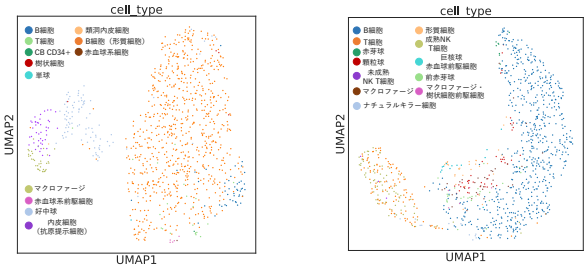
Organ	Types	Mouse-GF	Mix-GF
Brain	15	96.9	**97.6**
Heart	11	**97.8**	97.7
Kidney	18	94.9	**95.4**
Large_intestine	7	93.1	**94.6**
Limb muscle	9	99.5	**99.7**
Mammary gland	7	99.0	**99.1**
Spleen	10	**98.7**	98.6
Thymus	6	97.0	**97.6**
Tongue	3	94.9	**95.3**

表 2: ヒトの細胞型分類精度

Organ	Types	Human-GF	Mix-GF
Spleen	6	98.9	**99.0**
Brain	6	96.8	**97.7**
Immune	10	94.4	**95.1**
Kidney	15	92.8	**93.3**
Large_intestine	16	92.7	**93.4**
Liver	12	91.1	**91.2**
Lung	16	93.4	**94.3**
Pancreas	15	93.0	**93.5**
Placenta	3	97.9	**98.2**

表 1 および表 2 より，マウスおよびヒトのいずれのデータにおいても，Mix-Geneformer は従来モデルと同等以上の分類精度を示した．この結果は，複数の生物種のデータを同時に学習に用いることで，細胞型分類に寄与する特徴を獲得した可能性を示唆している．

また，生物種の横断性を評価した UMAP 可視化結果を図 2 に示す．図 2 より，可視化対象と異なるデータでファインチューニングしたモデルであっても，UMAP 上で細胞型ごとに一定程度クラスタが分離していることを確認した．このことから，Mix-Geneformer はマウスとヒトのデータを同時に学習することで，生物種を横断した解析が可能であると示唆される．



(a) マウスで FT →ヒトの脾臓データ可視化　(b) ヒトで FT →マウスの脾臓データ可視化

図 2: 生物種の横断性に関する UMAP 可視化結果

4.2 in silico 摂動実験による評価

in silico 摂動実験とは，コンピュータ上で遺伝子の過剰発現や遺伝子削除を模擬し，細胞状態を目標状態へ近づける上で重要な遺伝子を同定する手法である．in silico 摂動実験の概要を図 3 に示す．

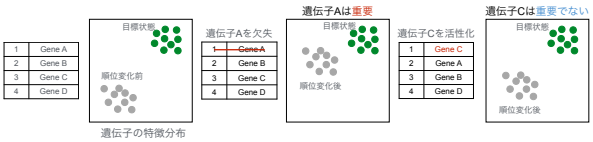


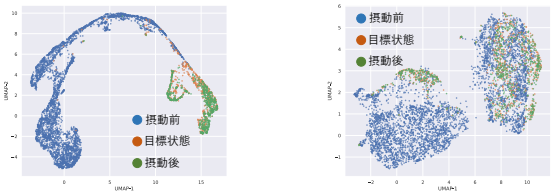
図 3: in silico 摂動実験の概要

本実験では，対象遺伝子の順位を変化させることで細胞に仮想的な摂動を与え，摂動後の特徴表現と目標状態の特徴表現間の類似度を算出する．その上で，目標状態への変化が大きい遺伝子を重要遺伝子として特定する手順で実験を行う．評価指標には，cosine_shift（↑）および p_value（↓）を用いた．cosine_shift はコサイン類似度の変化量であり，正の値が大きいほど目標状態へ近づくことを意味する．また，p_value は統計的有意性を評価する指標であり，本研究では $p < 0.05$ を統計的に有意とする．

本研究では，単一生物種内の評価として，マウスの心臓疾患データでファインチューニングしたモデルを用い，マウスにおける心臓疾患状態から正常状態へ変化させる in silico 摂動実験を行う．また，生物種間の横断性の評価として，マウスの心臓疾患データでファインチューニングしたモデルをヒトの心臓疾患データに適用し，心臓疾患状態から正常状態へ変化させる in silico 摂動実験を行う．前者の実験は遺伝子の削除，後者の実験は遺伝子の過剰発現による実験を行っている．これらの実験において，Mix-Geneformer が重要と判定し，実際の生物実験で有効性が確認された遺伝子の一部を表 3 に，摂動前，摂動後，および目標状態における細胞の特徴表現を UMAP により可視化した結果を図 4 に示す．

表 3: in silico 摂動実験で確認された有効遺伝子の一部

使用モデル	遺伝子名	cosine_shift	p_value
マウス	ALDOB	0.011	1.56E-2
マウス	ALDH3B2	0.011	9.03E-3
ヒト	MTRNR2L11	0.202	0.0
ヒト	NAP1L6	0.035	1.66E-03



(a) ヒトで FT したモデルによる実験　(b) マウスで FT したモデルによる実験

図 4: in silico 摂動実験における UMAP 可視化

表 3，図 4 から，マウスおよびヒトデータでファインチューニングしたモデルの両者ともに，摂動後の細胞の特徴表現は摂動前と比較して目標状態に近づいた．cosine_shift の摂動前後の変化は，図 4(a) においては約 0.43，図 4(b) においては約 0.05 であり，定量的，定性的評価ともに in silico 摂動実験の成功を確認した．一方で，ヒトデータでファインチューニングしたモデルでは，UMAP によるクラスタ間の分離がより明瞭であり，摂動前後における細胞の特徴表現の変化量も大きいことを確認した．これは，マウスデータで学習したモデルをヒトデータに適用する際に，生物種の違いに起因するドメインギャップが存在する可能性を示唆している．この差異を解消するには，マウスとヒト間でのデータ正規化や，種間の関係性学習のための損失を定義する必要があると考える．

5. おわりに

本研究では，ヒトおよびマウスの scRNA-seq データを同一の Transformer で事前学習するモデル Mix-Geneformer を提案した．細胞型分類タスクにおいて，Mix-Geneformer は従来モデルと同等以上の精度を示し，同一種内における性能の有効性を確認した．また，生物種の横断性の評価として，UMAP による特徴表現の可視化を行った結果，異なる生物種でファインチューニングしたモデルであっても，一定程度のクラスタ構造が保持されることを示した．in silico 摂動実験においても同様の傾向を確認し，マウスデータでファインチューニングしたモデルでもヒトの in silico 摂動実験が可能である．一方で，同一種のデータでファインチューニングを行った場合と比較すると性能に差が見られることから，生物種の差異が結果に影響する可能性が示唆される．以上より，Mix-Geneformer は生物種を横断した scRNA-seq 解析が可能である一方で，種間差異をより適切に扱うための学習手法の設計が今後の課題である．具体的には，生物種の差異に起因する影響の緩和，ヒトおよびマウス以外の生物種への拡張が挙げられる．

参考文献

[1] C. V. Theodoris, *et al.*, “Transfer learning enables predictions in network biology,” *Nature*, 2023.

[2] K. Ito, *et al.*, “Mouse-Geneformer: A deep learning model for mouse single-cell transcriptome and its cross-species utility,” *PLOS Genetics*, 2025.

研究業績

[1] 西尾優希，山下隆義，伊藤啓太，平川翼，藤吉弘亘，”Mix-Geneformer: Unified Representation Learning for Human and Mouse scRNA-seq Data”，IIBMP，2025