細胞型間の関係を考慮した mouse-Geneformer に関する研究

TP23001 伊藤 啓太　　指導教授：山下 隆義

1. はじめに

疾患の原因究明には，疾患を再現した生物の遺伝子やタンパク質の解析が行われる [1]．単一細胞レベルでの解析技術が進展したことで，細胞の遺伝子発現量が単一細胞データとして取得でき，遺伝子解析等に活用されている．ヒトの単一細胞データの遺伝子解析手法として，Geneformer [2] が提案されている．これはヒトの遺伝子間の関係を学習した深層学習モデルで，遺伝子発現量から細胞型の分類や in silico 摂動実験が可能である．一方で，疾患の再現にはマウスなどの生物が使用されるため，疾患研究の進展にはマウスの単一細胞データの解析が必要である．しかし，ヒトの単一細胞データを学習した Geneformer ではマウスの単一細胞データの解析は困難である．そこで本研究では，マウスの単一細胞データの解析を目的とした mouse-Geneformer を Geneformer に倣い構築する．

単一細胞データの解析は，遺伝子間だけでなく細胞型間の関係の解析も重要であり [3]，細胞型間の関係の解析により，胚の発生過程や疾患の進行による細胞状態に関する知見が得られる．そこで本研究では，マウスの遺伝子間の関係と細胞型間の関係を学習した mouse-Geneformer++ も構築する．mouse-Geneformer++ は，mouse-Geneformer よりもマウス細胞を細胞型や疾患状態に分類する事ができ，かつ mouse-Geneformer による in silico 摂動実験よりも正確な in silico 摂動実験が可能となる．

2. Geneformer

Geneformer [2] は，ヒトの単一細胞データの遺伝子解析を目的とした深層学習モデルである．大規模なヒトの単一細胞データセットにおいて各細胞の遺伝子発現量を降順に順位付けし，細胞特有の遺伝子群に変換する．この遺伝子群の一部の遺伝子（単語）をマスクしモデルへ入力する事で，マスクした遺伝子の特徴ベクトルを出力し，マスクした単語を予測する Masked Language Model (MLM) で正常なヒトの遺伝子間の関係を学習する．このモデルを，特定の臓器の単一細胞データで細胞型分類タスクにファインチューニングすることで，従来手法より正確な細胞型分類ができることを示している．また，疾患を患ったヒトの単一細胞データの in silico 摂動実験では，疾患原因遺伝子の抽出ができることを示している．

3. mouse-Genecorpus-20M

マウスの単一細胞データを深層学習モデルで解析するには，大規模なマウスの単一細胞データセットで事前学習する必要がある．しかし，大規模なマウスの単一細胞データセットは存在しない．そこで，本研究では，正常なマウスの単一細胞データで構成した大規模なデータセットとして mouse-Genecorpus-20M を構築する．具体的には，まず一般公開されている 6 つの単一細胞データベースから，臓器と年齢が多様なマウスの単一細胞データを 119,099,757 個取得する．次に，取得した単一細胞データから単一細胞ではないデータを除去する．残った 20,630,028 個の単一細胞データを mouse-Genecorpus-20M として構築する．

4. 提案手法

本研究では，マウスの単一細胞データの解析を目的とした Geneformer モデルを 2 つ提案する．1 つ目は Geneformer に倣い構築した mouse-Geneformer，2 つ目はマウスの遺伝子間の関係と細胞型間の関係を MLM と SimCSE++ で学習した mouse-Geneformer++ である．両モデルにより，マウス細胞の細胞型分類精度の向上やマウスを用いた in silico 摂動実験が可能となる．

4.1 mouse-Geneformer

mouse-Genecorpus-20M の各単一細胞データを細胞特有の遺伝子群に変換する．変換した遺伝子群に対して一部の遺伝子を隠すマスクトークンに置換し，遺伝子群の先頭に遺伝子群全体の特徴を表現するクラストークンを付与する．そして，6 層の Transformer Encoder Block で構成したモデルに入力する．モデルが出力するマスクトークンの特徴ベクトルを全結合層に入力しマスクした遺伝子を予測する MLM で事前学習する．この学習により，正常なマウスの遺伝子間の関係を学習できる．

4.2 mouse-Geneformer++

mouse-Geneformer は，マウスの遺伝子間の関係のみを学習する．しかし，単一細胞データの解析は，遺伝子間だけでなく細胞型間の関係の解析も重要である [3]．そのため，mouse-Geneformer++ では，MLM に加え細胞型間の関係を学習するために一般文章の対照学習手法である SimCSE++ [4] を mouse-Geneformer に導入する．

SimCSE++　SimCSE++ は，ドロップアウトによるランダムな変化を利用してデータ拡張を行い，文章の特徴ベクトルに対する対照学習と，各特徴ベクトルの同じ次元の特徴量に着目したベクトルに対する対照学習を行う手法である．これにより，文章の多様な特徴を捉えた特徴ベクトルを獲得する．図 1 に SimCSE++ の概要図を示す．

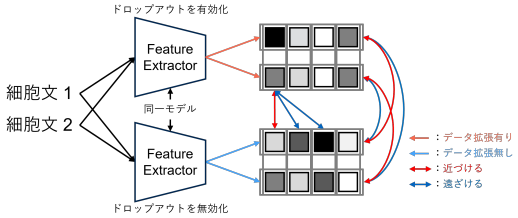


図 1：SimCSE++ の概要図.

特徴ベクトルに対する対照学習の正例ペアは，データ拡張した文章とデータ拡張していない同じ文章の特徴ベクトルである．負例ペアは，データ拡張した文章とデータ拡張していない別の文章の特徴ベクトルである．これにより，類似した意味を持つ文章をベクトル空間上で近い位置に，異なる意味を持つ文章をベクトル空間上で遠い位置に配置できる．

同じ次元の特徴量に着目したベクトルに対する対照学習は，データ拡張した文章の特徴ベクトルの同じ次元に着目したベクトル群と，データ拡張していない文章の特徴ベクトルの同じ次元に着目したベクトル群で行う．正例ペアは 2 つのベクトル群の同じ次元同士，負例ペアは 2 つのベクトル群の別の次元同士である．これにより，特徴量ごとの独立性が強調され，文章の多様な特徴を獲得できる．

事前学習　mouse-Genecorpus-20M の各単一細胞データを細胞特有の遺伝子群に変換する．変換した遺伝子群に対して一部の遺伝子を隠すマスクトークンに置換し，遺伝子群の先頭に遺伝子群全体の特徴を表現するクラストークンを付与する．そして，6 層の Transformer Encoder Block で構成したモデルに入力する．モデルの出力特徴ベクトルを用いて MLM と SimCSE++ で同時に事前学習する．SimCSE++ の学習は，クラストークンの特徴ベクトルに対して行う．この学習により，正常なマウスの遺伝子間の関係と細胞型間の関係を学習できる．

5. 評価実験

2 つの提案手法の有効性を検証するために事前学習の有無による細胞型分類実験，従来手法と提案手法による細胞型分類実験，提案手法による in silico 摂動実験を行う．

5.1 事前学習の有無による細胞型分類実験

本実験では，事前学習した 2 つの提案手法を細胞型分類タスクのデータセットを用いてファインチューニングする．使用するデータセットは，マウスの胚，腎臓-1，尿道と前立腺が混合（尿&前）した単一細胞データセットで，いずれも mouse-Genecorpus-20M に含まれていない．データセットを学習用として 80%，評価用として 20% に分割して

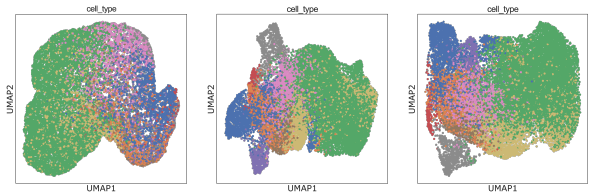
実験する．比較手法は，事前学習有りの mouse-Geneformer（MLM）と mouse-Geneformer++（MLM&SimCSE++），および事前学習なしの mouse-Geneformer である．なお，mouse-Geneformer と mouse-Geneformer++ の違いは SimCSE++ の有無のみであり，アーキテクチャは同一である．そのため，事前学習なしの場合は同一モデルとして評価ができる．定量評価は Accuracy を使用し，定性評価は胚データの特徴ベクトルを可視化する．

まず，事前学習の有無による細胞型分類の比較結果を表 1 に示す．表 1 から，事前学習有りの MLM と MLM&SimCSE++ の Accuracy が事前学習なしより高い事が分かる．このことから，MLM と MLM&SimCSE++ の事前学習は有効である事が分かる．また，事前学習有りの MLM&SimCSE++ の方が事前学習有りの MLM よりも Accuracy が高い．このことから，MLM&SimCSE++ の事前学習の方がより有効である．

次に，特徴ベクトルの可視化結果を図 2 に示す．図 2 の点は各細胞の特徴ベクトル，色は各細胞型を示す．図 2 の事前学習なしは細胞型に分離していない．一方，MLM や MLM&SimCSE++ は灰色や青色の細胞型が分かれている．中でも MLM&SimCSE++ は灰色の細胞型が分離している．このことから，MLM&SimCSE++ は，細胞型間の関係も学習したと言える．

表 1：事前学習の有無による細胞型の分類結果の比較

臓器	細胞型数	事前学習なし	事前学習有り	
			MLM	MLM&SimCSE++
尿&前	7	93.73	94.70	**95.04**
胚	9	66.56	75.24	**77.91**
腎臓-1	10	78.91	79.02	**80.14**



(a) 事前学習なし　(b) MLM　(c) MLM&SimCSE++

図 2：胚データの特徴ベクトルの可視化結果．

5.2 従来手法と提案手法による細胞型分類実験

本実験では，事前学習した 2 つ提案手法を細胞型分類タスクのデータセットを用いてファインチューニングする．使用するデータセットは，マウスの舌，胸腺，乳腺，大腸，筋肉，脾臓，心臓，脳，腎臓-2 の単一細胞データセットで，いずれも mouse-Genecorpus-20M に含まれていない．データセットを学習用として 80%，評価用として 20% に分割して実験する．比較には従来手法として，scDeepSort (scDS)，Single-cell VAE (scVAE)，提案手法として mouse-Geneformer (m-G)，mouse-Geneformer++ (m-G++) を用いる．定量評価は Accuracy を使用する．

従来手法と提案手法の細胞型分類の比較結果を表 2 に示す．表 2 から m-G と m-G++ の Accuracy が scDS や scVAE よりも高い事が分かる．このことから，m-G と m-G++ の細胞型分類タスクは有効である．また，m-G++ の Accuracy は m-G と同等かそれ以上であるデータが多い．このことから，m-G++ の方がより有効である．

表 2：従来手法と提案手法による分類結果の比較

臓器	scDS	scVAE	m-G	m-G++
舌	76.69	80.44	**94.87**	94.72
胸腺	54.94	74.95	**96.97**	**96.97**
乳腺	47.76	74.25	**99.02**	**99.02**
大腸	49.78	59.00	93.08	**94.46**
筋肉	90.82	79.58	99.52	**99.63**
脾臓	81.01	76.47	98.70	**98.82**
心臓	79.55	79.42	97.82	**97.87**
脳	58.46	76.19	96.92	**97.41**
腎臓-2	58.01	56.25	**94.88**	94.76
平均	66.33	72.95	96.86	**97.07**

5.3 in silico 摂動実験

in silico 摂動実験とは，コンピュータ上で遺伝子に摂動を加えて細胞状態の変化をシミュレートする実験である．遺伝子の摂動で変化した細胞状態と特定の細胞状態のコサイン類似度が 1 に近づいた時，摂動した遺伝子は疾患の原因となる重要な遺伝子と予測できる．

本実験では，事前学習した 2 つの提案手法を疾患状態を分類するタスクのデータセットを用いてファインチューニングする．使用するデータセットは，アルツハイマー型認知症と関係のある Cop1 遺伝子をノックアウト (KO) したマウスの単一細胞データと正常なマウスの単一細胞データが混合したデータセットで，いずれも mouse-Genecorpus-20M に含まれていない．データセットを学習用として 80%，評価用として 20% に分割して実験する．評価用データの分類精度が 85% 以上のファインチューニングモデルで Cop1 遺伝子を KO した細胞集団を正常な細胞集団に近づける in silico 摂動実験を行う．生物実験 (in vivo 実験) の結果と比較し評価をする．in silico 摂動実験で抽出する遺伝子はウィルコクソン検定で p 値が 5% 未満の遺伝子とする．

提案手法による in silico 摂動実験の結果と in vivo 実験の結果を表 3 に示す．表 3 の「in vivo 実験の結果」は in vivo 実験で確認した遺伝子を示し，「存在」は抽出した遺伝子群に in vivo 実験で確認した遺伝子があるかを示し，「類似度」は遺伝子を摂動した時のコサイン類似度を示す．表 3 から，in vivo 実験で確認した遺伝子が mouse-Geneformer と mouse-Geneformer++ で抽出した遺伝子群に存在する事が分かる．このことから，提案手法による in silico 摂動実験において異常な遺伝子の抽出が可能であると言える．また，mouse-Geneformer++ は mouse-Geneformer よりも多くの遺伝子を抽出し，かつそれぞれの遺伝子を摂動した時のコサイン類似が高い．このことから，mouse-Geneformer++ の in silico 摂動実験の方がより効果的であると言える．

表 3：in vivo 実験の結果と in silico 実験の結果の比較

in vivo 実験の結果	mouse-Geneformer		mouse-Geneformer++	
	存在	類似度	存在	類似度
Apoe	✓	0.197	✓	**0.372**
Fth1	✓	0.018	✓	**0.172**
Cle7a	-	-	-	-
Cst7	✓	**0.003**	✓	**0.003**
Ifitm1	-	-	✓	**0.038**
Ifitm3	-	-	-	-
Cxcl10	-	-	✓	**0.005**
Itgax	✓	0.004	✓	**0.023**
Il12b	-	-	✓	**0.003**

6. おわりに

本研究では，マウスの単一細胞データを解析する mouse-Geneformer と mouse-Geneformer++ を提案し，マウスの単一細胞データの解析に有効であることを示した．さらに，mouse-Geneformer++ の方がマウスの単一細胞データの解析に有効であることも示した．今後は，各生物間の遺伝子の関係を学習した Geneformer の構築を検討する．

参考文献

[1] Domínguez-Oliva A, *et al.*, “The Importance of Animal Models in Biomedical Research: Current Insights and Applications.” Animals, 2076-2615, 2023.

[2] Theodoris, *et al.*, “Transfer learning enables predictions in network biology.” Nature 618, 616–624 (2023).

[3] Miao, Z., *et al.* “Putative cell type discovery from single-cell gene expression data.” Nat Methods 17, 621–628 (2020).

[4] XU, *et al.*, “SimCSE++: Improving contrastive learning for sentence embeddings from two perspectives.”, arXiv preprint arXiv:2305.13192, 2023.

研究業績

K. Ito, *et al.*, “Mouse-Geneformer: A Deep Leaning Model for Mouse Single-Cell Transcriptome and Its Cross-Species Utility”, bioRxiv, 2024.