Mamba を用いたマウスの単一細胞解析のための基盤モデル構築

EP22064 小林 岳隼　　指導教授：山下 隆義

1. はじめに

次世代シーケンサを用いた Single-cell RNA sequencing (scRNA-seq) 解析の進展により，単一細胞内の遺伝子発現量の取得が可能となった．これに伴い，深層学習を用いた遺伝子解析技術も発展し，マウスの単一細胞データで事前学習を行った基盤モデルとして Mouse-Geneformer[1] が提案されている．本モデルは，遺伝子間の複雑な関係性を学習しており，細胞型の分類や in silico 摂動実験において高い性能を示す．しかし，そのアーキテクチャの基礎である Transformer は，計算量が入力長の 2 乗に比例して増大する．これにより，膨大なメモリ使用量がボトルネックとなり，現実的な計算リソースでは扱える遺伝子数に実質的な制約が生じる．そこで本研究では，入力長に対して線形な計算量で動作し，長い入力長でも効率的な学習が可能な Mamba[2] モデルを採用した，Mouse-GeneMamba を提案する．

2. Mouse-Geneformer

Mouse-Geneformer はマウスの単一細胞データの遺伝子解析を目的とした基盤モデルである．学習には大規模なマウスの単一細胞データセットである Mouse-Genecorpus-20M を用いる．Mouse-Genecorpus-20M では，各細胞内の遺伝子発現量の上位 2,048 個の遺伝子を抽出し，遺伝子トークン列に変換することで細胞文とする．作成した細胞文に対して，Masked Language Modeling (MLM) で学習を行うことで，正常なマウスの遺伝子間の関係を学習できる．さらに，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た．

3. 提案手法：Mouse-GeneMamba

本研究では，Mouse-Geneformer の Transformer Encoder を Mamba ブロックに置換した Mouse-GeneMamba を提案する．本手法の全体概要を図 1 に示す．入力データの構築において，順位情報を持つ遺伝子トークンと正規化した遺伝子発現量をそれぞれベクトル化して統合することで，各遺伝子の順位と大きさを両方含む細胞文を作成する．次に，この細胞文を Mamba ブロックへ入力して，長大な遺伝子配列の大域的な文脈学習を行う．学習タスクには Next Token Prediction (NTP) を採用し，過去の遺伝子配列から次の遺伝子を予測することで，遺伝子ネットワークの因果関係を獲得する．また，本モデルの学習には， Mouse-Genecorpus-20M を拡張し，正規化された遺伝子発現量の数値を保持した大規模データセットを用いる．

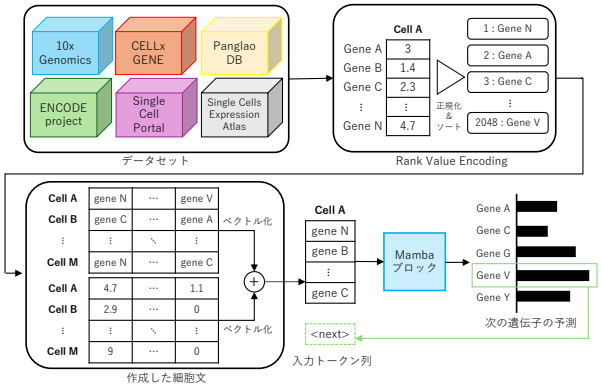

図 1： Mouse-GeneMamba の学習方法

4. 評価実験

提案手法の有用性を検証するために，複数の評価実験を行う．いずれの実験においても，事前学習モデルを細胞型分類タスクのデータセットでファインチューニングし，分類精度により評価する．事前学習タスクとして NTP を用いたモデルと MLM を用いたモデルを用いて，タスクの違いがモデル性能に与える影響を比較する．また，遺伝子発現量をモデル入力に統合することの有効性を検証する．具体的には，発現量の有無による精度比較に加え，Mouse-Geneformer との比較を行う．

4.1 評価実験結果

事前学習タスクとして NTP と MLM で学習したモデルの細胞型分類タスクの結果を表 1 に示す．表 1 より，事前学習タスクとして NTP を採用したモデルは多くの臓器で最も高い精度を達成した．

表 1： 入力長 2,048 における事前学習タスクによる性能比較

事前学習タスク	脳	四肢の筋肉	腎臓	胸腺	舌	乳腺	心臓	脾臓	大腸	平均
NTP	97.6	99.6	95.3	96.8	94.2	98.8	98.1	98.7	94.2	97.0
MLM	97.9	99.5	94.6	97.2	94.7	98.9	97.5	98.5	93.1	96.9

提案手法の発現量の有無と Mouse-Geneformer の細胞型分類タスクの結果を表 2 に示す．各臓器の分類において，最も高い精度を赤色，低い精度を青色で示す．表 2 より，遺伝子発現量の有無の観点では，発現量を考慮しない設定においてより高い分類精度を示した．この結果から，遺伝子発現量を学習に用いる場合，本研究で採用した方法とは異なる利用方法を検討する必要がある．一方で，Mouse-Geneformer との比較においては，提案手法の方が平均精度において最も高い精度を達成し，有用性を確認した．

各入力長における精度変化に着目すると，Mouse-Geneformer は入力長を 2,048 から 8,192 に拡張した際，平均精度が 0.53 ポイント低下したのに対し，提案手法は 0.20 ポイントの低下に留まった．以上の結果から，本手法に用いている Mamba モデルは長い入力長に対しても情報の損失を抑えつつ特徴を抽出できることを示した．

表 2： 提案手法と Mouse-Geneformer の細胞型分類精度

モデル	Mouse-GeneMamba						Mouse-Geneformer		
発現量	なし			あり			なし		
入力長	2,048	4,096	8,192	2,048	4,096	8,192	2,048	4,096	8,192
脳	97.6	98.0	97.8	96.7	96.4	95.7	96.7	96.2	95.3
四肢の筋肉	99.6	99.7	99.6	99.5	99.4	99.3	99.6	99.5	99.7
腎臓	95.3	94.4	95.1	93.8	93.2	92.5	94.8	94.7	94.1
胸腺	96.8	97.3	96.9	94.9	96.1	96.2	96.8	96.8	97.3
舌	94.2	94.4	93.7	92.8	92.2	91.1	94.6	94.4	93.9
乳腺	98.8	98.8	98.7	98.5	98.1	98.4	98.9	99.1	98.9
心臓	98.1	97.3	97.2	96.5	97.0	96.9	96.8	96.4	96.9
脾臓	98.7	98.5	98.6	98.3	97.9	97.8	98.6	98.6	98.5
大腸	94.2	94.3	93.9	93.3	93.4	93.4	92.3	92.3	89.8
平均	97.0	97.0	96.8	96.0	96.0	95.7	96.6	96.4	96.0

事前学習におけるメモリ使用量を表 3 に示す．表 3 より，提案手法は Mouse-Geneformer に比べてメモリ効率が向上しており，Mamba モデルを用いる有効性を確認した．

表 3： 事前学習におけるメモリ使用量

モデル	Mouse-GeneMamba			Mouse-Geneformer		
入力長	2,048	4,096	8,192	2,048	4,096	8,192
メモリ使用量 (↓)	10.4GB	24.4GB	36.6GB	16.8GB	32.5GB	out of memory

5. おわりに

本研究では，Mouse-Geneformer の高いメモリ使用量や入力長の制限という問題を解決するためのモデルである Mouse-GeneMamba を提案した．また，発現量の値を考慮した新たなデータセットを構築し，そのデータで学習および細胞型の分類実験を行うことで，発現量を考慮する学習の有効性を検証した．

今後は，別の発現量の入力方法での学習や Mamba の内部構造の変更，データセットの大規模化を実施することで，モデルの分類精度向上を目指す．加えて，多様な下流タスクによる検証を行うことで，基盤モデルとしての汎用性と有用性を実証していく．

参考文献

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

[2] Albert Gu, *et al.*, “Mamba: Linear-Time Sequence Modeling with Selective State Spaces”, COLM, 2024.