PO-002 **Mix-Geneformer : Unified Representation Learning for Human and Mouse scRNA-seq Data**

西尾 優希，山下 隆義，伊藤 啓太，平川 翼，藤吉 弘亘

研究背景

■ 単一細胞RNA-seqの解析技術発展により細胞ごとの詳細な解析が可能

- Geneformer，Mouse-Geneformer 等の遺伝子解析モデルが発達

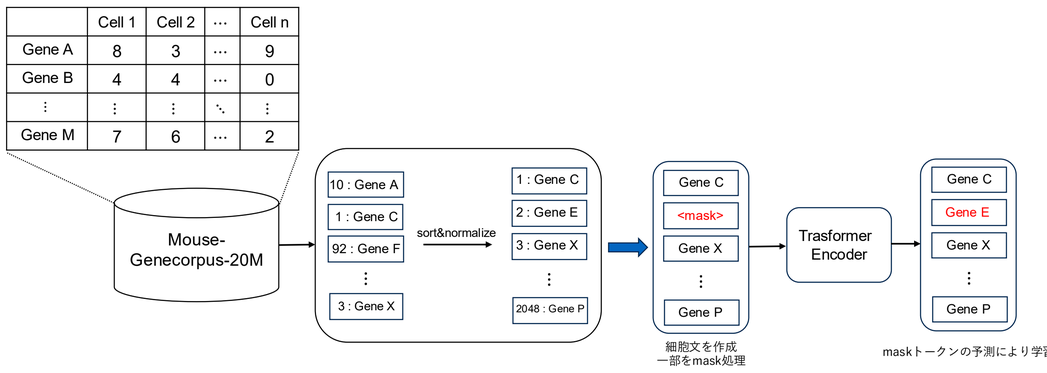
■ Geneformer [Theodoris+, nature'23]， Mouse-Geneformer [K.Ito+, PLOS Genetics'25]

- Transformerを遺伝子解析に応用したモデル
- 各生物の細胞データにおいて遺伝子ネットワークの正確な予測が可能
- ヒト，マウスの単一細胞データセットを作成
  - Masked Language Modeling (MLM)により学習

→ 単一の種に特化しており，マウス↔ヒトの横断性が課題

■ Mouse-Geneformer

- マウスの単一細胞データ Mouse-Genecorpus-20MをMLMにより学習



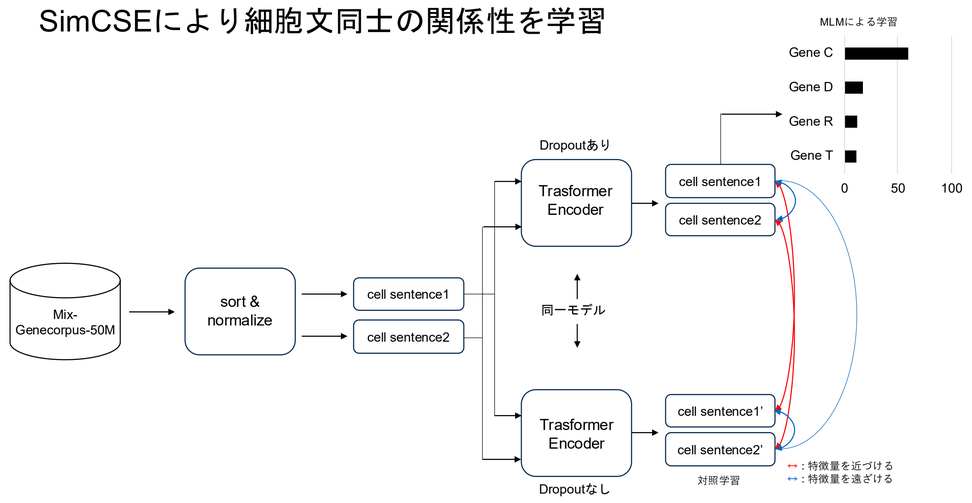
提案手法：Mix-Geneformer

■ Mix-Geneformerの概要

- マウスとヒトの単一細胞データで学習するTransformer
- ヒト，マウスの細胞型分類，in silico摂動実験が可能
  - in silico摂動実験：コンピュータ上で行う遺伝子のシミュレーション実験

■ Mix-Geneformerの事前学習

- MLMとSimCSEにより学習
  - MLMにより細胞文内の文脈を学習
  - SimCSEにより細胞文同士の関係性を学習

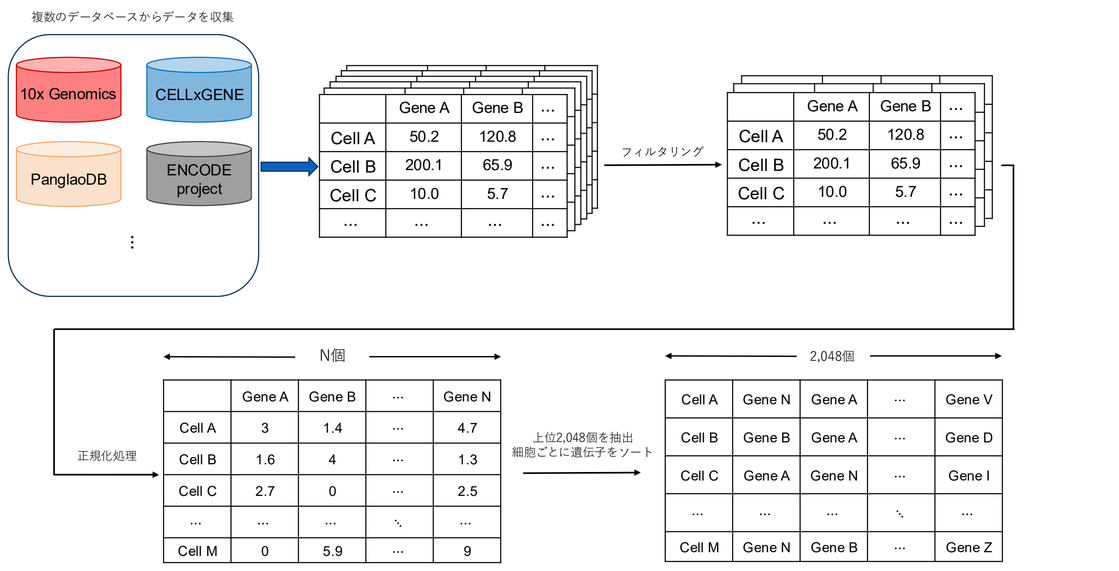


Mix-Geneformerの学習方法

学習データセット：Mix-Genecorpus-50M

■ マウスとヒトの統合データMix-Genecorpus-50Mを作成

- 収集データ内のノイズとなるデータをフィルタリング
- 正規化を行ったのちにN個のデータを降順ソート
- 細胞を特徴づける遺伝子を2,048個抽出



大まかな学習データセット作成の流れ

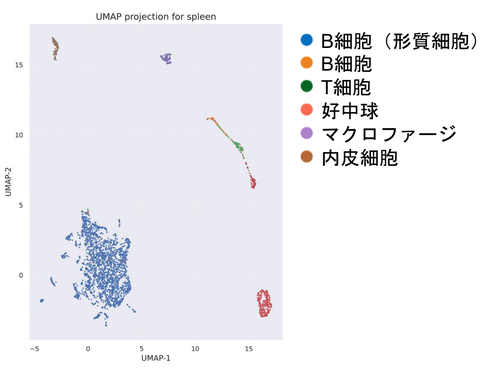
評価実験

■ 細胞型分類による評価

- 比較対象：Geneformer，Mouse-Geneformer，対照学習あり，なしのMix-Geneformer
- 評価指標：Accuracy

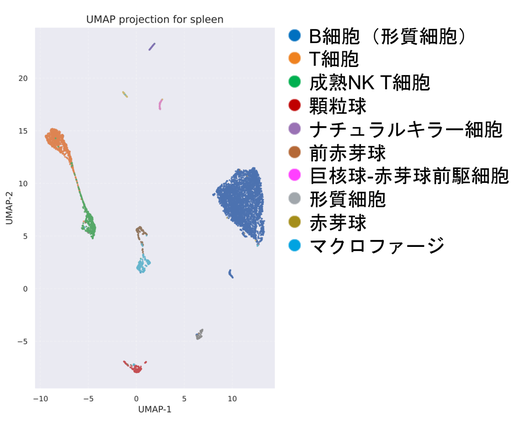
■ UMAPによる特徴空間の可視化

臓器名	細胞の種類	Mix-Geneformer (w/ CL)	Mix-Geneformer (w/o CL)	Human-Geneformer
脳	6	95.9	96.0	96.8
肝臓	12	88.0	87.9	91.1
腎臓	15	92.2	89.7	92.8
大腸	16	91.5	89.7	92.7
胎盤	10	97.4	96.1	97.9
膵臓	15	92.2	90.0	93.0
脾臓	6	98.3	97.9	98.9
肺	16	92.6	91.5	93.4
免疫系	10	93.6	91.1	94.4



GeneformerとMix-Geneformerの分類精度の比較とヒトの脾臓データにおけるMix-Geneformerの特徴空間

臓器名	細胞の種類	Mix-Geneformer (w/ CL)	Mix-Geneformer (w/o CL)	Mouse-Geneformer
脳	15	97.3	96.7	96.9
心臓	11	97.7	97.2	97.8
腎臓	18	95.6	94.4	94.9
大腸	7	94.7	93.1	94.3
四肢の筋肉	9	99.6	99.4	99.5
乳腺	7	99.1	99.1	99.0
脾臓	10	98.7	98.3	98.7
胸腺	6	97.5	97.2	97.0
舌	3	94.8	93.6	94.9



Mouse-GeneformerとMix-Geneformerの分類精度の比較とマウスの脾臓データにおけるMix-Geneformerの特徴空間

Geneformerと比較してやや精度が劣るが全ての臓器において90%ほどの分類精度を達成
Mouse-Geneformerと比較してほとんどの臓器において分類精度が向上
UMAPによる分析ではそれぞれの細胞の特徴を分離可能

→ Mix-Geneformerは複数種の細胞型分類を高精度に行うことが可能

■ in silico摂動実験による評価

- 評価方法：in vivo実験の結果とin silico摂動実験の結果を従来モデルと比較
  - in vivo実験：現実世界で行われる生物実験
  - 評価指標：cosine_shift（↑），p_value（↓）
- マウスデータでのin silico摂動実験　：腎臓病を正常にする実験
  - ADPKDとDKDの2種類を検証
- ヒトデータでのin silico摂動実験　：アルツハイマーを正常にする実験
- 比較対象：Geneformer，Mouse-Geneformer

遺伝子名	疾患名	Mix-Geneformer cosine_shift	Mix-Geneformer p_value	Mouse-Geneformer cosine_shift	Mouse-Geneformer p_value
Umod	ADPKD	0.0237	2.54e-64	-0.0048	2.67e-08
Sgk1	DKD	0.0161	1.47e-04	0.0019	0.174
Cfb	DKD	0.0132	1.72e-04	0.0014	6.08e-11
Rock2	DKD	0.0082	1.60e-05	0.0019	0.148

マウスデータでのin silico摂動実験

遺伝子名	疾患名	Mix-Geneformer cosine_shift	Mix-Geneformer p_value	Human-Geneformer cosine_shift	Human-Geneformer p_value
HSPB1	Alzheimer	0.0544	1.58e-13	0.0033	3.66e-04
S100B	Alzheimer	0.0425	5.78e-05	0.0004	0.962
HSP90AA1	Akzheimer	0.0143	9.99e-07	0.0095	1.22e-05

ヒトデータでのin silico摂動実験

Mix-Geneformerはマウスとヒトの複数種の生物を用いたin silico摂動実験が可能
→従来モデルよりも実用性が向上

臓器データや神経データ等の様々なドメインでのin silico摂動実験が可能
→幅広い応用性を獲得

■ まとめ

- マウスとヒトのデータで学習を行うMix-Geneformerを提案
- 複数の種での細胞型分類やさまざまなドメインでのin silico摂動実験が可能
- 今後の予定
  - ヒトとマウスの細胞間の特徴の差を考慮した学習方法の提案