Biologia Systemów 2024/25
Spatial Proteomics, zadanie dodatkowe

M. Możejko, K. Gogolewski

SCVI (3 pkt)

W zadaniu korzystaliśmy z redukcji wymiarowości opartej na PCA.
W tym dodatkowym zadaniu należy:

- Wytrenować model **SCVI** (`https://docs.scvi-tools.org/en/latest/user_guide/models/scvi.html`) na danych z ekspresji markerów,
- Użyć wytrenowanego modelu w algorytmie **CellCharter**,
- Porównać wyniki z wynikami uzyskanymi przy użyciu PCA — w podobny sposób jak porównywaliśmy je z bazowymi modalnościami (średnia ekspresja markerów, histogramy typów komórkowych).

BERT-Charter (7 pkt)

W tym zadaniu należy:

1. **Przygotować** `torch.Dataset`, który:
   - dla i-tej komórki oraz n_layers zwróci n_layers zbiorów komórek:
     - odległość 0 — oryginalna komórka,
     - odległość 1 — najbliżsi sąsiedzi,
     - ...,
     - odległość $n\_layers - 1$ — komórki oddalone o długość najkrótszej ścieżki $n\_layers - 1$ w grafie.
   - zwróci sekwencję wszystkich ekspresji markerów komórek oraz numer warstwy +1 każdej z komórek (centralna komórka = 1, bezpośredni sąsiedzi = 2, itd.).
2. **Przygotować** `NeighborhoodEmbedder` **(torch.Module)**, który:
   - przyjmie zbiór z poprzedniego punktu,
   - zwróci sekwencję o wymiarze `embedding_dim`,
   - dla każdej komórki:
     - obliczy sumę liniowego embeddingu ekspresji markerów,
     - doda uczalne zanurzenie numeru warstwy,
     - dla komórek z numerem warstwy 0 — zastąpi embedding ekspresji maskującym wektorem (`mask vector`) o wymiarze `embedding_dim`.
3. **Przygotować moduł** `BertCharter` **(torch.Module)**, który:
   - korzysta z `NeighborhoodEmbeddera`,
   - korzysta z `TransformerEncoder` z:
     - `hidden_dim` domyślnie 128 (równym `embedding_dim`),
     - 4 warstwami,
     - 4 głowami.
   - Wyjście ostatniej warstwy encodera przekazuje do warstwy liniowej o wymiarze liczby markerów (40),
   - Zwraca:
     - sekwencję wyjściową z ostatniej warstwy encodera,
     - sekwencję do predykcji markerów.
4. **Napisać pętlę trenującą**, która:

- bazuje na `DataLoaderze` z `Datasetu` z punktu 1,
- w każdej iteracji:
  - losowo maskuje określony procent komórek (domyślnie 15%),
  - zamienia numer warstwy zamaskowanych komórek na 0,
  - `BertCharter` przewiduje z powrotem wszystkie markery (zamaskowane i niezamaskowane),
  - funkcja kosztu: MSE lub MAE na wszystkich markerach,
  - wagi:
    - 1 dla zamaskowanych markerów,
    - 0.1 dla niezamaskowanych.
- Opcjonalnie: dodać zbiór walidacyjny.

5. **Przy pomocy wytrenowanego modelu**:

- obliczyć reprezentację otoczenia każdej komórki - przez uśrednienie wyjścia z ostatniej warstwy encodera wzdłuż sekwencji,
- porównać reprezentacje otoczeń:
  - z otoczeniami uzyskanymi z bazowej modalności,
  - z wynikami CellChartera.

Oddawanie rozwiązań Deadline na przesłanie rozwiązań - 31 maja 2025 roku do godziny 20:00.
Rozwiązania oraz ewentualne pytania należy przesyłać na adres: **marcin.mozejko@student.uw.edu.pl**, z kopią do: **k.gogolewski@mimuw.edu.pl**. W tytule wiadomości (zarówno z pytaniami, jak i oddawanymi rozwiązaniami) prosimy umieszczać: `[BiolSys] Spatial proteomics - zadanie dodatkowe`.