Incorporating Hierarchical Information into Multiple Instance Learning for Patient Phenotype Prediction with scRNA-seq Data

https://www.biorxiv.org/content/10.1101/2025.02.10.637389v1.full.pdf

ARTIFICIAL INTELLIGENCE AND DATA ANALYTICS LAB
명지대학교 융합소프트웨어학부 김민

* Our implementation is available at https://github.com/minhchaudo/hier-mil.

Experiments and Results

- DATA

1. *Cardio: 심근병증 환자 데이터를 활용한 다중 분류 (DCM, HCM, 정상; 세 가지 분류)*
2. *ICB: 면역항암제 치료 반응 여부를 예측하는 이진 분류*
3. **COVID: COVID-19 감염 여부를 예측하는 이진 분류**

https://singlecell.broadinstitute.org/single_cell/study/SCP1289/impaired-local-intrinsic-immunity-to-sars-cov-2-infection-in-severe-covid-19#study-download

20210701_NasalSwab MetaData.txt

Columns (27개)

컬럼명	설명
NAME	세포 또는 샘플의 고유 이름 (예: 바코드, cell ID 등)
donor_id	세포를 제공한 환자(또는 샘플 제공자)의 ID
Peak_Respiratory_Support_WHO_Score	WHO 기준에 따른 최대 호흡기 보조 수준 (0: 없음 ~ 7: 인공호흡기)
Bloody_Swab	채취한 면봉에 혈액이 있었는지 여부 (`Yes` / `No`)
Percent_Mitochondrial	세포 내 미토콘드리아 유전자 발현 비율 (품질 지표로 사용됨)
SARSCoV2_PCR_Status	환자 단위의 PCR 감염 여부 (`pos` / `neg`)
SARSCoV2_PCR_Status_and_WHO_Score	PCR 감염 여부 + WHO 점수를 합친 값 (예: `pos_5`)
Cohort_Disease_WHO_Score	질병 유형 + WHO 점수를 합친 값 (예: `COVID_WHO_5`)
biosample_id	동일한 샘플(코 swab 등)의 고유 ID
SingleCell_SARSCoV2_RNA_Status	개별 세포 내 SARS-CoV-2 RNA 유무 (`pos` / `neg`)
SARSCoV2_Unspliced_TRS_Total_Corrected	splicing되지 않은 바이러스 RNA 총량
SARSCoV2_Spliced_TRS_Total_Corrected	splicing된 바이러스 RNA 총량
SARSCoV2_NegativeStrand_Total_Corrected	음성가닥 바이러스 RNA 양 (복제 과정 중 생성됨)
SARSCoV2_PositiveStrand_Total_Corrected	양성가닥 바이러스 RNA 양
SARSCoV2_Total_Corrected	전체 바이러스 RNA 총합 (위 항목들 합산?)

species	종(species) ID (예: `NCBITaxon_9606` = 인간)
species__ontology_label	종 이름 (예: `Homo sapiens`)
sex	환자의 생물학적 성별 (`male` / `female`)
disease	질병 코드 (예: `PATO_0000461` = 정상?)
disease__ontology_label	질병 이름 (`normal`, `COVID-19`, 등)
organ	조직 ID (예: `UBERON_0001728`)
organ__ontology_label	조직 이름 (`nasopharynx` = 비인두)
library_preparation_protocol	실험에서 사용된 library prep 프로토콜의 ID
library_preparation_protocol__ontology_label	프로토콜 이름 (`Seq-Well`, `10x`, 등)
age	환자의 나이대 (`50-59`, `30-39` 등 범주형)
Coarse_Cell_Annotations	세포의 대분류 cell type (예: `Ciliated Cells`, `Goblet Cells`)
Detailed_Cell_Annotations	세포의 세분류 cell type
covid_label	너가 새로 만든 COVID 감염 여부 (PCR 기준): `1` = 감염, `0` = 비감염

COVID 데이터셋 (20210701_NasalSwab_MetaData.txt)

- **총 샘플 수: 32,588개**
- **총 환자 수(donor_id): 58명**
  - **SARSCoV2_PCR_Status**
    - pos 18073
    - neg 14515
  - **disease__ontology_label**
    - COVID-19 18073
    - normal 8874
    - respiratory failure 3335
    - long COVID-19 2306
  - Coarse_Cell_Annotations (18개)

Phenotype prediction from single-cell RNA-seq data using attention-based neural networks

https://academic.oup.com/bioinformatics/article/40/2/btae067/7613064

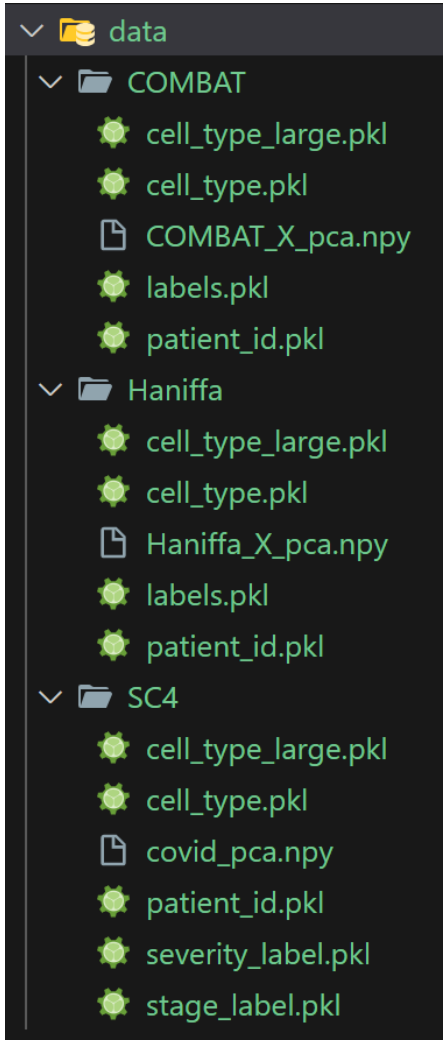
ARTIFICIAL INTELLIGENCE AND DATA ANALYTICS LAB
명지대학교 융합소프트웨어학부 김민

* The code of our proposed method ScRAT is published at https://github.com/yuzhenmao/ScRAT

Experiment

- For **COMBAT and Haniffa datasets**, we perform the task of disease diagnosis
  - COVID versus Non-COVID
- For **SC4** which includes mostly COVID samples
  - mild/moderate versus severe/critical (경증/중등증 vs. 중증/위중증)
  - convalescence versus progression (회복 vs. 진행)

Data File



파일 이름	역할 및 내용
COMBAT_X_pca.npy / Haniffa_X_pca.npy	PCA 차원 축소된 세포별 유전자 발현 데이터 (numpy 배열)
labels.pkl	표현형(Phenotype) 라벨 (COVID/Non-COVID)
patient_id.pkl	샘플에 대한 환자 ID 정보
cell_type_large.pkl	세포 유형(cell type) 정보 (상세한 라벨 포함)
cell_type.pkl	간략한 세포 유형 정보

COMBAT 데이터셋

- **총 세포 수: 836,148개**
- **총 환자 수: 124명**
  - **Unique Labels (8, object)**
    - ['COVID_SEV', 'COVID_MILD', 'COVID_HCW_MILD', 'COVID_CRIT', 'COVID_LDN', 'Sepsis', 'HV', 'Flu']
      - COVID_SEV → 중증 COVID-19 환자 247,799
      - COVID_MILD → 경증 COVID-19 환자 114,418
      - COVID_HCW_MILD → 경증 COVID-19 환자(보건의료 종사자) 88.898
      - COVID_CRIT → 위중증 COVID-19 환자 93.982
      - COVID_LDN → 런던 COVID-19 환자군 15,485
      - Sepsis → 패혈증(Sepsis) 환자 164,128
      - HV → 건강한 대조군(Healthy Volunteers) 92.205
      - Flu → 인플루엔자(Flu) 환자 19,233
  - **Unique Cell Types Large (41, object)**
    - ['NK.CD16hi', 'CD8.TEMRA', 'nan', 'ncMono', 'cMono', ..., 'GDT.VD2', 'CD8.TREG', 'PLT', 'RET', 'Mast']
  - **Unique Cell Types (18, object)**
    - ['NK', 'CD8', 'nan', 'ncMono', 'cMono', ..., 'HSC', 'DC', 'PLT', 'RET', 'Mast']

Haniffa 데이터셋

- **총 세포 수: 647,366개**
- **총 환자 수: 130명**
  - **Unique Labels (10, object)**
    - ['Moderate', 'Healthy', 'Death', 'Mild', 'Severe', 'LPS', 'Critical ', 'Non-covid', 'Asymptomatic', 'nan']
      - Moderate → 중등증 COVID-19 환자 179,012
      - Mild → 경증 COVID-19 환자 93,835
      - Severe → 중증 COVID-19 환자 40,235
      - Critical → 위중증 COVID-19 환자 63,854
      - Asymptomatic → 무증상 COVID-19 환자 33,601
      - Non-covid → COVID-19 감염되지 않은 환자 15,157
      - Healthy → 건강한 대조군(Healthy Control) 97,039
      - LPS → LPS(lipopolysaccharide) 염증 반응 실험군 7,884
      - Death → 사망자 데이터 41,836
      - nan → 결측값(missing value) 포함 74,913
  - **Unique Cell Types Large (51, object)**
    - ['CD8.TE', 'CD4.IL22', 'CD8.Naive', 'CD4.Naive', 'CD8.EM', ..., 'HSC_CD38neg', 'HSC_myeloid', 'HSC_MK', 'CD4.Th17', 'B_malignant']
  - **Unique Cell Types (18, object)**
    - ['CD8', 'CD4', 'CD14', 'B_cell', 'NK_16hi', ..., 'gdT', 'HSC', 'pDC', 'RBC', 'Mono_prolif']