GIST, 단일세포 분석의 새로운 패러다임 제시
AI 기술로 단일세포 RNA 시퀀싱 기술 한계 극복

- AI대학원 이현주 교수팀, 단일세포 RNA 데이터에서 모든 유전자를 학습시킬 수 있는 자기 지도 학습 방법론 통해 세포 종류에 대한 예측 성능 크게 향상된 'scRobust' 개발
- "소수의 세포에서만 발현되는 유전자와 같은 세포 유형에서 발생하는 미세한 특징도 비교·분석할 수 있게 돼" 국제학술지 《Briefings in Bioinformatics》 게재

▲ (왼쪽부터) AI대학원 이현주 교수, 전기전자컴퓨터공학부 박사과정 박세진 학생

개별 세포 단위로 유전자 발현량을 측정할 수 있는 단일세포 리보핵산(RNA)* 시퀀싱(sequencing)*은 최근 생물학, 신약 개발, 임상 연구 등 다양한 분야에서 급속한 발전을 이뤄 주목받고 있다.

단일세포 RNA 시퀀싱 기술은 세포 종류에 따른 유전자 분석이 가능해 질병에 대한 새로운 진단과 예후 예측 등 맞춤형 의료 서비스 제공에 적합하나, 여러 세포의 발현량을 합산하여 측정하는 **다세포 RNA 시퀀싱* 기술에 비해 정확도가 낮아 전체 유전자 중 일부만 검출되는 한계**가 있었다.

* **RNA:** DNA를 복사한 물질로 세포의 유전 정보를 담고 있다.

* **RNA 시퀀싱(**Single-cell RNA sequencing**):** 세포 안의 RNA의 양을 측정하는 기술로서 이를 통해 유전자의 발현(활성화) 정도를 추측할 수 있다. (예, 당뇨 환자의 경우 인슐린 분비와 관련된 유전자들은 발현이 감소함)

광주과학기술원(GIST, 총장 임기철)은 AI대학원 이현주 교수 연구팀이 단일세포 RNA 시퀀싱 기술의 근본적인 한계를 극복할 수 있는 **자기 지도 학습(self-supervised learning)* 방법론을 개발**했다고 밝혔다.

이를 적용한 결과, 동일한 세포 종류라도 당뇨병의 정도에 따라 구분되는 세부적인 특징까지도 발견할 수 있었다. 또한, 15개의 단일세포 RNA 데이터세트에서 실시한 세포 종류 분류 테스트 중 12개 데이터세트에서 가장 높은 F1 점수*를 보였다.

* **자기 지도 학습(self-supervised learning):** 일반적인 AI 모델을 학습시키기 위해서는 데이터와 라벨(그 데이터를 설명할 수 있는 정보)이 필요하고, 이러한 학습 방식을 지도 학습이라고 한다. 반면, 자기 지도 학습은 별도의 라벨 없이 AI 모델을 학습시키는 방법론을 말한다. 예를 들어, 챗봇 모델을 학습시키기 위해서는 대화 텍스트만 있으면 되며 별도의 라벨(해당 대화가 긍정적인지 혹은 부정적인지)은 필요 없다.

* **F1 점수:** 정밀도(얼마나 정확하게 맞췄는지, Precision)와 재현율(얼마나 놓치지 않았는지, Recall)의 균형을 평가하는 점수다. 해당 점수는 단순히 정답을 몇 개 맞췄는지에 초점을 맞추는 것이 아니고, 얼마나 정밀하게 정답을 맞췄는지를 측정하는 점수다. 예를 들어, 99명의 정상인과 1명의 암환자를 구분하는 테스트를 했을 때, 항상 정상인이라고 판단하면 99%의 정답률을 갖지만 F1 점수는 0이 된다. 왜냐하면 정상인은 정상인으로 판단했지만, 암환자를 암환자라고 진단하지 못했기 때문이다.

여러 세포의 RNA가 섞인 유전자 발현량을 측정할 수 있는 다세포 RNA 시퀀싱 기술에 반해 단일세포 RNA 시퀀싱은 단일세포만을 대상으로 하기 때문에 측정 정확도가 떨어진다. 따라서 **단일세포 RNA 데이터는 보통 30,000개가 넘는 유전자 중에서 2,000~3,000개 유전자의 발현량만을 얻는 경우가 많다.**

즉, 전체 유전자 중 10%만이 측정 가능한 높은 해상도를 가지고, 나머지 90% 정보는 낮은 해상도로 인하여 측정이 불가하다.

이에 따라 기존의 연구는 주로 여러 세포에서 공통으로 발현되는 **약 10%의 유전자만을 사용하여 세포 유형을 예측하고 분석**해 왔다.

하지만 특정 세포 종류에서만 발현되는 유전자가 오히려 해당 세포를 더 자세히 설명하는 경우가 많고, 현재 사용되는 **단일세포 RNA 시퀀싱 기술은** 약 90%의 유전자 정보를 사용하지 못하는 근본적인 문제점이 있었다.

연구팀은 단일세포 RNA 시퀀싱 데이터에 적합한 대조 학습(contrastive learning, 자기 지도 학습 방법론 중 하나)* 방법론을 활용하여 단 5% 미만의 유전자 정보만을 가지고도 각 단일세포의 보편적인 특징부터 세부적인 특징까지도 파악할 수 있는 기술, **'scRobust'를 개발**하였다.

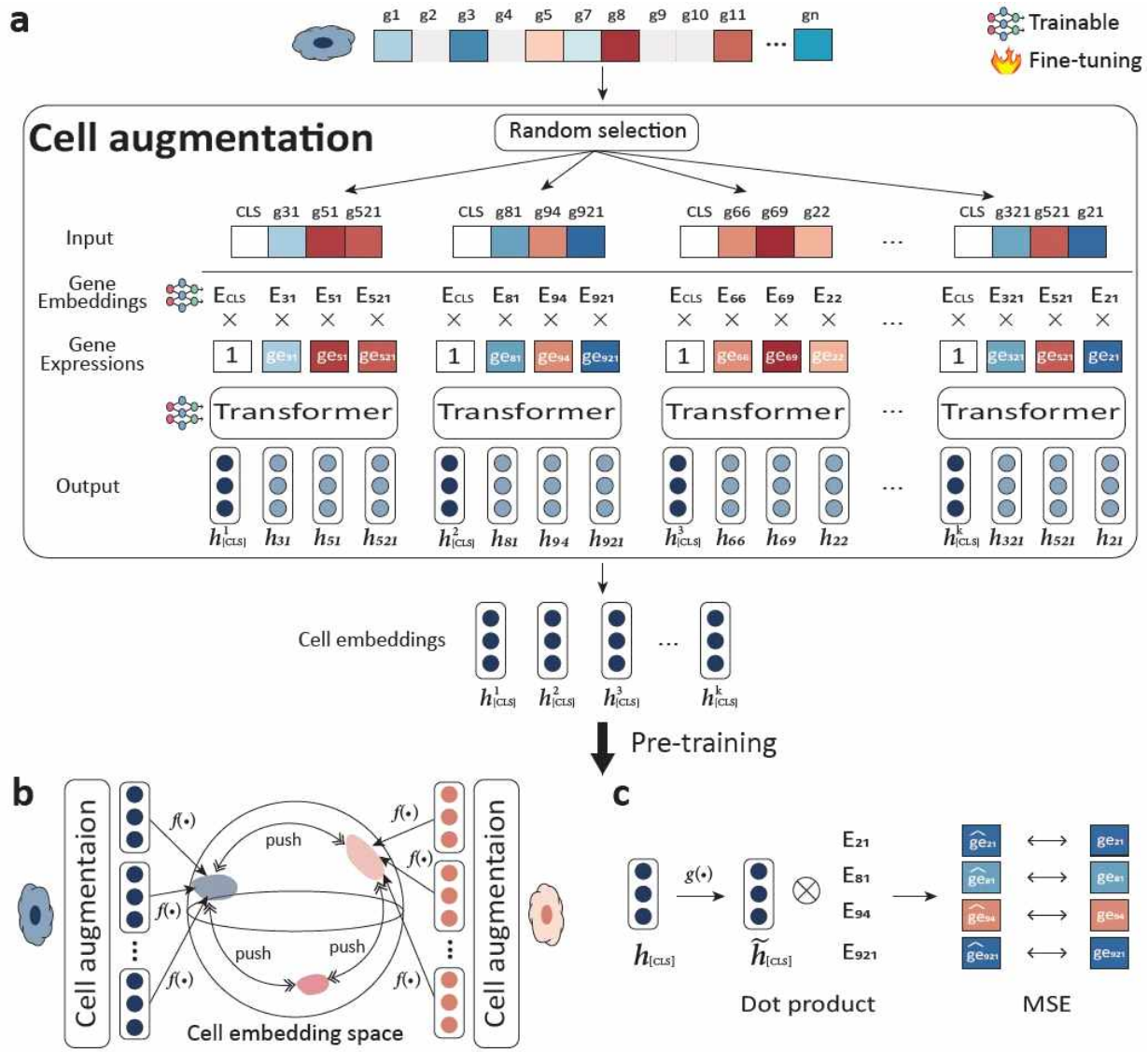
이를 통해 이전 방법론에서 활용되지 않았던 90%의 유전자 정보까지 사용할 수 있게 되어 **세포 종류에 대한 예측 성능이 향상되었을 뿐만 아니라 동일한 세포 내에서도 더욱 정밀한 분석이 가능**해졌다.

* **대조 학습 (contrastive learning):** 대조 학습은 주어진 데이터에 다양한 변화를 주어 여러 개의 데이터로 만들고, AI 모델이 이렇게 만들어진 수많은 데이터 중에서 같은 원본에서 나온 데이터들을 찾을 수 있도록 훈련시키는 방식이다.

이 기술은 하나의 세포로부터 다양한 유전자 조합을 만들어 여러 개의 세포 표현 벡터(cell representation vector)*를 생성할 수 있는 방법론을 기반으로 단일세포 RNA 시퀀싱 데이터에 적합한 데이터 증강(data augmentation)*을 하는 것이다.

* **표현 벡터(representation vector):** 임의의 데이터 샘플을 벡터 형태로 변환한 형태를 뜻한다. 예를 들어, cell representation vector는 세포를 벡터 형태로 변환하여 AI 모델이 세포라는 개념을 이해할 수 있게 된다.

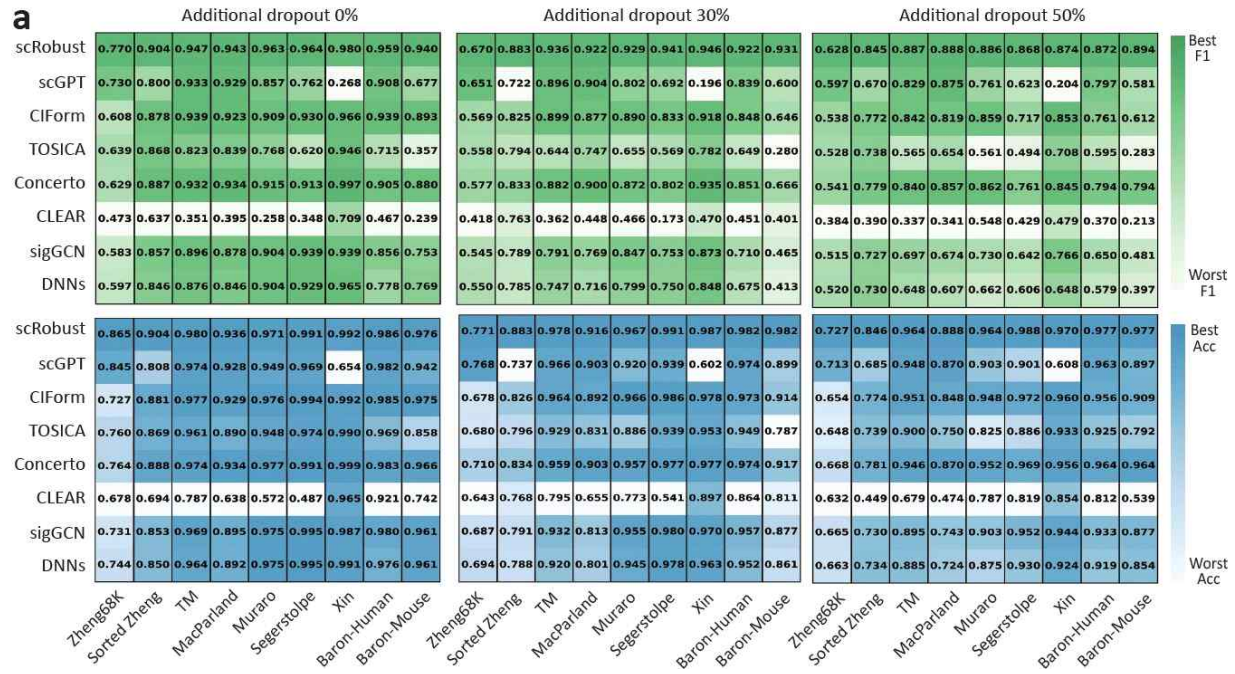
* **데이터 증가(data augmentation):** 주어진 데이터에 여러 가지 변환을 주어서(예: 하나의 사진을 회전시키거나 흑백 사진으로 바꿈) 여러 개의 데이터로 만드는 기술이다. 이 기술을 이용하면, AI 모델을 학습 시킬 수 있는 데이터의 수가 증가하는 효과가 생긴다.



▲ **본 연구에서 개발한 self-supervised learning의 개략도.** (a) data (cell) augmentation, 임의의 세포에서 다양한 세포 표현 벡터를 만들어낸다. (b) 대조 학습, 같은 세포에서 만들어진 세포 표현 벡터(cell embeddings)는 서로 유사한 위치에 존재한다. (c) 세포 표현 벡터를 이용하여, 임의의 유전자의 발현량을 예측한다.

대조 학습을 통해 AI 모델을 학습시키면, 서로 다른 유전자 조합으로 생성된 세포 표현 벡터라도 같은 세포에서 나온 것인지, 다른 세포에서 나온 것인지 구분할 수 있으며, 이 과정을 통해 다양한 유전자 조합으로 만든 세포 표현 벡터(local cell representation vector)들이 하나의 통일된 세포 표현 벡터(global cell representation vector)로 수렴하게 된다.

결과적으로 소수의 유전자만 사용하더라도 모든 유전자를 활용한 것과 유사한 세포 표현 벡터를 얻을 수 있어 **전체 유전자를 사용하는 효과를 기대**할 수 있다.



▲ **세포 유형 예측 결과**. 본 연구에서 개발한 모델 (scRobust)은 서로 다른 9개의 데이터 셋 중 8개에서 가장 높은 성능(F1 점수)을 보인다. 이는 scRobust이 특정 데이터 셋에서만 작동하는 것이 아니라 대부분에 데이터 셋에서 좋은 성능을 보임을 보여준다.

이현주 교수는 "이번 연구에서 개발된 알고리즘은 AI 모델이 유전자 일부만 학습하는 것이 아니라 **모든 유전자에 대해 학습하는 것이 가능**하다"면서 "이를 통해 **그동안 소수의 세포에서만 발현되는 유전자와 같은 세포 유형에서 발생하는 미세한 특징들까지도 비교, 분석**할 수 있게 되었다"고 설명했다.

또한 "다양한 세포 종류의 마커 유전자뿐만 아니라 **약물 저항성과 관련된 마커 유전자*까지 추출할 수 있어 향후 단일세포 분석의 패러다임을 바꿀 수 있을 것**으로 기대된다"고 말했다.

* **마커 유전자(marker gene):** 특정 세포나 조직에서만 활발하게 발현되어 해당 세포나 조직을 구분하는 데 도움이 되는 유전자를 말한다.

GIST AI 대학원 이현주 교수가 지도하고 박세진 박사과정생이 수행한 이번 연구는 정보통신기획평가원(IITP)의 지원을 받았으며, 생물정보학 분야 JCR 상위 4% 국제학술지 《Briefings in Bioinformatics》에 2024년 11월 16일 게재됐다.

논문의 주요 정보

1. 논문명, 저자정보

- 저널명 : Briefings in bioinformatics (IF: 6.8, 2023년 기준)
- 논문명 : Robust self-supervised learning strategy to tackle the inherent sparsity in single-cell RNA-seq data

- 저자 정보 : 박세진 (제1저자, 전기전자컴퓨터공학부 박사과정), 이현주 교수 (교신저자, AI 대학원, 전기전자컴퓨터공학부)