Title	Dimensionality reduction for visualizing single-cell data using UMAP

Paper Summary

본 연구는 초고차원인 single cell RNA sequencing 데이터셋과 세포 질량 분석 데이터셋에서 차원 축소를 수행하기 위하여 UMAP을 포함한 5개의 서로 다른 알고리즘을 적용하여 성능 차이를 비교한 연구이다. 기존에 존재하는 연구에서는 single cell RNA-seq 데이터에 IsoMap, Diffusion Map, t-SNE 등의 비선형 차원 축소 알고리즘을 적용했다. 이 비선형 차원축소 알고리즘들은 데이터의 국소 부위 구조와 전사체, 세포 데이터에서의 특이세포 분포 등을 잘 표현하였으나, 차원 축소를 하면서 발생하는 군집 간의 관계 손실, 느린 속도, 더 큰 데이터셋에서는 의미 있는 분석을 할 수 없음을 보여줬다. 본 연구에서는 local structure와 global structure 모두 최대한 보존하면서 수행 시간을 줄이기 위해 UMAP 알고리즘을 선택하였으며, t-SNE, Barnes-Hut t-SNE, Fit-SNE, scvis 알고리즘들과 비교하여 UMAP의 성능이 얼마나 좋은가를 나타내기 위해 시각화 및 정량화 자료를 제공했다. Fit-SNE는 기존 t-SNE에서 컨볼루션 단계 속도를 높여 실행 시간을 단축할 수 있도록 최적화된 알고리즘이다. 본 연구에서는 우선 가장 잘 알려진 UMAP, t-SNE 를 비교했고 이를 포함한 위 5개의 알고리즘을 통해 어느 것이 embedding 공간에서 local structure 혹은 global structure가 잘 보존되었는가를 파악하기 위한 분석을 진행하였다. 그 결과, 1. 2차원 임베딩 공간으로 차원을 축소시킨 뒤 랜덤 포레스트로 학습 결과를 평가해보았을 때, scvis를 제외한 알고리즘들 모두 90%가 넘는 정확도를 보였으며, 세포들의 군집 또한 깔끔하게 표현됐다. 2. Local structure 보존 정도를 비교한 분석에서는 UMAP 과 Fit-SNE가 보존 정도를 잘 표현했다. 3. Global structure 보존 정도에 대한 비교에서는 UMAP과 scivs가 구조를 가장 잘 표현했다. 즉, 종합적으로 판단했을 때, UMAP이 데이터의 구조를 가장 잘 보존하면서 차원축소를 수행했고,시각화, 정량화 자료를 통해 증명했다. 대상 데이터의 크기에 상관없이 다른 4가지 알고리즘보다 수행 속도 측면에서 비교 우위를 가지고 있기 때문에, single cell RNA sequencing과 같이 초고차원이면서 클러스터링이 필요한 데이터셋에 적합한 알고리즘임을 보였다.

Criticism for Research

본 연구 논문의 시각화 자료만 봐도 각 알고리즘마다 어느 정도로 클러스터링이 되었는지, 수행 시간이 얼마나 차이가 나는지, 초고차원에서의 구조를 얼마나 잘 유지하였는지 파악할 수 있다. 하지만, 적합한 차원 축소 알고리즘을 정하는 기준에 대한 설명이 제대로 이루어지지 않았다. 본 논문의 제목에서 확인할 수 있듯이, UMAP이 Single cell RNA sequencing 데이터의 차원 축소 알고리즘으로서 적합성을 증명하고자 하는게 이 연구의 목적이기 때문에, UMAP의 파라미터, 어떤 원리에 따라 local, global structure가 잘 보존될 수 있었는지에 대한 내용이 추가될 필요가 있다. 그리고 클러스터링이 잘 되었는가를 판단하기 위해 실루엣 스코어를 사용한 것인지 다른 방안으로 판단한 것인지에 대한 설명이 명시 될 필요가 있다.

Idea sketch

본 연구에서 UMAP은 single cell 데이터셋에 대해서 수행 속도가 빠르면서 local, global structure를 모두 잘 보존하는 알고리즘으로 증명되었다. UMAP이 어떠한 원리로 다른 차원 축소 알고리즘과 차이를 가지는지 알아봤다. UMAP은 min_dist라는 parameter로 각 데이터 간 최소 거리를 정의해야 한다. 최소 거리는 데이터를 중심으로 하는 임의의 원의 크기를 결정하기 때문에 값이 증가할수록 데이터를 중심으로 하는 원의 범위가 바깥쪽으로 확장되게 된다. 이 때, 다른 데이터와의 교차점이 생기면 두 데이터를 이어주고 같은 군집으로 고려한다. 여기에서 min_dist 값이 너무 작으면 많은 데이터들이 혼자 군집을 이루게 되고, 너무 크면 모든 데이터들이 하나의 군집을 이루게 되는 문제가 생긴다. 이러한 문제를 조율해주는 parameter가 UMAP의 n_neighbors이다. 이 값으로 UMAP이 local structure를 잘 보존할 것인지 global structure를 잘 보존할 것인지 결정하게 된다. 즉, 본 연구에서는 UMAP이 가장 적합함을 단순히 시각화만으로 설명하였지만, 위 두 파라미터를 제대로 설정하지 않는다면 t-sne보다도 성능이 훨씬 떨어질 수 있음을 알아야 한다.

UMAP vs t-SNE

UMAP	t-SNE
embed되는 차원에 제한이 없다.	2 혹은 3차원으로만 embed 하기 때문에 일반적으로 시각화에 많이 사용한다. 그래서 초고차원에 직접적으로 사용하지 않고 PCA, Autoencoder로 차원 축소시킨 데이터에 적용하는 것이 일반적이다.
n_neighbor, min_dist 두 개의 파라미터로 local, global structure 보존 정도를 조절할 수 있다.	기본적으로 local structure는 잘 보존해주나, global structure 보존에 신경을 써야 한다면 perplexity 라는 파라미터 값을 높여주어야 하는데 이 때 수행 시간이 길어지며 너무 많은 메모리를 차지하게 된다.
Two nested cluster 분리가 불가능하다.	Nested 된 cluster도 분리할 수 있다.