학부연구생_개인

2021학년도 1학기 CURT 프로그램 연구보고서

성 명	이진백	학년/학번	3학년
학과(전공)	컴퓨터융합소프트웨어학과	단과대학	과학기술대학
지도교수		멘 토	
연구기간	~ 2021.07.09.		
연구주제	장내 미생물군과 알츠하이머병의 연관성을 파악하기 위한 차원 축소 알고리즘 연구		

2021년 07월 12 일

지도교수: _____(서명)

멘토(대학원생) : _____(서명)

학부연구생 : _____이진백_____(서명)

Ⅰ. 연구 목적

사람 몸의 물질대사 균형을 유지해주고 건강 상태를 확인할 수 있는 장내 미생물이 알레르기 질환, 비만, 혈관질환에 이어, 심지어 뇌 질환에도 영향을 줄 수 있다는 사실이 밝혀지면서 장내 미생물에 대한 연구가 기하급수적으로 늘어나고 있다. 최근, 알츠하이머병(Alzheimer's disease, AD)을 유발하는 물질이 특정 장내 미생물의 증가와 감소에 영향을 받는다는 연구가 진행되었다. 해당 연구는 쥐를 AD 쥐와 건강한 쥐로 구분하였고, 두 집단 간에 장내 미생물 구성과 세포 배양을 통한 면역 체계의 차이를 비교하였다. 이 차이를 바탕으로 건강한 쥐의 배설 미생물종을 AD 쥐로 이동시키는 연구가 진행되었고 AD 쥐의 학습 행동이 개선된 것을 발견하였다. 이번 CURT 프로그램 연구 목적은 장내 미생물과 알츠하이머병의 연관성을 파악하기 위해 적합한 차원 축소 알고리즘을 연구하고 한계를 제시한다.

Ⅱ. 연구 내용

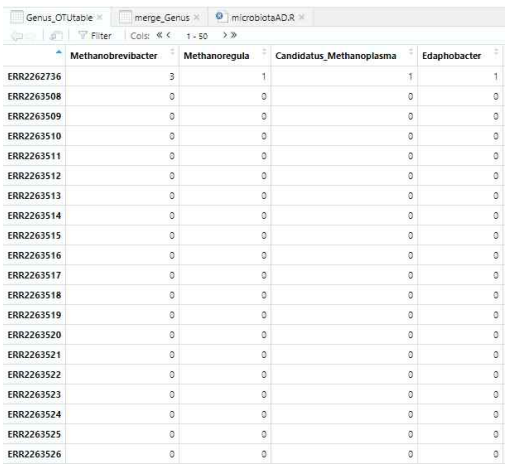
1. Data collection

https://www.ebi.ac.uk/metagenomics/studies/MGYS00002182#overview (Mgnify)

경희대학교에서 2018년 1월 26일에 업데이트된 쥐의 Fecal 데이터를 사용하였다.

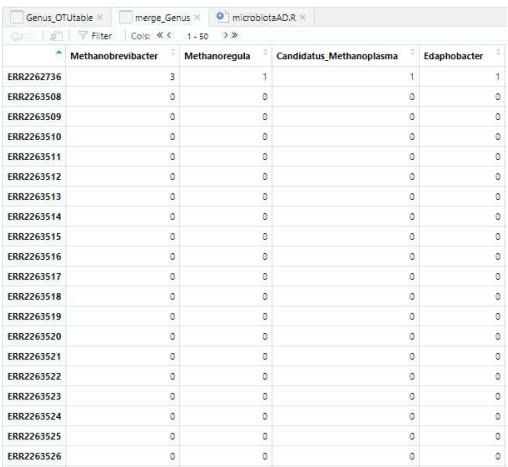
2. Data cleansing

앞서 수집한 데이터에서 각 각의 독립변수가 무엇을 의미하는지 알아야 한다. 장내 미생물(microbiome)은 taxonomy 라 불리는 계층적 생물 분류 체계를 따른다. 종 < 속 < 과 < 목 < 강 < 문 < 계 순으로 구성되며 각 level이 하위 level 들을 포함하는 개념이다. 이와 같은 계층적 구조로 표현된 장내 미생물 데이터를 OTU table이라 부르며 1번에서 수집된 데이터 역시 OTU table 형태로 클렌징 되어야 한다. 이때, '속'은 두 번째 작은 단위로, 미생물 군집의 다양성을 파악하면서 일정 수준의 군집화가 가능하다고 판단하여, 속(Genus) OTU table을 직접 구현하였다. 직접 구현한 OTU table의 정확성 판단을 위해 R에서 제공하고 있는 Phyloseq package를 통한 결과와 비교하여 일치함을 확인하였다.



	Methanobrevibacter	Methanoregula	Candidatus_Methanoplasma	Edaphobacter
ERR2262736	3	1	1	1
ERR2263508	0	0	0	0
ERR2263509	0	0	0	0
ERR2263510	0	0	0	0
ERR2263511	0	0	0	0
ERR2263512	0	0	0	0
ERR2263513	0	0	0	0
ERR2263514	0	0	0	0
ERR2263515	0	0	0	0
ERR2263516	0	0	0	0
ERR2263517	0	0	0	0
ERR2263518	0	0	0	0
ERR2263519	0	0	0	0
ERR2263520	0	0	0	0
ERR2263521	0	0	0	0
ERR2263522	0	0	0	0
ERR2263523	0	0	0	0
ERR2263524	0	0	0	0
ERR2263525	0	0	0	0
ERR2263526	0	0	0	0

[표1] 직접 구현된 Genus OTU table



	Methanobrevibacter	Methanoregula	Candidatus_Methanoplasma	Edaphobacter
ERR2262736	3	1	1	1
ERR2263508	0	0	0	0
ERR2263509	0	0	0	0
ERR2263510	0	0	0	0
ERR2263511	0	0	0	0
ERR2263512	0	0	0	0
ERR2263513	0	0	0	0
ERR2263514	0	0	0	0
ERR2263515	0	0	0	0
ERR2263516	0	0	0	0
ERR2263517	0	0	0	0
ERR2263518	0	0	0	0
ERR2263519	0	0	0	0
ERR2263520	0	0	0	0
ERR2263521	0	0	0	0
ERR2263522	0	0	0	0
ERR2263523	0	0	0	0
ERR2263524	0	0	0	0
ERR2263525	0	0	0	0
ERR2263526	0	0	0	0

[표2] Phyloseq package로 구현한 Genus OTU table

위 표는 각 OTU table의 일부를 나타낸다.

앞서 데이터 수집 단계에서 수집한 metadata와 Genus OTU table을 합쳐서 하나의 데이터로 만들어야 한다. metadata란 특정 데이터를 설명해주는 또 다른 데이터를 의미한다. 현재 OTU table은 행에 쥐 샘플, 열에 Genus type 장내 미생물로 구성되어 있으며, 각 쥐 샘플의 metadata에서 AD 여부를 나타내주는 변수를 합쳐서 최종 Genus_OTUtable 를 구성하였다.

Genus_OTUtable | microbiotaAD.R | merge_Genus

Filter | Cols: « < 1 - 50 > »

	SampleID	AD	Methanobrevibacter	Methanoregula	Candidatus_Methanoplasma	Edaphobacter
1	ERR2262736	1	3	1	1	1
2	ERR2263508	1	0	0	0	0
3	ERR2263509	1	0	0	0	0
4	ERR2263510	1	0	0	0	0
5	ERR2263511	1	0	0	0	0
6	ERR2263512	1	0	0	0	0
7	ERR2263513	1	0	0	0	0
8	ERR2263514	1	0	0	0	0
9	ERR2263515	1	0	0	0	0
10	ERR2263516	1	0	0	0	0
11	ERR2263517	1	0	0	0	0
12	ERR2263518	1	0	0	0	0
13	ERR2263519	1	0	0	0	0
14	ERR2263520	1	0	0	0	0
15	ERR2263521	1	0	0	0	0
16	ERR2263522	1	0	0	0	0
17	ERR2263523	1	0	0	0	0
18	ERR2263524	1	0	0	0	0
19	ERR2263525	1	0	0	0	0
20	ERR2263526	1	0	0	0	0

[표3] 최종 Genus OTU table 일부

3. Dimensionality reduction theory

최종 Genus OTU table은 (33,472) 형태를 띤다. 데이터의 샘플 수가 매우 적고 미생물 종류(변수)가 압도적으로 많다. 일반적으로 샘플 수가 적다면 그만큼 신뢰성 있는 정보를 추출하기 어렵고, 고차원일수록 “차원의 저주” 이론에 따라 데이터들의 특성 파악이 매우 어려워진다. 현재 OTU table에서 샘플 수를 늘리는 것은 불가능하기 때문에, 데이터의 특성을 가장 잘 표현 할 수 있는 합리적인 차원 축소 알고리즘을 적용해야 한다. 이를 위해 4가지 차원 축소 방법을 적용해 본다. 먼저 데이터에 PCA를 적용한 뒤 얻은 결과물에 t-sne, UMAP을 각 각 적용하고, PCA를 적용하지 않은 기존 데이터에 t-sne, UMAP을 적용하여 비교한다.

PCA 란?

PCA(Principal component analysis)는 가장 기본적인 차원 축소 알고리즘이다. 데이터의 분산을 최대한 보존하면서 기존 변수들의 상관관계에 기반하여 새롭게 추출된 변수(주성분)에 데이터들을 정사영 시키는 방법이다. PCA는 SVD(Singular Value Decomposition)이라 불리는 특이값 분해에 기반하여 수행될 수도 있고, 공분산 행렬에 기반하여 고유벡터와 고유값으로 수행될 수도 있다. 본 연구에서는 R prcomp를 사용하였고 이는 SVD 방식에 해당한다.

이 함수를 통해 생성된 주성분 PC는, 기존 변수들이 각 각 얼만큼 기여하였는가를 확인할 수 있

다.

3-1. Original data + t-sne

t-SNE(t distribution-stochastic neighbor embedding)는 데이터 간의 유사성을 이용하여 embedding 하는 차원 축소 알고리즘이다. 쉽게 말해, 원래 가까웠던 데이터는 더 가깝게, 멀었던 데이터는 더 멀어지게 만든다. 일반적으로 2 또는 3차원으로만 표현이 가능하기 때문에 시각화에 유용하다. 데이터에 여러 클러스터가 있는 경우 클러스터링이 제대로 되지 않을 수도 있는 PCA의 단점을 보완 한 것이다. 하지만 차원이 축소되는 과정이 기존 변수들과 무관하게 임의로 만들어진 2차원 공간에 초고차원 데이터를 뿌리는 것이기 때문에 아무리 PCA에 비해 클러스터링이 잘된 시각화 결과를 보여도, 데이터에 대한 해석은 각 클러스터 사이 거리에만 의존하여야 된다. 즉, embedding space의 각 변수가 의미하는 바는 전혀 파악할 수 없다.

3-2. Original data + UMAP

UMAP(Uniform Manifold Approximation and Projection)은 각 데이터를 중심으로 가중치와 min_dist 거리 값에 기반한 가상의 원의 범위를 갖는다. 해당 범위가 다른 데이터와 겹치면 두 데이터를 연결시키고 같은 클러스터로 판단하는 차원 축소 알고리즘이다. 데이터들의 전체 구조를 잘 보존하기 힘든 t-sne의 단점을 보완하였다. 즉, 기존 데이터의 각 군집 간의 거리와 군집 내 데이터 간의 거리를 모두 잘 보존하면서 차원 축소를 수행하게 된다. 하지만 UMAP 역시 embedding space의 각 변수가 의미하는 바를 전혀 파악할 수 없다.

3-3. PCA + t-sne

기존 데이터에 PCA를 적용한 뒤 t-sne를 적용한다. PCA와 t-sne 각 알고리즘의 장점을 모두 적용할 수 있다. PCA로 데이터의 특성을 파악할 수 있는 변수들로 축소시킨 뒤 t-sne로 군집 내 데이터들의 구조에 대한 보존력을 높이면서 2차원으로 재 축소 시킨다. 그냥 PCA를 적용하거나 t-sne를 적용했을 때보다 효과적 일 수 있다. 하지만 PCA는 정규화를 필요로 한다는 점을 유의해야 한다.

3-4. PCA + UMAP

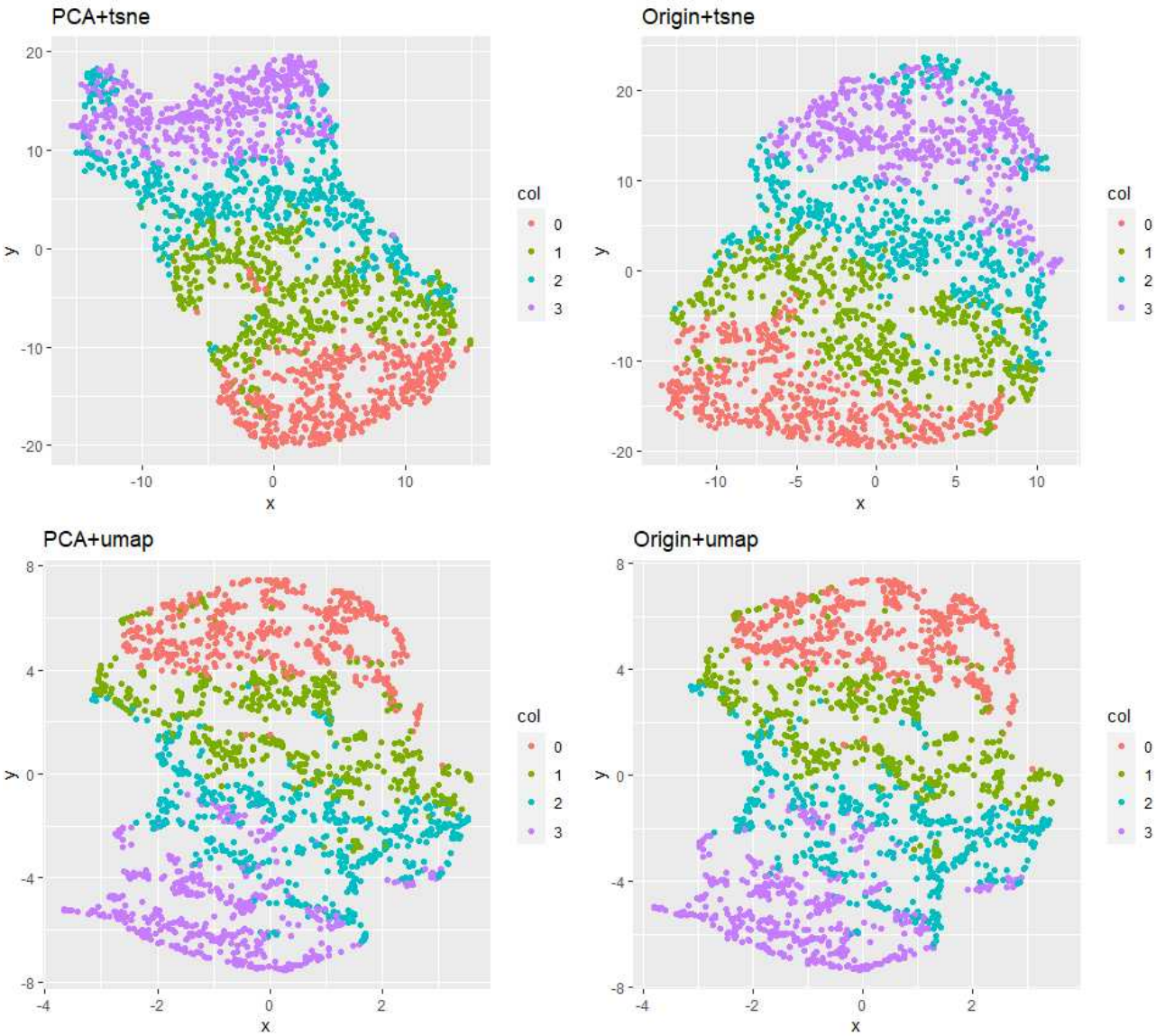
기존에 데이터에 PCA를 적용한 뒤 UMAP을 적용한다. 위 PCA+t-sne 와 동일한 논리이며 기존 데이터의 구조가 얼마나 보존되는가에만 차이가 있다.

4. Testing on temporary data

위 4가지 차원 축소 방법을 Genus OTU table에 적용해보기 전에 임의의 다른 데이터에 적용하여 실제로 효과가 있는지 파악해야 한다. Genus OTU table은 샘플 수가 매우 적기 때문에 4가지 차원 축소 방법이 실질적으로 차이를 갖게 될 것인지 알 수 없다.

데이터의 출처는 Ⅳ. 참고 문헌에 명시되어 있다. 데이터 선정 기준은 샘플 수가 충분한가, 차원 축소를 두 번 적용할 만큼의 차원으로 구성되어 있는가, 데이터가 명확하게 분류되어 있는가에 따라 결정하였다.

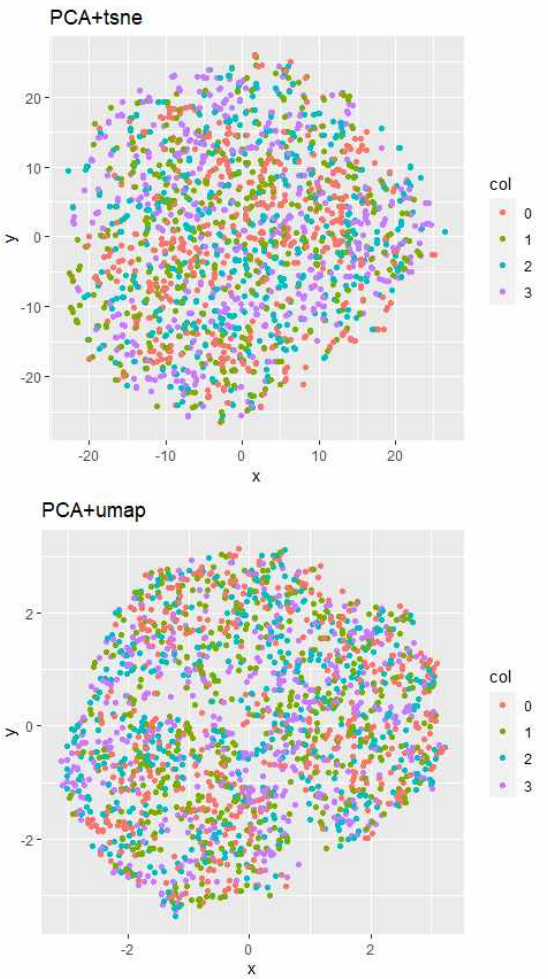
결정된 데이터를 클렌징 한 뒤에 4가지 방법을 적용하여 각 각 시각화 하였다.
이때, 적용된 PCA는 정규화를 시키지 않은 상태로 t-sne와 UMAP을 적용하였다. Original data 에서 바로 t-sne와 UMAP을 적용한 결과와 크게 다르지 않다. 하지만 이 경우에는 큰 오류가 생기게 된다. PCA 과정에서 정규화가 되지 않았기 때문에 분산이 큰 변수가 주성분으로 선택되지 않고 단순히 큰 값을 가지는 변수가 주성분으로 선택되게 된다.



[그림 1] PCA 과정에서 정규화를 적용하지 않은 4가지 알고리즘 plot

다시 말해, PCA 과정에서 정규화를 해주지 않으면 의미 있는 변수를 얻을 수 없다. 그래서 R

prcomp 함수로 PCA를 진행하면서 동시에 scaling(=정규화)을 해주었다.
이 점을 고려하여 정규화를 진행한 뒤 t-sne와 UMAP을 적용한 plot은 다음과 같다.



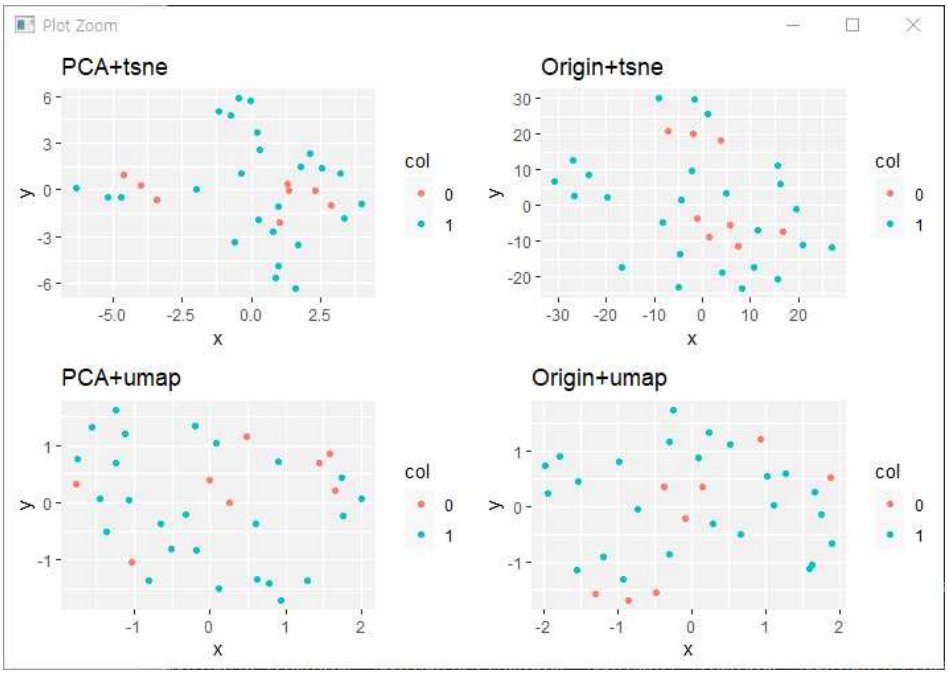
옆 그림에 보이는 결과와 같이 정규화를 진행한 뒤 t-sne, UMAP을 적용하게 되면 모든 값이 유사한 수준의 범위 값을 가지며 표준 편차가 일정하게 맞춰지기 때문에 거리 기반으로 동작하는 두 알고리즘에서 클러스터링이 제대로 되지 않게 된다. 즉, PCA에서 아무리 기존 변수들로 유의미한 새로운 PC들을 추출 해내었다고 해도 각 군집을 확실하게 분류시켜주지 못한다.

[그림 2] PCA 과정에서 정규화를 적용한 뒤 t-sne와 UMAP을 적용한 plot

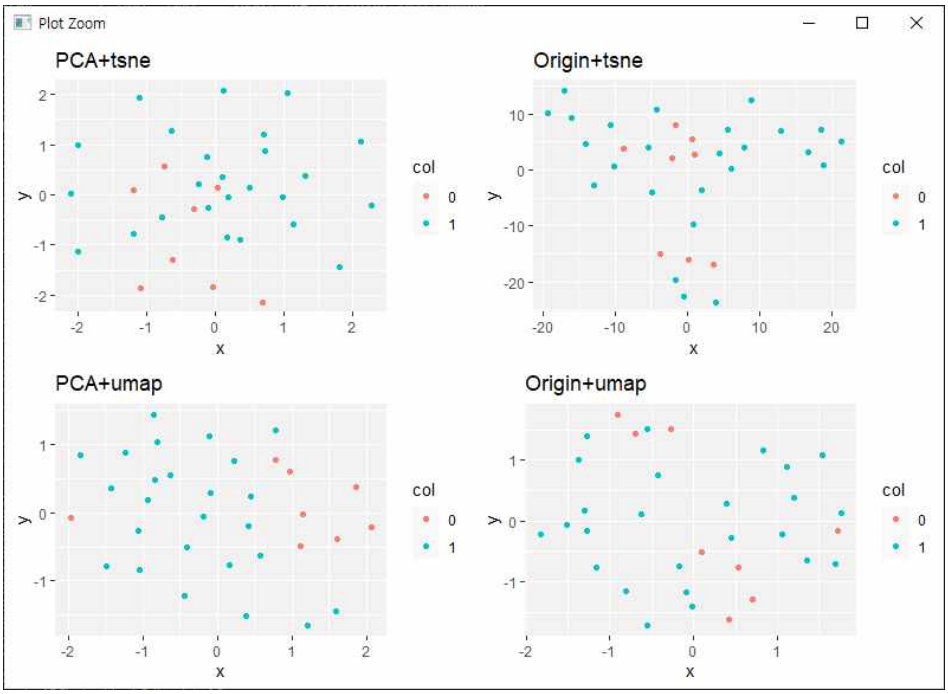
Original data에 t-sne와 UMAP을 적용한 그래프처럼 데이터 간의 클러스터링을 올바르게 해주기 위해서는 정규화의 반대과정을 거쳐주어야 제대로 클러스터링이 될 것이다. 이 경우 PCA, 즉 새로운 에서 정규화가 되었기 때문에

5. Testing on Genus OTU table data

이제 위와 같은 접근 방식으로 Genus OTU table 데이터에 3-1 ~ 3-4 차원 축소 알고리즘을 적용하였다.



[그림 3] PCA 정규화가 되지 않은 plot



[그림 4] PCA 정규화가 된 plot

Genus OTU table 데이터는 4단계에서 사용한 temporary data에 반해 차원의 개수가 훨씬 많고 샘플 수는 훨씬 적다. 그림 3과 그림 4를 통해 볼 수 있듯이 Original data에 바로, 혹은 PCA를 적용한 뒤에 t-sne와 UMAP을 적용한 방법 모두 클러스터를 정확하게 확인할 수 없었다. 즉, 샘플 수가 충분하지 않다면 어떠한 경우에도 데이터의 특성을 정확하게 파악할 수 없다.

Ⅲ. 연구 결과

본 연구에서는 PCA+t-sne, PCA+UMAP, origin+t-sne, origin+UMAP 총 4가지 차원 축소 알고리즘을 사용하였다. [그림 1]의 일부와 [그림 2]는 임의 데이터에 4가지 알고리즘을 적용한 뒤 시각화 한 것이고, [그림 3]과 [그림 4]는 OTU table에 4가지 알고리즘을 적용한 뒤 시각화한 것이다. 그 결과 PCA가 얼마나 유의미한 변수들로 추출을 하였는가, t-sne와 UMAP 알고리즘이 얼마나 데이터의 특성을 유지하면서 군집화를 하였는가와 관계없이 샘플 수가 적다면 데이터 특성에 대한 예측이 불가능하다 라는 결과를 얻었다.

하지만 Temporary data에서 PCA+t-sne/UMAP 적용, Genus OTU table에 PCA+t-sne/UMAP이 적용된 두 그래프를 비교해보면, 오히려 샘플이 적은 OTU table에서 군집화가 비교적 잘 된 것을 확인할 수 있다. 물론 샘플 수가 적었기 때문일 가능성이 매우 크지만, 만약 샘플 수가 많아졌을 때도 PCA+t-sne 와 PCA+UMAP의 결과가 잘 군집화된다면 microbiome OTU table 데이터의 특성을 잘 설명할 수 있는 차원 축소 방법으로 고려할 수 있다.

Ⅳ. 참고 문헌

https://www.kaggle.com/iabhishekofficial/mobile-price-classification
https://data-newbie.tistory.com/295
https://www.ebi.ac.uk/metagenomics/