

복부암 다중 오믹스 유전체 데이터 기반의 생물학기작 활성화 정도 분석 방법

전재민, 정인욱*
IT 대학 컴퓨터학부

경북대학교

jaemin1479@knu.ac.kr, inukjung@knu.ac.kr*

Multi Omics Pathway Activation Analysis

Jae Min Jeon, Inuk Jung*
School of Computer Science and Engineering
Kyungpook National University

요 약

본 논문에서는 PARAFAC 을 이용한 텐서 분해기법을 활용하여 다중 오믹스 유전체 데이터 기반 유전자 패스웨이 활성화 정도를 측정하기 위한 방법을 제안한다. 단일 생물체로부터 다양한 유전체 정보가 존재하는데 단일 오믹스에 비해 다중(또는 복수) 오믹스를 활용했을 때 복부암의 기작을 더 잘 설명할 수 있음을 보였다. 또한, 본 방법을 활용하여 복부암 환자들의 임상 변수를 라벨로 정의하여 이들에 대한 분류 정확도가 다른 방법들에 비해 높은 성능을 가졌음을 보였다.

I. 서 론

바이오의학 기술의 발전으로 생물체로부터 다양한 종류의 오믹스(omics) 데이터와 세부정보를 수집할 수 있게 되었다. 오믹스는 유전체 정보를 뜻하며 여기에는 유전자 (e.g., DNA, RNA), 단백질, 대사체 등의 다양한 생체물질을 포괄한다. 기존에는 단일 오믹스를 통하여 생물학적 복잡성을 설명하였다면 다중 오믹스를 사용했을 때는 보다 다양한 관점에서 이를 설명 할 수 있게 된다. 환자들의 생물학적 복잡성의 정보를 패스웨이(pathway)를 통하여 설명을 할 수 있다. 오믹스 데이터를 통해 환자들의 Gene set analysis (GSA)는 오믹스 데이터를 통하여 특정 유전자들의 특징을 발굴하고 패스웨이를 통하여 분석하는 기법이다. 패스웨이는 생체 활동에서 중요한 경로에 해당되는 유전자들을 각각 모아 놓은 유전자 집합이라고 할 수 있다. 기존에는 하나의 오믹스를 사용한 single sample Gene Set Enrichment Analysis (ssGSEA)[1], Gene Set Variation Analysis(GSVA)[2], Zscore 등의 기법이 있다. 다중 오믹스를 사용하여 패스웨이에 대하여 하나의 샘플 당 스코어를 구하는 경우는 Multi Omics Gene set Analysis(MOGSA)[3]가 있다. Multi Factor Analysis(MFA)[4]를 사용한 MOGSA는 다중 오믹스를 합하여 패스웨이 점수를 계산한다. 그러나 MOGSA 에서는 패스웨이 스코어를 통하여 subtype 분류 성능이 부족하였고 PCA (Principal component analysis) 사용하여 설명을 잘하는 component 만 뽑아서 하나의 오믹스에 편중화 되는 경향이 있다. 본 논문에서는 mRNA 발현량, microRNA 발현량, DNA methylation level 의 오믹스 데이터를 사용하여 패스웨이 활성화를 분석한다. 본 논문은 3 가지 오믹스를 텐서화를 통해 유전자들을 서열화해서 각

환자의 패스웨이 활성화를 분석한다. 이 활성화 점수를 통해 각 임상정보마다 분류 정확도와 클러스터링 정도를 기존 기법들과 비교하여 성능을 보여준다.

II. 본론

2.1 데이터 및 정규화

데이터는 암유전체지도 TCGA (The Cancer Genome Atlas)[5] 프로젝트에 있는 복부암(Stomach adenocarcinoma, STAD)에 해당되는 환자 유전자 정보를 사용하였다. 복부암 환자에 대한 7 가지 임상 타입 데이터를 사용한다 (표 1).

표 1. 복부암 임상타입에 따른 데이터

임상타입	샘플 수	아형	유전자 개수
Hypermethylation category	311	[CIMP-H, CIMP_EBV, GEA_CIMP-L, Non-CIMP]	14,454
Hypermethylated	312	[0,1]	
Molecular subtype	305	[MSI-H,MSI-L,MSS]	
Pathologic_N	312	NO,N1,N2,N3,N3a	
Pathologic_M	299	M0,M1,Mx	
Stage	310	'I', 'II', 'III', 'IV'	
ABSOLUTE	396	['AC', 'AF', 'SNP']	

오믹스 데이터는 mRNA, miRNA, methylation level 오믹스를 사용한다. 유전자 집합 데이터는 Reactome Pathway Database[6]에서 암과 관련된 300 개 데이터를 사용한다. 3 개의 오믹스 원시 데이터를 quantile 정규화를 이용하여 정규화를 한다

2.2 다중 오믹스 패스웨이 분석 모델

본 논문에서는 텐서 분해를 통한 다중 오믹스 유전자 집합 스코어 모델을 제안한다. 이 모델은 크게 텐서 분해와 그 텐서를 사용한 패스웨이 별 각 샘플의 점수를 계산하는 과정으로 구성한다.

$$T = \sum_{f=1}^R g_f \otimes p_f \otimes o_f \quad (1)$$

(1)번 수식에서의 유전자 텐서 g_f , 샘플 텐서 p_f , 오믹스 텐서 o_f , 3 개의 텐서로 하나의 T 가 구성되는 것을 볼 수 있다. $\otimes$ 수식은 행렬 곱셈이다. 3 개의 정보가 섞인 값으로 분해된 각 텐서를 이용하여 패스웨이 활성화 점수를 구한다. 앞의 수식에서 만든 텐서들과 패스웨이 바이너리 데이터를 입력 값으로 받는다. 본 모델에서 오믹스별 텐서값은 유전자 $g_{(i)} = \{g_{ji}, \dots, g_{ji}\}$, 환자 텐서 $s_{(k)} = \{s_{jk}, \dots, s_{jk}\}$ 으로 표시한다. j는 랭크, i는 유전자, k는 샘플을 뜻한다. $g_{(i)}$ 는 평균 0 표준편차 1의 표준화를 한다. 그리고 음수 값을 제거하기 위하여 전체 텐서에서 최소값을 빼 모든 텐서 값이 양수가 되도록 한다. 환자 텐서 값에서 각 랭크마다 높은 값을 가지는 환자들을 뽑는다. 이 기준은 가우시안 커널 함수를 통하여 최빈값을 가지는 샘플 이상의 샘플 텐서들을 선택한다. 한 환자는 여러 랭크에 속하는 것을 볼 수 있고 (2)번과 같이 각 랭크에 해당되는 유전자 텐서 값과 환자 텐서 값을 곱하여 더한다. (2) 수식으로 다중 오믹스를 통합한 환자에 따른 유전자 중요도가 구해지며 이를 유전자 서열화를 하는데 사용 한다.

$$r_{(j)} = \sum s_{(j)} \times g_{(j)} \quad (2)$$

유전자 서열화 데이터 $r_{(j)}$ 를 GSVA에 사용된 Kolmogorov-Smirnov (KS) like random walk statistic[1]를 변형하여 각 환자 당 패스웨이의 스코어를 계산한다. 기존의 KS에서는 $r_{(j)}$ 의 순서를 사용하였지만 본 모델에서는 $r_{(j)}$ 값 자체의 차이를 이용한다.

$$E_{jt} = \frac{\sum_{j=1}^l r_{(j)} I(G_{(j)} \in p_{(t)})}{\sum_{j=1}^p r_{(j)} I(G_{(j)} \in p_{(t)})} - \frac{\sum_{j=1}^l r_{(j)} I(G_{(j)} \notin p_{(t)})}{|G_{(j)}| - |p_{(t)}|} \quad (3)$$

$p_{(t)}$ 는 유전자 집합이고 $G_{(j)}$ 는 입력 값이 되는 유전자들이다. $I(G_{(j)} \in p_{(t)})$ 은 지시 함수로 패스웨이에 해당 유전자가 포함되는지를 표시하는 함수이다. 속한다면 1, 아니면 0의 값을 가지게 된다.

$$\text{diff} = |E_{jt}^+| - |E_{jt}^-| \quad (4)$$

(3)번 수식에서 나오는 값들 중에 최대값 (E_{jk}^+)과 최소값 (E_{jk}^-)의 차이를 (4)번 수식에 대입을 한다. +와 -는 패스웨이에 속하는 유전자들이 아닌 유전자들 보다 높은 값을 가지는 여부에 대한 수식이다. (4)번 수식의 결과가 한 유전자 집합의 한 샘플의 스코어이다. 스코어의 값은 -1~1 사이이며 높을수록 유전자 집합에 해당되는 유전자의 중요도가 높음을 알 수 있다.

2.3 실험 결과

멀티 오믹스를 활용한 패스웨이 스코어가 실제 데이터에서도 아형에 따른 분류를 도와주는지 알아보았다. 이를 위해 복부암의 7 가지 임상타입에 따른 데이터를 사용하여 분석을 하였다.

그림 1은 TCGA에서 제공하는 STAD의 7 가지 clinical data에 따라 활성화 스코어가 얼마나 잘 구분하는지를 레이더 차트로 표시를 하였다.

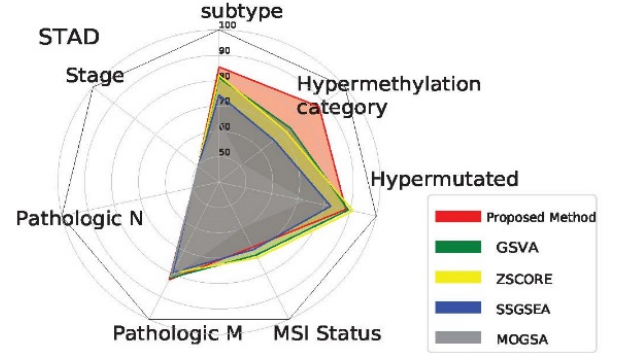


그림 1. STAD 임상정보에 따른 모델들의 MLP f1-score

Multilayer perceptron(MLP) 분류기를 사용하여 10 교차검증을 하였고 테스트 샘플과 훈련 데이터를 9:1로 나누어서 f1-score 계산을 하였다. f1-score를 사용한 이유는 아형마다 샘플 수의 차이가 많이 차이났다. 본 모델이 다른 모델들과 비슷한 정확도를 가지지만 subtype, Hypermethylation category 임상에서는 f1-score가 높은 것을 볼 수 있다. 새로운 모델은 methylation 오믹스를 사용하였기 때문에 mRNA만 사용하는 기법들보다는 높은 분류 성능을 보였다. 그러나 같은 오믹스들을 사용한 MOGSA보다도 정확도가 높은 것을 볼 수 있다. 그림 2의 (B) hypermethylation category에서 클러스터링이 어느정도 되는지를 ARI(Adjusted Rand Index)와 AMI(Adjusted Mutual information)를 사용하여 보여준다.

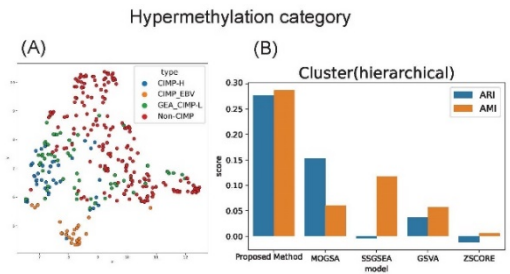


그림 2. Hypermethylation category에 대한 (A) tsne_plot (B) 클러스터링 스코어

본 논문에서 제안하는 모델이 다른 기존 모델 보다 높은 점수를 가지는 것을 볼 수 있다. 그리고 본 논문의 모델의 활성화 스코어의 TSNE-plot으로 얼마나 잘 구분되는지를 확인한다.

III. 결론

본 논문에서는 질병을 가진 환자들의 패스웨이 분석을 하기 위하여 텐서 기반 다중 오믹스 패스웨이 활성화 기법을 소개한다. 각 환자에 따라 값이 높은 텐서를 추출하고 패스웨이에 따른 생물학 기작을

활성화 정도를 구하기 때문에 아형끼리 비슷한 활성화 정도를 가지며 클러스터링 성능이 뛰어났다. 이를 복부암 환자 데이터셋을 사용하여 성능 평가한 결과 기존 기법보다 높은 성능을 보이는 것을 볼 수 있었다.

ACKNOWLEDGMENT

이 논문은 2020 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 한국연구재단에서 부여한 과제번호: 2020M3C9A5085604)

참 고 문 헌

- [1] Barbie, David A., et al. "Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1," *Nature* 462.7269(2009):108-112.
- [2] Hanzelmann, Sonja, Robert Castelo, and Justin Guinney. "GSVA : gene set variation analysis for microarray and RNA-seq data," *BMC bioinformatics* 14.1 (2013): 1-15
- [3] Meng, Chen, et al. "MOGSA: integrative single sample gene-set analysis of multiple omics data," *Molecular & Cellular Proteomics*, 18(8), S153-S168
- [4] Abdi, Herve, Lynne J. Williams, and Dominique Valentin. "Multiple factor analysis: principal component analysis for multitable and multiblock data sets," *Wiley Interdisciplinary reviews: computational statistics* 5.2(2013):149-179
- [5] Tomczak, Katarzyna, Patycja Czerwiriska, and Maciej Wizenerowicz. "The Cancer Genome Atlas(TCGA): an immeasurable source of knowledge," *Contemporary oncology* 19.1A(2015).A68
- [6] Fabregat, Antonio, et al. "The reactome pathway Knowledgebase," *Nucleic acids research*, 46.D1(2018):D649-D655.