멀티 오믹스 데이터 및 생물학적 네트워크 정보를 이용한 드라이버 유전자 분류

박정호[1], 조겨리[1]
[1]충북대학교 컴퓨터공학과
yu065812@naver.com, kyurijo@chungbuk.ac.kr

Cancer driver gene using multi-omics data and biological network information

Jeong-Ho Park[1], Kyuri Jo[1]
[1]Department of Computer Engineering, Chungbuk National University

요 약

시퀀싱(sequencing) 기술의 발달로 다양한 오믹스(omics) 데이터의 축적과 인공 지능 기술의 발달로 인하여 다양한 드라이버 유전자 분류기법이 제안되어왔다. 최근에는 암 데이터가 대용량으로 축적되며 기계 학습 기반의 다양한 기법들이 활발히 제안되었다. 특히 다양한 오믹스 데이터를 결합한 고차원 데이터에서 높은 정확도를 확보하기 위한 시도가 활발히 이루어지고 있다. 본 논문에서는 멀티 오믹스와 네트워크 관련 특징을 기반으로 암의 증식 및 발생에 중요한 역할을 하는 드라이버 유전자를 분류하는 딥러닝 모델을 제시한다. 또한 The Cancer Genome Atlas(TCGA) 데이터를 통해서 모델 학습 후 기존 통계 및 머신러닝 기반 기법과 비교하여 성능이 개선되었음을 확인하였다.

1. 서 론

암은 유전자의 돌연변이로 인하여 세포가 비정상적으로 증식되는 질병이다. 지난 수십년간 암의 발생 및 증식에 중요한 역할을 하는 드라이버 유전자를 식별하기 위한 연구가 활발히 시도되었다. 최근에는 다수 축적된 암 데이터와 기계 학습 및 통계적 방법을 기반으로 드라이버 유전자를 식별하기 위한 연구들이 활발하게 진행중이다. 기계 학습 기반의 분류 기법으로 GUST[1], 통계적 방법 기반으로 DiffMUT[2] 등이 개발된 바 있다. 또한 최근 딥러닝 기법을 활용하여 고차원의 멀티 오믹스 데이터를 기반으로 드라이버 유전자 분류를 수행하는 DeepDriver[3] 등이 제안되었다.

본 연구에서는 TCGA 데이터베이스에서 공개된 암 멀티 오믹스 데이터와 네트워크 데이터를 통한 딥 러닝 기반의 드라이버 유전자 분류 기법을 제안한다. 드라이버 유전자 분류를 위한 입력 특징으로 유전자 내 돌연변이 정보, 유전자의 차등 발현량, 유전자의 특정 생물학적 패스웨이 소속 정보, 유전자의 단백질 상호작용 정보를 활용한다. 본 논문에서 제시한 모델의 우수성을 평가하기 위해 성능을 측정 후 GUST[1], DiffMut[2], DeepDriver[3]와 성능을 비교한다.

2. 본 론

2.1 데이터셋 구성

TCGA 암 환자의 돌연변이 데이터와 유전자 발현 데이터를 UCSC Xena[4], 암 연관 패스웨이 정보를 KEGG[5] 데이터베이스, 단백질 상호작용 네트워크를 HumanNetv3[6]로부터 수집하였다. 또한, CGCv95[7]에서 암 관련 드라이버 유전자로 알려진 유전자 목록을 받아, 분류 모델의 정답 레이블(label)로 사용하였다. 최종 데이터셋은 활용 가능한 드라이버 유전자가 12개 이상인 14개 암에 대해서 생성되었다.

2.2 유전체 기반 돌연변이 입력 특징

돌연변이 정보는 암 발생 및 촉진에 있어서 가장 중요한 정보 이기 때문에 드라이버 유전자 분류에서도 중요한 역할을 한다. 본 모델에서는 표 1과 같이 해당 유전자에 존재하는 돌연변이의 특성을 반영하는 9개의 입력 특징을 활용하였다.

<표 1> 돌연변이 기반 입력 특징 및 설명 대응표

특징	설명
과오 돌연변이 (missense variant) 비율	과오 돌연변이(missense variant)의 비율
스플라이스 부위 돌연변이 (splice region variant) 비율	스플라이스 부위 돌연변이(splice region variant)의 비율
인프레임 돌연변이 비율	인프레임 삽입 돌연변이(inframe insertion)와 보수적 인프레임 삭제 돌연변이(conservative inframe deletion) 그리고 인프레임 삭제 돌연변이(inframe deletion)를 종합한 비율
분실 돌연변이 비율	시작 코돈 분실 돌연변이(start lost)와 멈춤 코돈 분실 돌연변이(stop lost)를 종합한 비율
동의 돌연변이 (synonymous	동의 돌연변이(synonymous

variant) 비율	variant) 비율의 비율
멈춤 코돈 획득 돌연변이 (stop gained) 비율	멈춤 코돈 획득 돌연변이(stop gained)의 비율
틀이동 돌연변이 (frame shift variant) 비율	틀이동 돌연변이(frameshift variant)의 비율
과오 돌연변이 (missense variant) 상대적 비율	동의 돌연변이(synonymous variant 대비 과오 돌연변이(missense variant)의 비율
중요 돌연변이 합	중요 돌연변이 종류의 (스플라이스 부위 돌연변이(splice region variant), 멈춤 코돈 획득 돌연변이(stop gained), 과오 돌연변이(missense variant), 틀이동 돌연변이(frameshift variant)) 빈도수를 합한 후 표준화 한 수치

2.3 네트워크 기반 유전자 정보

네트워크 기반 특징으로는 KEGG[5]의 패스웨이 정보 및 HumanNetv3[6] 단백질 상호 작용 네트워크를 활용하였다.

<표 2> 네트워크 정보에 이용한 각 암 별 패스웨이 목록

암	패스웨이 ID	패스웨이 이름	연관 패스웨이 개수
BRCA	hsa05224	Breast cancer	8
PAAD	hsa05212	Pancreatic cancer	9
PRAD	hsa05215	Prostate cancer	8
BLCA	hsa05219	Bladder cancer	6
GBM	hsa05214	glioma	7
READ	hsa05210	Colorectal cancer	9
SKCM	hsa05218	melanoma	6
KIRP	hsa05211	Renal cell carcinoma	6
KIRC	hsa05211	Renal cell carcinoma	6
LUSC	hsa05223	Non-small cell lung cancer	7
LUAD	hsa05223	Non-small cell lung cancer	7
SARC	hsa05200	Pathway in cancer	21
COAD	hsa05210	Colorectal cancer	9
HNSC	hsa05200	Pathway in cancer	21

생물학적 패스웨이는 특정한 생물학적 기작에 관여하는 것으로 밝혀진 유전자들의 집합 및 상호작용을 나타낸 네트워크로, 각각의 암과 연관된 패스웨이에 많이 속하는 유전자일수록 드라이버 유전자일 가능성이 높다고 가정하고 암 관련 패스웨이(표 2)에 속하는지 여부(dir_pathway)와 KEGG에 명시된 암 연관 패스웨이에 속하는 횟수(related_pathway)를 표준화하여 특징으로 활용하였다.

유전자는 단백질 사이의 상호작용을 통해 동작하기 때문에 단백질 상호작용 네트워크는 드라이버 유전자 분류에 도움이 될 수 있다. 본 논문에서는 HumanNetv3[6] 단백질 상호작용 네트워크를 필터링 후 예측을 하고자 하는 유전자의 이웃 유전자에 대한 정보를 입력 특징으로 활용하였다. 먼저 단백질 상호작용 네트워크의 간선 점수를 기준으로 상위 5% 미만의 간선들은 제거하였다. 그 후 암 발생에 있어서 중요 돌연변이(스플라이스 부위 돌연변이, 멈춤 코돈 획득 돌연변이, 과오 돌연변이, 틀이동 돌연변이)가 존재하는 유전자의 이웃 유전자 중 중요 돌연변이 합 특징이 상위 25%에 해당되는 유전자의 수를 해당 유전자의 값으로 하고, 중요 돌연변이가 존재하지 않는 유전자는 0을 해당 유전자의 값으로 한다. 그 후 각 유전자들의 값을 표준화하여 입력 특징으로 활용한다(식 3).

$$n_j = \begin{cases} 1 & if\ gene\ j \in P \\ 0 & otherwise \end{cases} \quad (1)$$

$$D_i = \begin{cases} \sum_{j \in N_i} n_j & if\ gene\ i \in X \\ 0 & otherwise \end{cases} \quad (2)$$

$$g_i = \frac{D_i - m}{\sigma} \quad (3)$$

X는 중요 돌연변이가 존재하는 유전자 집합이며 P는 중요 돌연변이 합 특징이 상위 25%에 해당되는 유전자들의 집합, N_i는 유전자 i의 이웃 유전자 집합이다.

2.4 전사체 기반 유전자 발현량 정보

전사체 기반 특징으로 본 논문에서는 각 유전자의 차등 유전자 발현량을 활용하였다. 각 유전자의 상대적인 발현량을 측정하기 위해 비교군의 유전자 별 평균 발현량(treatment)과 대조군의 유전자 별 평균 발현량(control)을 통해 상대적 값을 얻을 수 있는 log2 fold change($log_2(FC)$)를 계산하였다(식 4).

$$\boldsymbol{log_2(FC) = log_2(\frac{Treatment}{Controal})} \quad (4)$$

$log_2(FC)$의 비교군으로는 TCGA 암 RNA 시퀀싱 데이터 (HTSeq) 중 암세포 샘플의 유전자 발현량, 대조군으로는 정상 세포 샘플에서의 유전자 발현량 평균을 사용하였다. 즉, 차등 유전자 발현량은 해당 유전자가 정상 세포보다 암 세포에서 얼마나 특이적으로 발현되는지를 나타낸다.

2.5 클래스 불균형

본 논문에서는 학습 데이터의 클래스간의 불균형(class imbalance)을 해소하기 위해 클래스 가중치(class weight)를 사용하였다. 각 클래스 별 가중치 계산 식은 식 5와 같다.

$$\mathrm{CW} = \frac{1}{N_C} \times \frac{N_t}{2} \quad (5)$$

N_c는 해당 클래스의 샘플 수이며, N_t는 전체 클래스의 샘플 수이다.

2.6 학습 모델

본 논문에서는 학습 모델로 깊은 신경망 네트워크를 사용하였다. 모델은 총 4개의 전결합 계층(fully connected layer)으로 구성되었으며, 학습률(learning rate)은 0.001, 최적화 알고리즘(optimizer)으로는 Adam, 손실 함수(loss function)로는 희소 범주형 교차 엔트로피(sparse categorical cross entropy)를 사용하였다. 각 암 데이터 별로 최적의 구조를 찾기 위해 그리드 검색(grid search)을 수행하였다.

3. 실험 결과

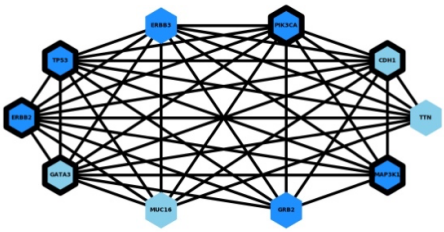
3.1 모델 성능 비교

표 3은 멀티 오믹스 데이터 및 네트워크 데이터를 학습시킨 드라이버 유전자 분류 모델의 ROC-AUC(area under the curve)을 측정한 표이다. 교차 검증(cross validation)기법을 통하여 3개의 각기 다른 학습 데이터셋 및 테스트 데이터 셋을 조합 생성하여 3번 반복 실험하였다. 실험 결과 GUST[1], DiffMut[2], deepDriver[3]와 비교하여 본 논문 모델의 성능이 우수한 것을 확인할 수 있었다.

<표 3> TCGA 데이터를 통한 성능 비교

Cancer name	Our method	GUST	DiffMut	deepDriver
BRCA	**0.9304±0.03**	0.8022±0.03	0.7464±0.05	0.8016±0.02
PAAD	**0.9399±0.03**	0.7045±0.1	0.6742±0.11	0.7010±0.11
PRAD	**0.9053±0.05**	0.7089±0.06	0.5720±0.07	0.7060±0.09
BLCA	**0.9541±0.02**	0.8004±0.07	0.8487±0.12	0.7448±0.09
GBM	**0.9380±0.04**	0.7058±0.09	0.8685±0.12	0.7721±0.1
READ	**0.9029±0.05**	0.7038±0.16	0.7787±0.05	0.6115±0.12
SKCM	**0.8267±0.03**	0.7766±0.04	0.6864±0.05	0.6375±0.05
KIRP	**0.9104±0.05**	0.7934±0.06	0.7845±0.02	0.6685±0.12
KIRC	**0.8406±0.04**	0.5839±0.02	0.7139±0.06	0.6908±0.14
LUSC	**0.8904±0.02**	0.6381±0.04	0.6724±0.06	0.7028±0.01
LUAD	**0.8943±0.02**	0.6538±0.05	0.6662±0.04	0.7042±0.03
SARC	**0.6584±0.10**	0.6230±0.11	0.6191±0.17	0.4795±0.12
COAD	**0.8914±0.06**	0.5839±0.09	0.6825±0.08	0.7339±0.09
HNSC	**0.8731±0.06**	0.6802±0.08	0.5801±0.13	0.7053±0.08

3.2 잠재적 암 유발 유전자 탐지 및 높은 예측 점수 유전자 사이의 연결성 확인



(그림 1) 유방암(BRCA)에서 예측 점수 기준 상위 10 개의 유전자의 네트워크: 파란색 노드의 경우 KEGG[5]에 명시된 유방암(BRCA)와 관련된 패스웨이에 속하는 유전자, 굵은 테두리 노드들의 경우 CGCv95[7]에 속하는 유전자

높은 예측 점수를 가지며 CGCv95[7]에 포함되지 않는 유전자는 높은 가능성을 가진 잠재적 드라이버 유전자이다. 본 논문에서는 실험 3.1 의 3 개의 테스트 데이터 셋에 대한 결과를 합치고 CGCv95[7]에 포함되는 유전자는 제외한 후 예측 점수 기준 상위 5 개의 유전자를 뽑아 해당 유전자들에 대한 문헌을 탐색하였다. 유방암(BRCA)의 경우 상위 5 개의 유전자(MUC16, GRB2, ERBB3, TTN, UBC) 중 3 개의 유전자에서 관련 문헌을 찾을 수 있었다. MUC16 유전자는 유방암 확산에 있어서 이중적인 역할을 하고[8], GRB2 유전자 발현의 하향 조절을 통해서 유방암 세포의 성장을 억제할 수 있다[9]. ERBB3 는 이량체 기능에 의해 유방암 세포의 증식에 기여한다[10]. 이러한 문헌 검색 결과를 보아, 본 논문에서 제시한 예측 모델이 신빙성 있는 예측을 한다는 것을 알 수 있다.

또한 본 논문에서는 높은 점수의 유전자 사이에 연결성을 확인하기 위해 유방암 데이터에서 예측 점수 기준 상위 10 개 유전자를 뽑은 후 STRING[11] 그래프를 활용하여 서브 그래프(subgraph)를 구성하였다. 서브 그래프(subgraph)는 그림 1 에 표현되었다. 실험 결과 상위 10 개의 유전자 중 6 개의 유전자가 CGCv95[7]에 포함되었으며, 상위 10 개의 유전자 중 6 개의 유전자가 KEGG[5]에 명시되어 있는 유방암(BRCA) 관련 패스웨이에 속하였다. 예측 점수 기준 상위 10 개의 유전자들은 유방암에 있어서 중요한 역할을 함과 동시에 높은 연결성을 보이며, 이를 보았을 때 상호간에 높은 가능성으로 영향력을 행사함을 볼 수 있다.

4. 결론 및 향후 연구 방향

본 연구에서는 유전체, 전사체, 네트워크 관련 입력 특징을 이용하여 드라이버 유전자를 분류하는 모델을 제시하였다. 실험 결과 기존 통계 및 머신러닝 기법에 비해 우수한 성능을 보였으며, 클래스 불균형 문제를 효과적으로 해결할 수 있었다. 그러나 본 연구에서 활용된 모델은 네트워크 중심성(centrality)를 간접적으로 반영하고 있으므로 이를 개선할 필요성이 있다. 또한, 생물학적 네트워크 정보는 실험적으로 규명되지 않은 상호작용이 배제되어 있는 등 데이터의 불완전성 및 편향이 존재할 수 있으므로 이러한 정보를 보완할 수 있는 모델을 후속연구를 통해 개발하고자 한다.

5. Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 지역 지능화혁신인재양성(Grand ICT연구센터) 사업의 연구결과로 수행되었음 (IITP-2023-2020-0-01462).

참고문헌

[1] Chandrashekar, P., Ahmadinejad, N., Wang, J., Sekulic, A., Egan, J. B., Asmann, Y. W., ... & Liu, L. Somatic selection distinguishes oncogenes and tumor suppressor genes. Bioinformatics, 36(6), 1712-1717, 2020.

[2] Przytycki, P. F., & Singh, M. Differential analysis between somatic mutation and germline variation profiles reveals cancer-related genes. Genome medicine, 9(1), 1-11, 2017

[3] Luo, P., Ding, Y., Lei, X., & Wu, F. X. deepDriver: predicting cancer driver genes based on somatic mutations using deep convolutional neural networks. Frontiers in genetics, 10, 13, 2019.

[4] Goldman, M. J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., ... & Haussler, D. Visualizing and interpreting cancer genomics data via the Xena platform. Nature biotechnology, 38(6), 675-678, 2020.

[5] Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., & Tanabe, M. KEGG: integrating viruses and cellular organisms. Nucleic acids research, 49(D1), D545-D551, 2021.

[6] Kim, C. Y., Baek, S., Cha, J., Yang, S., Kim, E., Marcotte, E. M., ... & Lee, I. HumanNet v3: an improved database of human gene networks for disease research. Nucleic acids research, 50(D1), D632-D639, 2022.

[7] Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., ... & Campbell, P. J. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic acids research, 43(D1), D805-D811, 2015.

[8] Lakshmanan, I., Ponnusamy, M. P., Das, S., Chakraborty, S., Haridas, D., Mukhopadhyay, P., ... & Batra, S. K. MUC16 induced rapid G2/M transition via interactions with JAK2 for increased proliferation and anti-apoptosis in breast cancer cells. Oncogene, 31(7), 805-817, 2012.

[9] Tari, A. M., Hung, M. C., Li, K., & Lopez-Berestein, G. Growth inhibition of breast cancer cells by Grb2 downregulation is correlated with inactivation of mitogen-activated protein kinase in EGFR, but not in ErbB2, cells. Oncogene, 18(6), 1325-1332, 1999.

[10] Holbro, T., Beerli, R. R., Maurer, F., Koziczak, M., Barbas III, C. F., & Hynes, N. E. The ErbB2/ErbB3 heterodimer functions as an oncogenic unit: ErbB2 requires ErbB3 to drive breast tumor cell proliferation. Proceedings of the National Academy of Sciences, 100(15), 8933-8938, 2003.

[11] Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., ... & Mering, C. V. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic acids research, 47(D1), D607-D613, 2019.