KIPS Trans. Softw. and Data Eng.
Vol.10, No.12 pp.547~554
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2021.10.12.547

Cancer Patient Specific Driver Gene Identification by Personalized Gene Network and PageRank

Jung Hee Won† · Park Ji Woo† · Ahn Jae Gyoon††

ABSTRACT

Cancer patients can have different kinds of cancer driver genes, and identification of these patient-specific cancer driver genes is an important step in the development of personalized cancer treatment and drug development. Several bioinformatic methods have been proposed for this purpose, but there is room for improvement in terms of accuracy. In this paper, we propose NPD (Network based Patient-specific Driver gene identification) for identifying patient-specific cancer driver genes. NPD consists of three steps, constructing a patient-specific gene network, applying the modified PageRank algorithm to assign scores to genes, and identifying cancer driver genes through a score comparison method. We applied NPD on six cancer types of TCGA data, and found that NPD showed generally higher F1 score compared to existing patient-specific cancer driver gene identification methods.

Keywords : Cancer Driver Gene, Gene Network, PageRank, Multi Omics Analysis

개인별 유전자 네트워크 구축 및 페이지랭크를 이용한 환자 특이적 암 유발 유전자 탐색 방법

정 희 원† · 박 지 우† · 안 재 균††

요 약

암을 유발하는 유전자는 모든 암 환자에게 공통적인 것은 아니며, 이러한 환자 특이적 암 유발 유전자의 탐색은 개인 맞춤형 암 치료 및 항암제 개발에 있어서 매우 중요하다. 환자 특이적 암 유발 유전자를 찾기 위한 생물 정보학 연구들이 있어왔지만, 아직 정확도 면에서는 발전의 여지가 있다. 본 논문에서는 환자 특이적 암 유발 유전자를 탐색하기 위하여 NPD (Network based Patient-specific Driver gene identification)라는 방법을 제안한다. NPD는 환자 특이적 유전자 네트워크를 구축하고, 여기에 수정된 PageRank 알고리즘을 적용하여 유전자에 점수를 부여한 후, 유전적 변이 데이터를 사용한 승률 계산 방법을 통하여 암 유발 유전자를 찾는 세 단계로 이루어진다. TCGA 데이터베이스의 여섯 개의 암 데이터에 NPD를 적용한 결과, NPD가 기존의 환자 특이적 암 유발 유전자 탐색 방법들보다 전체적으로 높은 F1 점수를 보여줌을 확인할 수 있었다.

키워드 : 암 유발 유전자, 유전자 네트워크, 페이지랭크, 다중 오믹스 분석

1. 서 론

암 유발 유전자(cancer driver gene)를 정확하게 식별하는 것은 암에 대한 보다 깊은 이해를 가능하게 하며, 표적 항암제나 더 나은 암 치료법의 개발의 시작이 될 수 있다는 점에서 매우 중요하다. Next Generation Sequencing (NGS) 게놈 및 전사체 데이터의 축적에 따라 암 유발 유전자 또는 돌연변이를 식별하기 위한 많은 컴퓨터 기반 방법의 개발이 가능해졌다.

이러한 방법은 크게 세 가지 그룹으로 나누어진다. 첫 번째 그룹은 빈도별로 암 유발 유전자 또는 돌연변이를 식별한다[1]. 이 방법들의 주요 단점은 방대한 양의 데이터가 제공되지 않는 한 희귀한 암 유발 유전자나 돌연변이를 찾을 수 없다는 것이다. 두 번째 그룹은 알려진 암 유발 돌연변이 또는 유전자의 유전적 또는 전사체 패턴을 학습하는 기계 학습 모델을 기반으로 한다[2-5]. 기계 학습 기반 접근 방식은 최근 많은 연구 덕분에 높은 정확도를 보일 것이라 예상된다는 장점이 있지만, 알려진 암 유발 돌연변이나 유전자의 수가 한정적이기 때문에 사용할 수 있는 훈련 데이터의 수가 한정적이다는 한계가 있다. 세 번째 그룹은 암 유발 유전자를 식별하기 위해 유전자 조절 네트워크 또는 단백질-단백질 상호작용 네트워크와 같은 유전자 네트워크에 다양한 네트워크 검색 알고리즘을 적용한다[6,7].

※ 이 논문은 인천대학교 자체연구비(2021-0097) 의하여 연구되었음.
※ 정희원과 박지우는 같은 정도로 논문 작성에 기여하였음.
† 비 회 원 : 인천대학교 컴퓨터공학과 석사과정
†† 정 회 원 : 인천대학교 컴퓨터공학부 부교수
Manuscript Received : September 16, 2021
Accepted : October 28, 2021
* Corresponding Author : Ahn Jae Gyoon(jgahn@inu.ac.kr)

위에서 언급한 방법의 대부분은 암 코호트의 암 유발 유전자를 식별하는 데 중점을 두고 있다. 그러나 동일한 암 유형을 가진 개별 암 환자는 이질적인 암 유발 요인을 가질 가능성이 매우 높다[8,9]. 이러한 암 유발 유전자 중 일부는 돌연변이 빈도가 높으며 그중에서 많은 부분이 잘 연구되어 있지만, 대부분은 드물고 식별하기가 어렵다[10,11]. 개별 암 환자 데이터에서 희귀한 환자 고유의 암 유발 유전자를 찾기 위해 여러 방법이 개발되었다. 이러한 방법은 대부분 세 번째 그룹인 네트워크 검색을 기반으로 하며, 대표적으로 DawnRank[12], Personalized Network Control(PNC)[13], Single-sample Controller Strategy (SCS)[14], PRODIGY[15] 등의 방법들이 있다.

PNC는 Single Sample Network (SSN)라는 방법을 이용하여 환자 특이적 유전자 네트워크에 Maximum Matching Set 알고리즘을 적용하여 전체 네트워크에 영향을 미치는 최소 암 유발 유전자를 찾는다. SSN은 모든 간선이 정상과 암 샘플 사이의 유전자 발현 상관관계에서 유의미한 변화를 보이는 환자 특이적 유전자 네트워크를 만들 수 있다. SCS는 CTC(Constrained Target Control)라는 방법을 사용하여 차등적으로 발현된 유전자를 제어하는 데 필요한 최소 돌연변이 유전자를 식별하며, 이는 하나의 암 유발 유전자와 하위 조절 유전자로 구성된 여러 유전자 모듈을 발생시킨다. PRODIGY는 deregulated path에 대한 돌연변이 유전자의 영향을 계산하여 암 유발 유전자의 순위를 매기는 방법이다.

DawnRank는 변형된 형태의 PageRank[16]를 기반으로 하는데, 변형된 PageRank에서의 댐핑 팩터(damping factor)는 네트워크에서 한 노드로 들어오는 간선의 수를 기반으로 계산된다. 이때 유전자 네트워크의 방향성은 원래 네트워크의 반대로 설정되는데, 이는 원인과 결과를 뒤집는 역할을 한다. 네트워크의 각 유전자는 그 발현 값이 암과 정상 샘플에서 많이 차이를 보일수록 높은 초기 점수를 가지게 되며, 이를 암 유발 유전자의 결과로 해석할 경우 변형된 PageRank는 이러한 결과를 초래한 원인을 찾도록 동작하게 된다. 따라서 DawnRank는 결과에서 원인을 찾아가는 방법이라고 생각할 수 있으며, 변형된 PageRank의 랭크 값이 유전자가 암 유발 유전자일 가능성을 의미하게 된다.

본 논문에서는 환자 특이적 암 유발 유전자를 식별하기 위하여 NPD(Network based Patient-specific Driver gene identification)라는 유전자 네트워크 기반 모델을 제안한다. NPD는 DawnRank와 마찬가지로 결과에서 원인을 찾아가는 방법이다. 하지만 DawnRank와는 달리 환자별로 유전자 가중치 네트워크를 구축하며, 유전자 네트워크의 셀프-루프(self-loop)들의 가중치를 이용하여 댐핑 팩터를 계산하게 된다. 좀 더 구체적으로, 하나의 쌍 지어진 (paired) 암 샘플 및 정상 샘플의 유전자 발현량 사이의 단일표본검증을 통해 얻은 t-통계량으로 셀프-루프의 가중치를 구하게 되며, 1에서 계산된 가중치를 뺀 만큼이 유전자 노드들의 댐핑 팩터가 된다. 그 후 전체 환자와 특정 환자 모두의 상호작용의 강도를 고려하여 유전자 네트워크의 모든 간선의 가중치가 구해지며, 이것이 환자별 유전자 네트워크가 된다. 이러한 환자별 유전자 네트워크에 PageRank 알고리즘을 적용하여 각 유전자에 대한 점수를 매긴 후, 체세포 돌연변이(somatic mutation), 유전자 복제수 변이(copy number alteration), DNA 메틸화(methylation) 등의 유전적 변이 데이터를 이용하여 다시 점수를 매겨 최종 점수를 결정한다.

TCGA[17]의 6가지 암(유방암, 결장암, 간암, 폐암, 췌장암, 위암)에 NPD를 적용한 결과, NPD가 기존의 환자 특이적 암 유발 유전자 식별 방법보다 적은 편차로 높은 F1 점수를 보여줌을 확인할 수 있었다.

2. 방 법

2.1 개요

NPD는 환자 특이적 암 유발 유전자를 식별하기 위한 네트워크 기반 방법으로 1) 환자 특이적 유전자 네트워크 구축, 2) PageRank를 통한 환자별 유전자의 점수 부여, 3) 환자별 유전적 변이 (체세포 돌연변이, 유전자 복제 수 변이, DNA 메틸화) 데이터를 토대로 한 유전자 점수 수정의 세 단계로 구성된다. Fig. 1에 이상의 세 가지 단계를 통하여 유전자에 점수를 부여하는 방법이 제시되어 있다.

2.2 환자 특이적 유전자 네트워크 구축

첫 번째 단계인 환자 특이적 유전자 네트워크를 구축하기 위하여 우선 Reactom[18], RegNetwork[19], TRRUST[20]의 세 가지 유전자 네트워크에서 방향성이 있는 간선만 합친 뒤 간선이 없는 유전자를 삭제해서 통합 네트워크를 만든다. 여기에 각 환자 샘플의 유전자 발현 데이터를 사용하여 환자 특이적 유전자 네트워크를 구축한다. 이 네트워크는 인접 행렬 W로 표현하며, Equation (1)로 구해진다.

$$W=(I-\Psi)\Phi+\Psi \tag{1}$$

Equation (1)에서 I는 단위 행렬이며, Ψ는 대각 행렬로써, 각 대각 원소는 해당 유전자가 주어진 암 샘플 및 모든 정상 샘플 들에서 얼마나 발현 차이를 보이는지를 보여준다. 이 차이는 단일 표본 t-검정을 이용한 t-통계량으로 나타나며, 이 t-통계량은 최소-최대 스케일링(min-max scaling) 방법을 통해서 0.1에서 0.9의 범위로 정규화된다. 또한 한 유전자에서 들어오는 간선이 없을 경우에는 1의 값을 가지

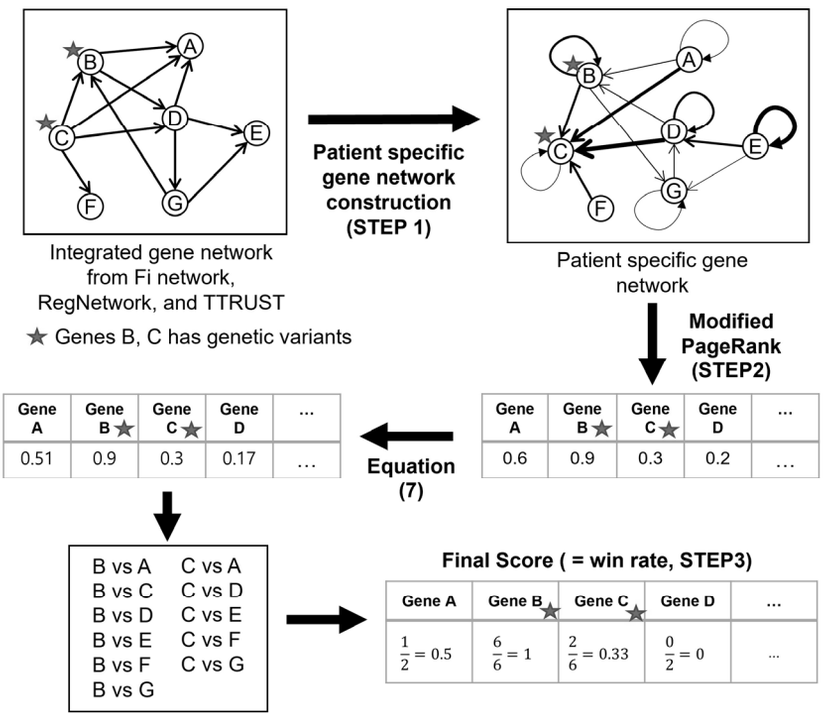

Fig. 1. Methods for Generating Patient-specific Gene Networks and Scoring Cancer Driver Genes

게 된다. 이때 Ψ는 셀프-루프들의 가중치를 의미한다. 그리고 다음 단계에서 사용될 PageRank의 댐핑 팩터를 d라고 할 때, 가중치는 $1-d$와 같다.

행렬 Φ는 셀프-루프 외의 모든 간선의 가중치를 의미하며, Equation (2)에서와 같이 A, W_{CORR}, W_{DIFF}, W_{sample}의 4개 행렬의 요소별 곱(element-wise multiplication)으로 정의된다.

$$\Phi = A \otimes W_{CORR} \otimes W_{DIFF} \otimes W_{sample} \tag{2}$$

A는 통합 유전자 네트워크의 인접 행렬 표현이며, i번째 유전자와 j번째 유전자가 네트워크상에서 연결되어 있다면 A_{ij}= 2 혹은 1의 값을 가지게 되며 그렇지 않다면 0의 값을 가지게 된다. 만약 i번째 유전자가 유전적 변이를 가지고 있다면 A_{ij}= 2가 되며 그렇지 않다면 A_{ij}= 1이다.

W_{CORR}는 Equation (3)과 같이 암 샘플 그룹의 유전자 간의 피어슨 상관 계수(PCC)를 사용하여 계산되며, W_{DIFF}는 Equation (4)와 같이 암 샘플 그룹의 유전자 간의 PCC와 정상 샘플 그룹의 유전자 간의 PCC의 차이로 계산된다.

$$W_{CORR}[i,j] = \left| PCC(X_{cancer}[i], X_{cancer}[j]) \right| \tag{3}$$

$$W_{DIFF}[i,j] = 0.5 * \left| \begin{matrix} PCC(X_{cancer}[i], X_{cancer}[j]) \\ -PCC(X_{normal}[i], X_{normal}[j]) \end{matrix} \right| \tag{4}$$

Equation (3) 및 (4)에서 X_{cancer} 및 X_{normal}은 각각 암과 정상 샘플들의 유전자 발현 행렬을 뜻한다. $X_{DIFF}[i,j]$의 값은 암 샘플 그룹에서의 두 유전자의 발현 상관관계와 정상 샘플 그룹에서의 두 유전자의 발현 상관관계의 차이가 적을수록 0에 가깝고 많을수록 1에 가깝게 된다.

W_{sample}는 다른 암 샘플들과 비교하여 해당 샘플에서 특히 중요한 상호작용을 보여준다. $W_{sample}[i,j]$는 암 샘플들의 PCC와 비교하여 해당 암 환자가 i번째와 j번째 유전자 사이의 선형 상관 패턴이 유사한 경우 1에 가깝다.

$$W_{sample}[i,j] = sigmoid\left(\frac{(X_{cancer}[i,k]-\mu_i)(X_{cancer}[j,k]-\mu_j)}{\sigma_i \sigma_j} * sgn(\rho_{cancer}[i,j]) \right) \tag{5}$$

Equation (5)에서 $X_{cancer}[i,k]$는 k번째 샘플 (주어진 샘플)의 i번째 유전자 발현량 값이고, μ_i와 σ_i는 암 샘플 그룹의 i번째 유전자 발현량의 평균과 표준편차를 나타낸다. 또한 sgn은 입력값의 부호를 반환하는 함수로 1 또는 -1값을 반환한다. $sigmoid$함수는 특정 간선의 두 유전자가 서로 관련이 없을 경우 그 가중치를 0으로 만듦으로써 간선을 없애기 쉽도록 채택된 함수이다.

2.3 페이지랭크 알고리즘을 통한 유전자 점수 계산

환자별 유전자 점수(S_i)는 환자 특이적 유전자 네트워크 W중 셀프 루프를 제외한 부분인 Φ와 댐핑 팩터(d), 초기값

(f)를 통해 Equation (6)과 같이 계산된다.

$$S_t = (1-d)f + d\Phi \times S_{t-1} \quad (6)$$

Equation (6)에서 f는 정상 샘플과 암 샘플 쌍의 유전자 발현량 데이터의 차의 절대값으로 계산한다. 또한 S_t에서 t는 t번째 반복을 의미한다.

그리고 특정 노드에서의 댐핑 팩터 d는 2.2절에서 구했던 셀프-루프의 가중치를 ψ라고 할 때, $1-\psi$와 같다. ψ는 특정 환자와 정상 샘플의 유전자 발현량의 차이와 비례한다. 이 차이가 클수록 암의 유발 유전자라기보다는 암의 결과를 보여주는 유전자이며, 동시에 주변 유전자의 영향을 많이 받는 유전자일 가능성이 크다고 가정할 수 있다. 하지만 앞에서 언급했듯 NPD는 DawnRank와 마찬가지로 결과에서 원인을 찾아가는 방법이므로 환자 특이적 유전자 네트워크를 구성할 때 유전자 네트워크의 방향을 반대로 하여 구성하였다. 따라서 셀프-루프의 가중치가 클수록 주변 유전자의 영향을 덜 받아야 하며, Equation (6)에서 이는 f의 영향력이 커짐으로 해석 가능하다.

2.4 유전적 변이 데이터를 통한 환자별 유전자 점수 보정

마지막으로, 2.3절에서 계산한 유전자 점수를 보정하기 위해 우선 Equation (7)과 같이 유전적 변이가 존재하지 않는 유전자에 대하여 페널티를 부여한다.

$$S_i = \begin{cases} S_i & ,\text{if } gene_{\ i} \ has \ \ genetic \ \ variation \\ S_i \times p\,, & other wise \end{cases} \quad (7)$$

Equation (7)에서 S_i는 i번째 유전자의 유전자 점수를 의미하며, p는 페널티로써, DawnRank에서와 같이 0.85로 설정하여 실험하였다.

보정한 유전자 점수를 바탕으로 유전적 변이가 발생한 유전자들에 대해서, 해당 유전자를 제외한 모든 유전자의 가능한 모든 쌍에 대해 점수를 비교한다. 이로써 유전자의 승률을 계산할 수 있으며, 이 승률이 유전자의 최종 점수가 된다. Fig. 1의 하단에 예제를 통한 보다 자세한 계산법이 제시되어 있다.

3. 결 과

3.1 데이터

6가지 암 유형(유방암, 결장암, 간암, 폐암, 췌장암, 위암)의 4가지 유형의 omics 데이터(유전자 발현, 체세포 변이, 유전자 복제 수 변이 및 DNA 메틸화 데이터)는 TCGA 데이터 포털에서 다운로드하였다. 메틸화 데이터의 경우 각 샘플의 상위 5% 메틸화 수준은 “1”로, 나머지는 “0”으로 대체되었다. 유전자 발현 데이터의 경우 샘플의 80% 이상에서 FPKM이 0인 유전자는 제외되었다. 또한 실험 시 모든 방법은 암세포와 정상 세포가 쌍을 이루는 (paired) 샘플 데이터만 사용하였다. 이상의 데이터 정보가 Table 1에 제시되어 있다.

또한 Reactom에서 제공하는 FI 네트워크, RegNetwork 데이터베이스의 유전자 조절 네트워크, TRRUST 데이터베이스에서 유전자 규제 네트워크를 다운로드하여 방향성이 있는 간선만을 사용하여 통합했다. 알려진 암 유발 유전자 정보는 Intogen[21], CGC[22] 및 NCG[23](Repana el al., 2019)에서 다운로드하였다. CGC에서 제공하는 암 유발 유전자는 2개의 Tier로 나뉘는데 Tier 1에 해당하는 암 유발 유전자가 Tier 2보다 암 발생에 대한 증거가 많기 때문에 Tier 1 유전자만을 사용했다. 통합 네트워크에 대한 자세한 내용은 Table 2에 기술하였다.

Table 1. Detailed Description of Data Downloaded from TCGA Data Portal and Number of Driver Gene by Cancer Type

Cancer Type	Number of Samples (tumor / normal)	Number of Genes	CGC	Intogen	NCG	Num of Known Driver Gene
BRCA	972 / 113	16,473	54	99	711	713
COAD	278 / 41	16,209	58	72		717
LIHC	360 / 50	15,853	29	31		714
LUAD	510 / 59	16,332	27	42		714
PAAD	171 / 4	17,019	32	52		714
STAD	407 / 35	16,526	35	35		718

Table 2. Detailed Description of Network Data

	Num of Nodes	Num of Edges	Num of Total Edges
FI network	14,071	110,721	268,857
Regnetwork	23,336	372,774	372,774
TRRUST	2,852	9,383	9,383
Integrated network	25,167	490,200	648,336

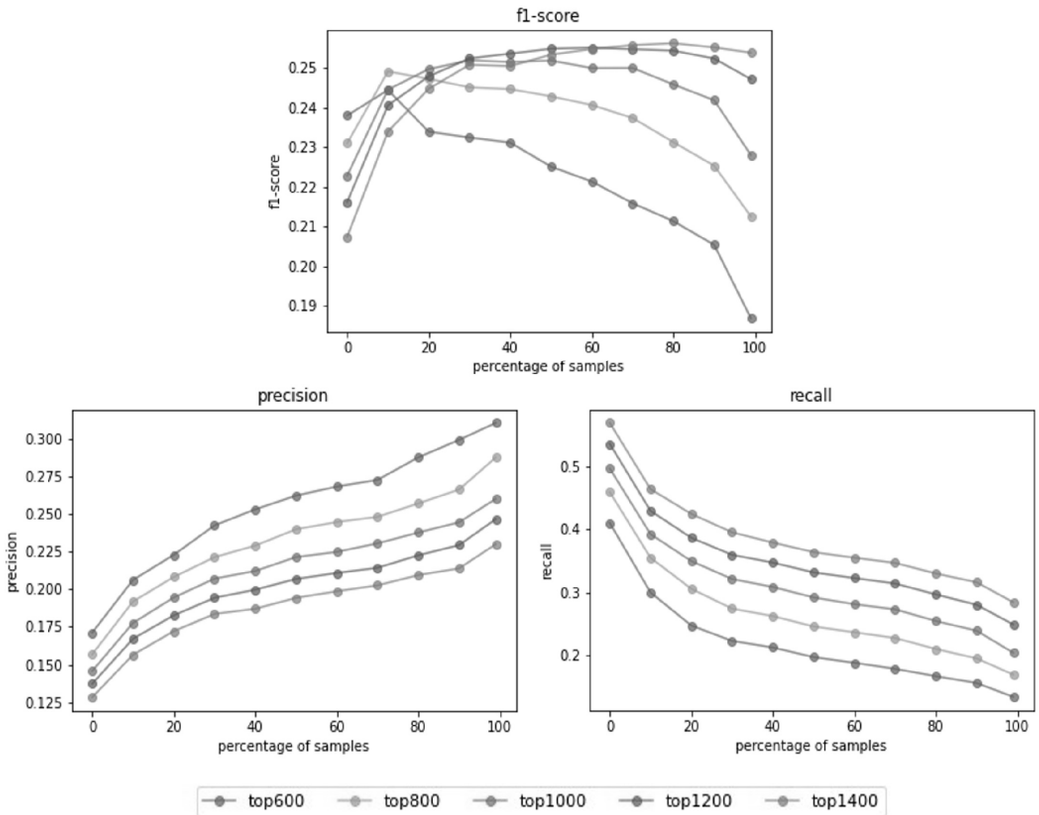

Fig. 2. Comparison of F1 Scores, Precision and Recall Varying the Number of Cancer Driver Gene

3.2 환자별 암 유발 유전자의 선택

NPD를 통해 유전자들에 점수를 부여한 후에는 상위 k개의 유전자를 암 유발 유전자라고 할 수 있다. 최적의 k를 찾기 위하여 6개의 암 종에 대해서 상위 600, 800, 1000, 1200, 1400 개의 유전자를 선택하였다. 그 후 선택된 유전자가 전체 샘플 중 몇 퍼센트 이상의 샘플에서 출현했는지를 계산하여 그 역치(threshold)를 넘는 유전자들에 대해 평가 점수를 계산하였다. 역치는 {0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 99}로 구성된다. 역치가 0인 경우 1개 이상의 샘플에서 나온 유전자를 모두 합친 것을 의미하며, 역치가 99인 경우 는 전체 샘플에서 모두 출현한 유전자만을 의미한다. 각 역치 별로 나온 유전자 집합에 대하여 CGC, Intogen, NCG의 알려진 암 유발 유전자 목록을 이용하여 precision, recall, F1 점수를 계산한 후 평균을 내었다. 그 결과가 Fig. 2에 제시되어 있다.

결과적으로는 전체적으로 높은 F1 점수를 보이고, precision 및 recall이 균형적인 k = 1000을 선택하였다. k = 1200 및 1400의 경우 역치가 50 이상일 경우 1000보다 높은 F1 점수를 보여주고 있지만, 역치가 낮을 경우 낮은 성능을 보여주고 있으므로 선택하지 않았다.

3.3 환자 특이적 암 유발 유전자 식별 방법 비교

NPD를 평가하기 위해 기존의 암 환자 특이적 암 유발 유전자 식별 모델인 DawnRank, PNC, Prodigy와 F1 점수(Fig. 3), precision(Fig. 4), recall(Fig. 5)을 비교하였다. NPD는 3.2절에서 설명하였듯이 환자별로 상위 1000개의 유전자를 암 유발 유전자로 선택하였으며, 나머지 방법은 자동적으로 그 개수가 정해진다.

Fig. 3에서는 6가지 암종에서 NPD의 F1 점수가 다른 방법들과 비교하였을 때 전체적으로 높으며, 그 편차가 적음을 확인할 수 있다. 역치가 0%이 아닌, 즉 적은 수의 샘플에서 발견되는 희귀한 암 유발 유전자들을 제외한 경우 NPD의 F1 점수가 가장 높음을 확인할 수 있다.

DawnRank의 경우 precision이 높지만 그 편차가 크며(Fig. 4), recall이 작은 것을 확인할 수 있다(Fig. 5). 또한 PNC의 경우 반대로 recall이 높지만(Fig. 5), precision이 낮은 것을 확인할 수 있다(Fig. 4).

NPD는 DawnRank를 개선한 방법이나, Fig. 3~5에서 NPD는 DawnRank보다는 PNC와 그 결과의 양상이 더 비슷한 것을 확인할 수 있다. 이는 DawnRank에서는 모든 환자 샘플에 대하여 동일한 가중치 네트워크를 사용됨에 반해

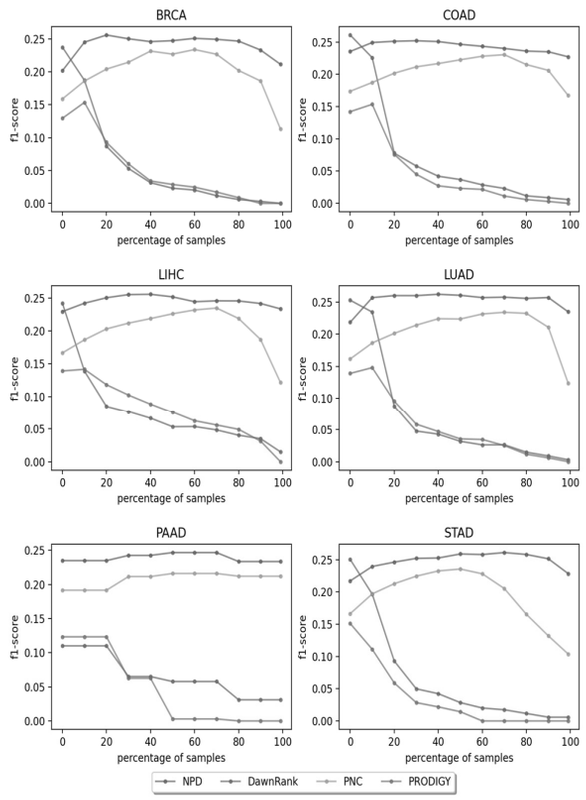

Fig. 3. Comparison of F1 Scores of Patient-specific Cancer Driver Gene Identification Methods

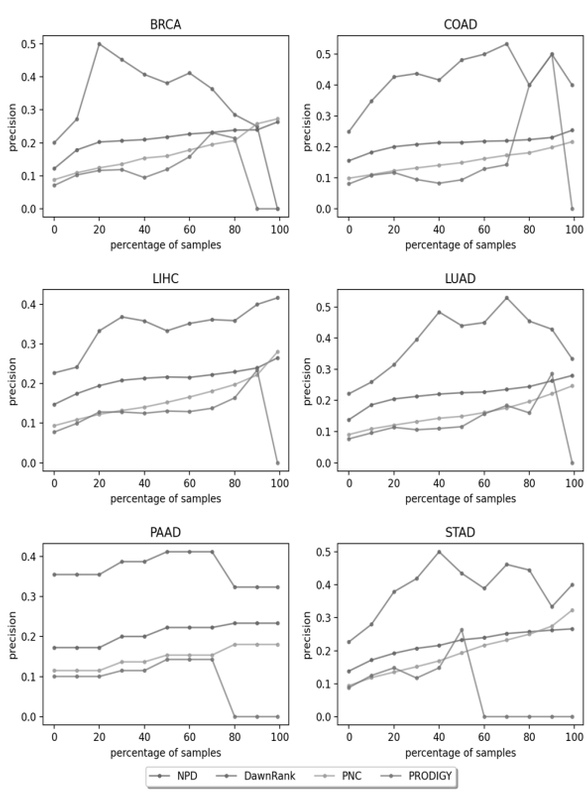

Fig. 4. Comparison of Precision of Patient-specific Cancer Driver Gene Identification Methods

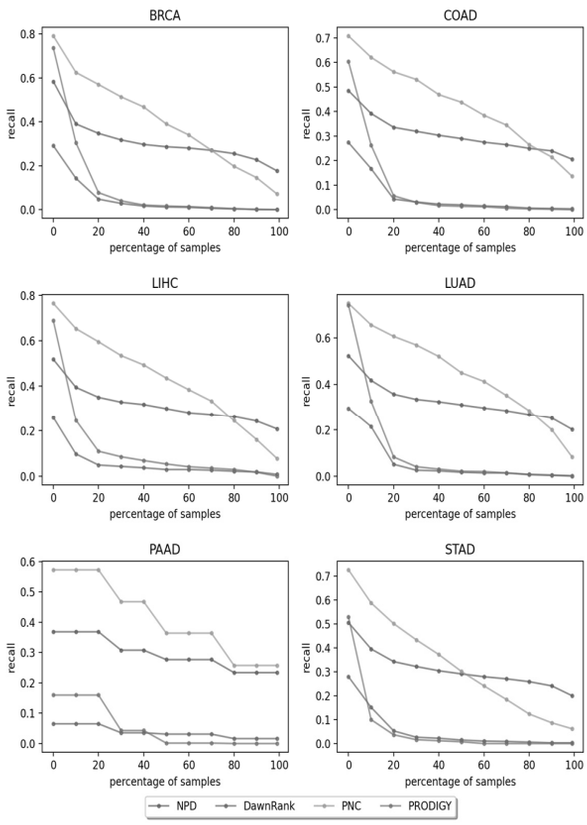

Fig. 5. Comparison of Recall of Patient-specific Cancer Driver Gene Identification Methods

서 NPD과 PNC는 바탕으로 환자별로 다른 네트워크를 구성하여 사용한다는 차이가 있기 때문이라고 추측된다.

4. 결 론

본 논문에서는 환자 특이적 암 유발 유전자를 찾기 위하여 NPD (Network based Patient-specific Driver gene identification)라는법을 제안한다. NPD는 기존의 방법 중 하나인 Dawnrank와 비슷하게 PageRank를 사용하여 암 유발 유전자를 찾는 방식을 채택하고 있으나, Dawnrank와는 다르게 환자 특이적인 유전자 네트워크를 구축한다는 점에서 차이를 보인다. TCGA 데이터베이스의 여섯 개의 암 데이터 및 알려진 암 유발 유전자 데이터베이스를 이용하여 NPD와 기존의 환자 특이적 암 유발 유전자 탐색 방법들을 비교한 결과, NPD가 전체적으로 높은 F1 점수를 보여줌을 확인 가능하였다. 특히 NPD는 Dawnrank 보다 월등히 높은 recall로 인한 높은 F1 점수를 보여주었으며, 이것은 환자 특이적 유전자 네트워크에 기인한다고 예상된다.

References

[1] M. S. Lawrence, et al., "Mutational heterogeneity in cancer and the search for new cancer-associated genes," *Nature*, Vol.499, No.7457, pp.214-218, 2013.

[2] C. Arnedo-Pac, L. Mularoni, F. Muinos, and A. Gonzalez-Perez, and N. Lopez-Bigas, "OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers," *Bioinformatics*, Vol.35, No.22, pp.4788-4790, 2019.

[3] P. Luo, Y. Ding, X. Lei, and F. X. Wu, "deepDriver: Predicting cancer driver genes based on somatic mutations using deep convolutional neural networks," *Frontiers in Genetics*, Vol.10, pp.13, 2019.

[4] J. Nulsen, H. Misetic, C. Yau, and F. D. Ciccarelli, "Pan-cancer detection of driver genes at the single-patient resolution," *Genome Medicine*, Vol.13, No.1, pp.1-14, 2021.

[5] H. Yang, Q. Wei, X. Zhong, H. Yang, and B. Li, "Cancer driver gene discovery through an integrative genomics approach in a non-parametric Bayesian framework," *Bioinformatics*, Vol.33, No.4, pp.483-490, 2017.

[6] D. Bertrand, et al., "Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles," *Nucleic Acids Research*, Vol.43, No.7, pp.e44-e44, 2015.

[7] V. V. Pham, et al., "DriverGroup: A novel method for identifying driver gene groups," *Bioinformatics*, Vol.36, No.(Supplement_2), pp.i583-i591, 2020.

[8] D. Pe'er and N. Hacohen, "Principles and strategies for developing network models in cancer," *Cell*, Vol.144, No.6, pp.864-873, 2011.

[9] M. R. Stratton, "Journeys into the genome of cancer cells," *EMBO Molecular Medicine*, Vol.5, No.2, pp.169-172, 2013.

[10] L. Ding, M. C. Wendl, D. C. Koboldt, and E. R. Mardis, "Analysis of next-generation genomic data in cancer: Accomplishments and challenges," *Human Molecular Genetics*, Vol.19, No.R2, pp.R188-R196, 2010.

[11] J. Reimand and G. D. Bader, "Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers," *Molecular Systems Biology*, Vol.9, No.1, pp.637, 2013.

[12] J. P. Hou and J. Ma. "DawnRank: Discovering personalized driver genes in cancer," *Genome Medicine*, Vol.6, No.7, pp.1-16, 2014.

[13] W. F. Guo, S. W. Zhang, T. Zeng, Y. Li, and J, Gao, "A novel network control model for identifying personalized driver genes in cancer," *PLoS Computational Biology*, Vol.15, No.11, pp.e1007520, 2019.

[14] W. F. Guo, et al., "Discovering personalized driver mutation profiles of single samples in cancer by network control strategy," *Bioinformatics*, Vol.34, No.11, pp.1893-1903, 2018.

[15] G. Dinstag and R. Shamir, "PRODIGY: Personalized prioritization of driver genes," *Bioinformatics*, Vol.36, No.6, pp.1831-1839, 2020.

[16] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," Stanford InfoLab, 1999.

[17] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge," *Contemporary Oncology*, Vol.19, No.1A, pp.A68, 2015.

[18] D. Croft, et al. "Reactome: A database of reactions, pathways and biological processes," *Nucleic Acids Research*, Vol.39, No.(suppl_1), pp.D691-D697, 2010.

[19] Z. P. Liu, C. Wu, H. Miao, and H. Wu, "RegNetwork: An integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse," *Database*, 2015, 2015.

[20] H. Han, et al., "TRRUST v2: An expanded reference database of human and mouse transcriptional regulatory interactions," *Nucleic Acids Research*, Vol.46, No.D1, pp.D380-D386, 2018.

[21] G. Gundem, et al., "IntOGen: Integration and data mining of multidimensional oncogenomic data," *Nature Methods*, Vol.7, No.2, pp.92-93, 2010.

[22] Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, L. Dunham, and S. A. Forbes, "The COSMIC Cancer Gene Census: Describing genetic dysfunction across all human cancers," *Nature Reviews Cancer*, Vol.18, No.11, pp.696-705, 2018.

[23] D. Repana, et al., "The Network of Cancer Genes (NCG): A comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens," *Genome Biology*, Vol.20, No.1, pp.1-12, 2019.

정 희 원

https://orcid.org/0000-0002-5157-1407
e-mail : shs871@inu.ac.kr
2020년 인천대학교 컴퓨터공학과(학사)
2020년 ~ 현 재 인천대학교
컴퓨터공학과 석사과정
관심분야 : 데이터베이스, 바이오인포메틱스, 데이터마이닝, 기계학습

박 지 우

https://orcid.org/0000-0001-5045-1369
e-mail : jw_95@inu.ac.kr
2021년 인천대학교 조형예술학부(학사)
2021년~현 재 인천대학교
컴퓨터공학과 석사과정
관심분야 : 데이터베이스, 바이오인포메틱스,
데이터마이닝, 기계학습

안 재 균

https://orcid.org/0000-0002-7020-7002
e-mail : jgahn@inu.ac.kr
2006년 연세대학교 컴퓨터과학과(학사)
2009년 연세대학교 컴퓨터과학과(석사)
2013년 연세대학교 컴퓨터과학과(박사)
2013년~2015년 UCLA Department of Integrative Biology and Physiology Postdoctoral Research Associate.
2015년~2019년 인천대학교 컴퓨터공학부 조교수
2019년~현 재 인천대학교 컴퓨터공학부 부교수
관심분야 : 데이터베이스, 바이오인포메틱스, 데이터마이닝,
기계학습