



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0148944
(43) 공개일자 2023년10월26일

(51) 국제특허분류(Int. Cl.)

G16B 5/00 (2019.01) C12Q 1/6883 (2018.01)
G16B 20/00 (2019.01) G16B 25/10 (2019.01)
G16H 10/60 (2018.01) G16H 50/20 (2018.01)
G16H 50/50 (2018.01)

(52) CPC특허분류

G16B 5/00 (2019.02)
C12Q 1/6883 (2022.01)

(21) 출원번호 10-2022-0047950

(22) 출원일자 2022년04월19일

심사청구일자 2022년04월19일

(71) 출원인

인하대학교 산학협력단

인천광역시 미추홀구 인하로 100(용현동, 인하대학교)

(72) 발명자

한경숙

서울특별시 서초구 사임당로 137, 3동 309호(서초동, 신동아아파트)

이옥

인천광역시 미추홀구 한나루로463번길 66, 201호(용현동)

이석우

서울특별시 영등포구 여의동로3길 10, 25층 101동 2502호(여의도동, 여의도자이)

(74) 대리인

특허법인태백

전체 청구항 수 : 총 5 항

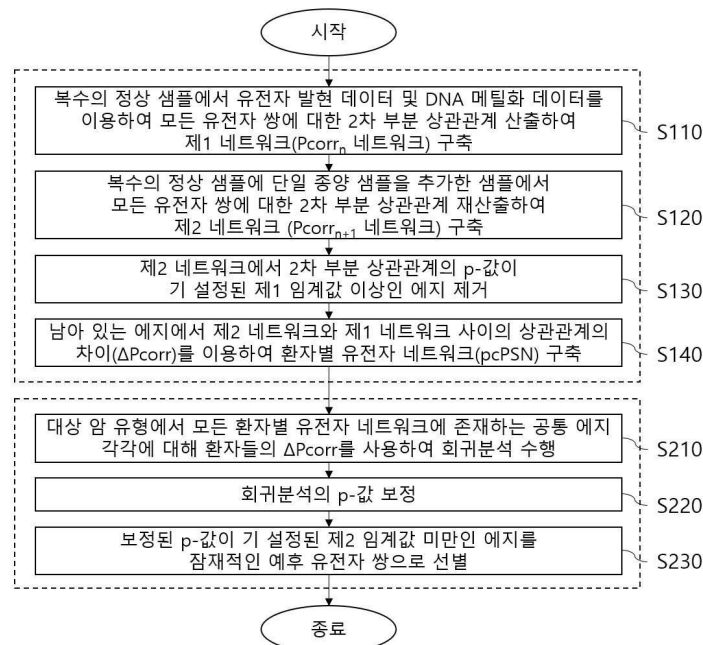
(54) 발명의 명칭 유전자 발현과 DNA 메틸화의 2차 부분 상관관계에 기초한 암 환자별 네트워크 구축 방법

(57) 요약

본 발명은 유전자 발현과 DNA 메틸화의 2차 부분 상관관계에 기초한 암 환자별 네트워크 구축 방법에 관한 것이다. 본 발명에 따르면, 복수의 정상 샘플에서 유전자 발현 데이터 및 DNA 메틸화 데이터를 이용하여 모든 유전자 쌍에 대한 2차 부분 상관관계 산출하여 제1 네트워크(P_{corr_n} 네트워크) 구축

(뒷면에 계속)

대표도 - 도1



플을 추가한 샘플에서 모든 유전자 쌍에 대한 2차 부분 상관관계를 재산출하여 제2 네트워크를 구축하는 단계; 상기 제2 네트워크에서 2차 부분 상관관계의 p-값이 기 설정된 제1 임계값 이상인 에지를 제거하는 단계; 및 남아 있는 에지에서 상기 제2 네트워크와 상기 제1 네트워크 사이의 2차 부분 상관관계의 차이를 이용하여 환자별 유전자 네트워크를 구축하는 단계를 포함한다.

이와 같이 본 발명에 따른 2차 부분 상관관계 기반의 네트워크는 다양한 유형의 암을 분류하고 잠재적인 예후 유전자 쌍을 도출하는데 매우 효과적이다. 특히, 개별 유전자 분석이나 유전자 발현 데이터 분석으로는 찾을 수 없는 새로운 암 관련 유전자 또는 유전자 쌍을 발굴할 수 있다. 또한, 네트워크에서 도출된 잠재적인 예후 유전자 쌍의 정보를 기초로 암 환자의 생존 시간을 예측할 수 있다.

(52) CPC특허분류

G16B 20/00 (2019.02)
G16B 25/10 (2019.02)
G16H 10/60 (2021.08)
G16H 50/20 (2018.01)
G16H 50/50 (2018.01)
C12Q 2600/156 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711128786
과제번호	2020R1A2B5B01096299
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)(R&D)
연구과제명	바이오 빅데이터로부터 개인별 분자관계 네트워크 추론과 암의 조기 진단과 예후 예측을 위한 네트워크 바이오마커 발굴
기 여 율	1/1
과제수행기관명	인하대학교
연구기간	2021.03.01 ~ 2022.02.28

명세서

청구범위

청구항 1

복수의 정상 샘플에서 유전자 발현 데이터 및 DNA 메틸화 데이터를 이용하여 모든 유전자 쌍에 대한 2차 부분 상관관계를 산출하여 제1 네트워크를 구축하는 단계;

상기 복수의 정상 샘플에 단일 종양 샘플을 추가한 샘플에서 모든 유전자 쌍에 대한 2차 부분 상관관계를 재산출하여 제2 네트워크를 구축하는 단계;

상기 제2 네트워크에서 2차 부분 상관관계의 p-값이 기 설정된 제1 임계값 이상인 에지를 제거하는 단계; 및

남아 있는 에지에서 상기 제2 네트워크와 상기 제1 네트워크 사이의 2차 부분 상관관계의 차이를 이용하여 환자별 유전자 네트워크를 구축하는 단계를 포함하는 유전자 발현과 DNA 메틸화의 2차 부분 상관관계에 기초한 암 환자별 네트워크 구축 방법.

청구항 2

제 1 항에 있어서,

상기 2차 부분 상관관계는 하기의 수학적식

$$Pcorr(g_i, g_j | m_i, m_j) = \rho_{g_i g_j . m_i m_j} \\ = \frac{\rho_{g_i g_j . m_i} - \rho_{g_i m_j . m_i} \cdot \rho_{g_j m_j . m_i}}{\sqrt{(1 - \rho_{g_i m_j . m_i}^2)(1 - \rho_{g_j m_j . m_i}^2)}}$$

에 따라 산출하며,

상기 $Pcorr(g_i, g_j | m_i, m_j)$ 는 유전자 i 및 j 사이의 2차 부분 상관관계를 나타내고, g_i, g_j 는 각각 유전자 i 및 j의 발현 수준을 나타내고, m_i, m_j 는 각각 유전자 i 및 j의 메틸화 수준을 나타내며, ρ 는 피어슨 상관관계를 나타내는 유전자 발현과 DNA 메틸화의 2차 부분 상관관계에 기초한 암 환자별 네트워크 구축 방법.

청구항 3

제 2 항에 있어서,

상기 2차 부분 상관관계의 p-값은 하기의 수학적식으로 정의되는 t-값과 $(n - k - 2)$ 자유도를 갖는 t-분포 함수를 사용하여 산출되며,

$$t = \rho_{g_i g_j . m_i m_j} \sqrt{\frac{n - k - 2}{1 - \rho_{g_i g_j . m_i m_j}^2}}$$

상기 n은 샘플의 수를 나타내고, k는 상관관계의 차수를 나타내는 유전자 발현과 DNA 메틸화의 2차 부분 상관관계에 기초한 암 환자별 네트워크 구축 방법.

청구항 4

제 1 항에 있어서,

대상 암 유형에서 모든 환자별 유전자 네트워크에 존재하는 공통 예지 각각에 대해 환자들의 2차 부분 상관관계의 차이를 사용하여 회귀분석을 수행하는 단계;

상기 회귀분석의 p-값을 보정하는 단계; 및

보정된 p-값이 기 설정된 제2 임계값 미만인 예지를 잠재적인 예후 유전자 쌍으로 선별하는 단계를 더 포함하는 유전자 발현과 DNA 메틸화의 2차 부분 상관관계에 기초한 암 환자별 네트워크 구축 방법.

청구항 5

제 4 항에 있어서,

상기 회귀분석을 수행하는 단계는 상기 공통 예지의 각 유전자 쌍의 2차 부분 상관관계의 차이로 콕스 비례 위험 모델을 실행하는 유전자 발현과 DNA 메틸화의 2차 부분 상관관계에 기초한 암 환자별 네트워크 구축 방법.

발명의 설명

기술 분야

[0001] 본 발명은 유전자 발현과 DNA 메틸화의 2차 부분 상관관계에 기초한 암 환자별 네트워크 구축 방법에 관한 것으로, 더욱 상세하게 설명하면 암 환자 샘플의 유전자 발현 데이터와 DNA 메틸화 데이터를 모두 사용하여 암 환자별 유전자 상관관계 네트워크를 구축하고 암 환자의 예후를 예측할 수 있는 유전자 쌍을 발굴할 수 있는 유전자 발현과 DNA 메틸화의 2차 부분 상관관계에 기초한 암 환자별 네트워크 구축 방법에 관한 것이다.

배경 기술

[0002] 대부분의 인간 암은 많은 유전적 또는 후성적 메커니즘에 의해 발생하는 체세포 돌연변이의 결과로 발생한다. 유전자 돌연변이와 대조적으로, 후성적 유전자 돌연변이는 DNA 구조나 유전자 활성의 변화를 의미하며, 이것은 DNA 배열 자체의 변화와는 관련이 없다. 암은 복잡하고 이질적인 질병이기 때문에, 많은 치료법들은 종양 세포의 발달을 돕는 특정 유전자 또는 후성유전적 변화를 가진 환자들에게만 효과적이다. 따라서, 개별 환자에 대해 특정 유전자 또는 후성적 유전자 돌연변이를 식별하는 것은 암 환자에 대한 효과적인 치료법을 선택하는데 필수적이다.

[0003] 기술의 발달로 암 환자의 다량의 유전자 및 후성유전학적 데이터를 이용할 수 있게 되었다. 다량의 데이터의 가용성은 암 관련 유전자를 식별하거나 종양 샘플과 정상 샘플 사이의 유전자-유전자 간 상호작용 차이에서 샘플 특이적 유전자 상관관계 네트워크를 구축하기 위한 여러 컴퓨터 연구를 촉진하였다.

[0004] 샘플 특이적 유전자 상관관계 네트워크는 환자 맞춤형 치료법을 찾거나 환자의 예후를 예측하는 데 도움이 되기 때문에 환자 샘플에 고유한 유전자 간 관계를 식별하는데 유용하다. 샘플 특이적 유전자 상관관계 네트워크는 일반적으로 복수의 대조군 샘플에 대한 기준 네트워크와 새로운 샘플에 의해 교란된 네트워크 간의 차이를 계산함으로써 구축된다.

[0005] 샘플 특이적 유전자 상관관계 네트워크에 대한 종래의 연구는 단일 유전자 또는 단일 유전자들의 집합보다 유전자 상관관계가 암 환자의 예후를 예측하는데 더 안정적이고 강력하다는 것을 보여주었다.

[0006] 그러나 종래의 연구에 의해 제안된 대부분의 샘플 특이적 유전자 상관관계 네트워크는 후성적 유전자 데이터를 고려하지 않고 유전자 발현 데이터만을 기반으로 구축된다. 이에 종래 기술에 의해 구축된 샘플 특이적 유전자 상관관계 네트워크는 직접적인 유전자 상호작용과 DNA 메틸화 등에 의한 간접적인 상호작용을 구분할 수 없다는 한계가 있다.

발명의 내용

해결하려는 과제

[0007] 이와 같이 본 발명에 따르면, 암 환자 샘플의 유전자 발현 데이터와 DNA 메틸화 데이터를 모두 사용하여 암 환자별 유전자 상관관계 네트워크를 구축하고 암 환자의 예후를 예측할 수 있는 유전자 쌍을 발굴할 수 있는 유전자 발현과 DNA 메틸화의 2차 부분 상관관계에 기초한 암 환자별 네트워크 구축 방법을 제공하기 위한 것이다.

과제의 해결 수단

- [0008] 이러한 기술적 과제를 이루기 위한 본 발명의 실시예에 따르면 유전자 발현과 DNA 메틸화의 2차 부분 상관관계에 기초한 암 환자별 네트워크 구축 방법에 있어서, 복수의 정상 샘플에서 유전자 발현 데이터 및 DNA 메틸화 데이터를 이용하여 모든 유전자 쌍에 대한 2차 부분 상관관계를 산출하여 제1 네트워크를 구축하는 단계; 상기 복수의 정상 샘플에 단일 종양 샘플을 추가한 샘플에서 모든 유전자 쌍에 대한 2차 부분 상관관계를 재산출하여 제2 네트워크를 구축하는 단계; 상기 제2 네트워크에서 2차 부분 상관관계의 p-값이 기 설정된 제1 임계값 이상인 에지를 제거하는 단계; 및 남아 있는 에지에서 상기 제2 네트워크와 상기 제1 네트워크 사이의 2차 부분 상관관계의 차이를 이용하여 환자별 유전자 네트워크를 구축하는 단계를 포함한다.
- [0009] 상기 2차 부분 상관관계는 하기의 수학적식

$$Pcorr(g_i, g_j | m_i, m_j) = \rho_{g_i g_j . m_i m_j} \\ = \frac{\rho_{g_i g_j . m_i} - \rho_{g_i m_j . m_i} \cdot \rho_{g_j m_j . m_i}}{\sqrt{(1 - \rho_{g_i m_j . m_i}^2)(1 - \rho_{g_j m_j . m_i}^2)}}$$

- [0010]
- [0011] 에 따라 산출하며, 상기 $Pcorr(g_i, g_j | m_i, m_j)$ 는 유전자 i 및 j 사이의 2차 부분 상관관계를 나타내고, g_i, g_j 는 각각 유전자 i 및 j의 발현 수준을 나타내고, m_i, m_j 는 각각 유전자 i 및 j의 메틸화 수준을 나타내며, ρ 는 피어슨 상관관계를 나타낼 수 있다.
- [0012] 또한, 상기 2차 부분 상관관계의 p-값은 하기의 수학적식으로 정의되는 t-값과 $(n - k - 2)$ 자유도를 갖는 t-분포 함수를 사용하여 산출되며,

$$t = \rho_{g_i g_j . m_i m_j} \sqrt{\frac{n - k - 2}{1 - \rho_{g_i g_j . m_i m_j}^2}}$$

- [0013]
- [0014] 상기 n은 샘플의 수를 나타내고, k는 상관관계의 차수를 나타낼 수 있다.
- [0015] 한편, 본 발명의 실시예에 따른 유전자 발현과 DNA 메틸화의 2차 부분 상관관계에 기초한 암 환자별 네트워크 구축 방법은, 대상 암 유형에서 모든 환자별 유전자 네트워크에 존재하는 공통 에지 각각에 대해 환자들의 2차 부분 상관관계의 차이를 사용하여 회귀분석을 수행하는 단계; 상기 회귀분석의 p-값을 보정하는 단계; 및 보정된 p-값이 기 설정된 제2 임계값 미만인 에지를 잠재적인 예후 유전자 쌍으로 선별하는 단계를 더 포함할 수 있다.
- [0016] 상기 회귀분석을 수행하는 단계는 상기 공통 에지의 각 유전자 쌍의 2차 부분 상관관계의 차이로 콕스 비례 위험 모델을 실행할 수 있다.

발명의 효과

- [0017] 이와 같이 본 발명에 따른 2차 부분 상관관계 기반의 네트워크는 다양한 유형의 암을 분류하고 잠재적인 예후 유전자 쌍을 도출하는데 매우 효과적이다. 특히, 개별 유전자 분석이나 유전자 발현 데이터 분석으로는 찾을 수 없는 새로운 암 관련 유전자 또는 유전자 쌍을 발굴할 수 있다. 또한, 네트워크에서 도출된 잠재적인 예후 유전자 쌍의 정보를 기초로 암 환자의 생존 시간을 예측할 수 있다.

도면의 간단한 설명

- [0018] 도 1은 본 발명의 일 실시예에 따른 유전자 발현과 DNA 메틸화의 2차 부분 상관관계에 기초한 암 환자별 네트워

크 구축 방법의 흐름도이다.

도 2는 암 환자별 부분 상관관계를 기반으로 암 환자별 유전자 상관관계 네트워크를 구축하는 과정을 개략적으로 도시하는 도면이다.

도 3은 본 발명의 일 실시예에 따른 유전자 발현과 DNA 메틸화의 2차 부분 상관관계에 기초한 암 환자별 네트워크 구축 방법의 실험 과정을 개략적으로 도시한 도면이다.

도 4는 본 발명의 실험에 따라 Kaplan-Meier 플롯을 사용하여 시각화한 두 그룹의 생존율을 도시하는 도면이다.

도 5는 비교 실험에 따라 잠재적인 예후 유전자 쌍에 관련된 모든 유전자를 대상으로 로그 순위 테스트를 실시한 결과를 도시하는 도면이다.

도 6은 2명의 KIRC 환자의 pcPSN에서 추출된 서브네트워크를 비교하여 도시하는 도면이다.

도 7은 10가지 암 유형의 4,321개 pcPSN의 계층적 클러스터링 결과를 도시하는 도면이다.

도 8은 BRCA 환자의 5년 생존 예측에서 각 분류 모델의 성능을 도시하는 도면이다.

도 9는 LIRI 데이터 세트의 ROC곡선과 ROC 곡선 아래 면적을 도시하는 도면이다.

도 10은 간암에 대한 두 개의 예후 유전자 쌍에 대한 두 개의 데이터 세트에서 간암 환자의 생존율을 도시하는 도면이다.

발명을 실시하기 위한 구체적인 내용

[0019] 이하 첨부된 도면을 참조하여 본 발명에 따른 바람직한 실시예를 상세히 설명하기로 한다. 이 과정에서 도면에 도시된 선들의 두께나 구성요소의 크기 등은 설명의 명료성과 편의상 과장되게 도시되어 있을 수 있다.

[0020] 또한 후술되는 용어들은 본 발명에서의 기능을 고려하여 정의된 용어들로서, 이는 사용자, 운용자의 의도 또는 관례에 따라 달라질 수 있다. 그러므로 이러한 용어들에 대한 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다.

[0021] 도 1은 본 발명의 일 실시예에 따른 유전자 발현과 DNA 메틸화의 2차 부분 상관관계에 기초한 암 환자별 네트워크 구축 방법의 흐름도이다.

[0022] 도 1을 참조하면, 우선 복수의 정상 샘플에서 유전자 발현 데이터 및 DNA 메틸화 데이터를 이용하여 모든 유전자 쌍에 대한 2차 부분 상관관계(2nd-order partial correlation coefficient; Pcorr)를 산출하여 제1 네트워크(Pcorr_n 네트워크)를 구축할 수 있다(S110).

[0023] 일 실시예에 따르면, 유전자 발현 데이터로서 RNA-seq 유전자 발현 데이터를 이용할 수 있고, 후성 유전자 데이터 중에서는 후성적 메커니즘이 가장 일반적이고 널리 연구되어 많은 양의 데이터를 이용할 수 있는 DNA 메틸화 데이터를 이용할 수 있다. 여기서, 포유류 게놈의 DNA 메틸화는 시토신의 C5 위치에 메틸기를 추가하는 것을 의미하며, 유전자 억제에 관여하는 단백질을 모집하거나 전사인자가 DNA에 결합하는 것을 억제함으로써 유전자 발현을 조절한다.

[0024] 또한, 유전자 상관관계 네트워크에서 유전자 쌍에 대한 상관관계를 산출하기 위해 피어슨 상관관계가 널리 사용된다. 그러나, 피어슨 상관관계는 교란 요인이 두 변수에 미치는 영향을 고려하지 않은 두 변수 간의 상관관계를 나타내는 것으로, 0차 상관관계이다. 이와 달리 부분 상관관계는 두 변수에 대한 다른 변수의 영향을 제거한 후의 두 변수 간의 직접적인 상관관계를 나타낸다. 하기의 수학적 식 1에서 정의된 1차 부분 상관관계

$\rho_{xy.z}$ 는 다른 변수 z를 제어하면서 x와 y 사이의 상관관계를 측정한다. 여기서, ρ_{xy} 는 x와 y 사이의 피어슨 상관관계를 나타낸다.

수학식 1

$$\rho_{xy.z} = \frac{\rho_{xy} - \rho_{xz} \cdot \rho_{yz}}{\sqrt{(1 - \rho_{xz}^2)(1 - \rho_{yz}^2)}}$$

[0025]

[0026]

본 발명에서 이용하는 2차 부분 상관관계 $\rho_{xy.zw}$ 는 다른 두 변수 z 와 w 를 제어하면서 x 와 y 사이의 상관관계를 측정한다.

수학식 2

$$\rho_{xy.zw} = \frac{\rho_{xy.z} - \rho_{xw.z} \cdot \rho_{yw.z}}{\sqrt{(1 - \rho_{xw.z}^2)(1 - \rho_{yw.z}^2)}}$$

[0027]

[0028]

본 발명의 실시예에서는 하기의 수학식 3에 따라 각 유전자의 메틸화를 제어한 후 각 유전자 쌍에 대한 2차 부분 상관관계 $Pcorr$ 를 산출할 수 있다. 여기서, g_i , g_j 는 2개의 유전자의 발현 수준을 나타내고 m_i , m_j 는 이들의 메틸화 수준을 나타낸다.

수학식 3

$$\begin{aligned} Pcorr(g_i, g_j | m_i, m_j) &= \rho_{g_i g_j . m_i m_j} \\ &= \frac{\rho_{g_i g_j . m_i} - \rho_{g_i m_j . m_i} \cdot \rho_{g_j m_j . m_i}}{\sqrt{(1 - \rho_{g_i m_j . m_i}^2)(1 - \rho_{g_j m_j . m_i}^2)}} \end{aligned}$$

[0029]

[0030]

이후, 복수의 정상 샘플에 단일 종양 샘플을 추가한 샘플에서 모든 유전자 쌍에 대한 2차 부분 상관관계를 계산 출하여 제2 네트워크($Pcorr_{n+1}$ 네트워크)를 구축할 수 있다(S120). 이는 S110 단계에서 상술한 바와 동일한 방식으로 수행될 수 있는 바, 이에 대한 중복적인 설명은 생략한다.

[0031]

이후, 제2 네트워크에서 2차 부분 상관관계의 p -값이 기 설정된 제1 임계값 이상인 에지를 제거할 수 있다(S130).

[0032]

일 실시예에 따르면, 하기의 수학식 4로 정의되는 t -값과 $(n - k - 2)$ 자유도를 갖는 t -분포 함수를 사용하여 2차 부분 상관관계 p -값을 산출할 수 있다. 여기서, n 은 샘플의 수를 나타내고, k 는 상관관계의 차수를 나타내는 것으로, 본 발명에서는 2차 부분 상관관계이므로 $k = 2$ 이다.

수학식 4

$$t = \rho_{g_i g_j . m_i m_j} \sqrt{\frac{n - k - 2}{1 - \rho_{g_i g_j . m_i m_j}^2}}$$

[0033]

[0034]

또한, 제1 임계값은 0.05로 설정될 수 있으며, 이에 따라 2차 부분 상관관계의 p-값이 0.05 이상인 에지들을 제거할 수 있다.

[0035]

이후, 남아 있는 에지에서 제2 네트워크와 제1 네트워크 사이의 상관관계의 차이($\Delta Pcorr$)를 이용하여 환자별 유전자 네트워크(partial correlation-based patient-specific network; pcPSN)를 구축할 수 있다(S140). pcPSN에는 제2 네트워크에서 2차 부분 상관관계의 p-값이 제1 임계값 보다 작은 에지들이 포함될 수 있으며, pcPSN의 각 에지의 값은 $\Delta Pcorr$ 을 나타낸다.

수학식 5

$$\Delta Pcorr = |Pcorr_{n+1} - Pcorr_n|$$

[0036]

[0037]

도 2는 암 환자별 부분 상관관계를 기반으로 암 환자별 유전자 상관관계 네트워크를 구축하는 과정을 개략적으로 도시하는 도면이다.

[0038]

도 2를 참조하면, 위에서 구체적으로 설명한 바와 같이, n개의 정상 샘플에서 유전자 발현과 메틸화의 2차 부분 상관관계($Pcorr_n$)를 기반으로 기준 네트워크인 제1 네트워크를 구축하고, 그 다음 n개의 정상 샘플에 단일 종양 샘플을 추가하고 p-값이 0.05 미만인 2차 부분 상관관계($Pcorr_{n+1}$)을 선택하여 교란 네트워크인 제2 네트워크를 구축할 수 있다. 이후, 교란 네트워크에서 기준 네트워크를 감산하여 환자별 유전자 네트워크 pcPSN를 구축할 수 있다. 여기서, 상관관계의 차이, 즉 $\Delta Pcorr$ 는 전적으로 추가된 단일 종양 샘플에 기인하는 것이므로, 종양 샘플마다 상이한 $\Delta Pcorr$ 를 나타내며, 각 종양 샘플을 샘플 고유의 네트워크로 나타낼 수 있다.

[0039]

상술한 과정에 따라 환자별 유전자 네트워크가 구축되고 나면, 이로부터 암 환자의 예후를 예측할 수 있는 잠재적인 예후 유전자 쌍을 발굴하기 위해 후술하는 과정을 수행할 수 있다.

[0040]

먼저, 대상 암 유형에서 모든 환자별 유전자 네트워크에 존재하는 공통 에지(즉, 유전자 쌍) 각각에 대해 환자들의 $\Delta Pcorr$ 를 사용하여 회귀분석을 수행할 수 있다(S210). 일 실시예에 따르면, 콕스회귀분석(Cox regression analysis)을 수행할 수 있으며, 각 유전자 쌍의 $\Delta Pcorr$ 로 콕스 비례 위험 모델을 실행할 수 있다.

[0041]

이후, 회귀분석의 p-값을 보정할 수 있다(S220). 일 실시예에 따르면, 콕스 비례 위험 모델의 p-값은 다중 비교 문제에서 발생할 수 있는 거짓 양수 비율(false positive rate, FDR)을 제어하기 위해 공지된 Benjamini-Hochberg 방법에 따라 보정될 수 있다.

[0042]

이후, 보정된 p-값이 기 설정된 제2 임계값 미만인 에지를 잠재적인 예후 유전자 쌍으로 선별할 수 있다(S230). 일 실시예에 따르면, 제2 임계값은 0.05로 설정될 수 있으며, 이에 따라 보정된 p-값이 0.05 미만인 에지의 유전자 쌍을 대상 암 유형에 대한 잠재적인 예후 유전자 쌍으로 선별할 수 있다.

[0043]

도 3은 본 발명의 일 실시예에 따른 유전자 발현과 DNA 메틸화의 2차 부분 상관관계에 기초한 암 환자별 네트워크 구축 방법의 실험 과정을 개략적으로 도시한 도면이다.

[0044]

도 3을 참조하면, 먼저 유전자 발현 데이터와 DNA 메틸화 데이터는 모두 GDAC Firehose의 TCGA(The Cancer

Genome Atlas) 데이터 세트에서 획득하였다. TCGA 데이터 세트에서 유전자 발현 데이터, DNA 메틸화 데이터 및 10개 이상의 고형 정상 조직이 있는 방광암(bladder urothelial carcinoma, BLCA), 유방암(breast invasive carcinoma, BRCA), 결장암(colon adenocarcinoma, COAD), 두경부암(head and neck squamous cell carcinoma, HNSC), 투명세포형 신세포암(kidney renal clear cell carcinoma, KIRC), 유두상 신세포암(kidney renal papillary cell carcinoma, KIRP), 간암(liver hepatocellular carcinoma, LIHC), 폐선암(lung adenocarcinoma, LUAD), 전립선암(prostate adenocarcinoma, PRAD) 및 자궁내막암(uterine corpus endometrial carcinoma, UCEC)을 포함하는 10가지 유형의 암을 선택하였다. 각 암 유형에 대해 기준 네트워크의 정상 샘플을 위해서 TCGA에서 환자의 종양 인접 조직 샘플을 수집하였다.

[0045] 유전자 발현 데이터는 RSEM(RNA-Seq by Expectation Maximization)을 사용하여 정규화되었으며 log2 스케일로 변환되었다. 전체 종양 샘플 중에서 원발 종양 샘플이 선택되었다. 각 암 유형에 대해 MalaCards 데이터베이스에서 암 관련 유전자를 얻었다.

[0046] 이후, 도 1을 참조하여 상술한 바에 따라 환자별 유전자 네트워크를 구축하였다. 또한, 본 발명과의 성능 비교를 위해, 환자별 0차 유전자 상관관계 네트워크(cPSN)도 구축하였다. cPSN을 구축하는 방법은 pcPSN과 유사하지만, cPSN과 pcPSN의 차이는 다음과 같다: (1) pcPSN은 유전자 발현 데이터와 DNA 메틸화 데이터를 모두 사용하여 구축되는 반면, cPSN은 유전자 발현 데이터만으로 구축된다. (2) pcPSN은 2차 부분 상관관계를 사용하여 구축되지만, cPSN은 모든 유전자 쌍에 대해 0차 상관관계(즉, 피어슨 상관관계수, PCC)를 사용하여 구축되었다. 따라서 cPSN의 모든 에지는 하기의 수학적 식 6과 같이 환자 샘플과 정상 샘플 사이의 두 유전자의 PCC 차이(즉, $\Delta PCC(g_i, g_j)$)를 나타낸다.

수학적 식 6

$$\Delta PCC(g_i, g_j) = |PCC_{n+1}(g_i, g_j) - PCC_n(g_i, g_j)|$$

[0047]

[0048] 이후, 도 1을 참조하여 상술한 바에 따라 각 암 유형에 대해 잠재적인 예후 유전자 쌍을 발굴하였다.

[0049] 이후, 도출된 잠재적인 예후 유전자 쌍에 대해 암 환자를 다음과 같이 두 그룹으로 분류하였다.

[0050] 고 ΔPcorr 그룹: '유전자 쌍의 ΔPcorr ≥ 동일한 암 유형의 모든 환자의 대응하는 유전자 쌍의 평균 ΔPcorr'인 환자

[0051] 저 ΔPcorr 그룹: '유전자 쌍의 ΔPcorr < 동일한 암 유형의 모든 환자의 대응하는 유전자 쌍의 평균 ΔPcorr'인 환자

[0052] 이후, 로그 순위 테스트(log-rank test)에 의해 두 그룹의 전체 생존율을 분석하고, Kaplan-Meier 플롯을 사용하여 결과를 시각화하였다.

[0053] pcPSN은 ΔPcorr 값에 상관없이 Pcorr_{n+1}의 p-값이 < 0.05인 모든 유전자 쌍을 포함한다. 위 두 그룹(고 ΔPcorr 그룹과 저 ΔPcorr 그룹)의 pcPSN을 비교하기 위해 pcPSN에서 서브네트워크를 추출하였다. 서브네트워크는 유전자 쌍의 2개의 유전자와 그 인접 유전자로 구성되며, 서브네트워크에서 유전자를 연결하는 에지는 ΔPcorr ≥ 모든 종양 샘플의 평균 ΔPcorr인 에지로 제한된다. 두 그룹의 서브 네트워크의 토폴로지 특성을 분석하였다.

[0054] 또한, 그룹 간 유전자의 기능적 차이를 비교하기 위해 유전자 온톨로지 분석을 수행하였다. 각 그룹의 서브네트워크에서 가장 빈번한 100개의 유전자를 이용하여 주석, 시각화 및 통합 발견 데이터베이스(Database for Annotation, Visualization and Integrated Discovery, DAVID)의 기능 주석 클러스터링을 수행하였다. 지정된 유전자에 대해 DAVID 기능 주석 클러스터링은 여러 데이터베이스의 주석을 통합하고 각 클러스터에 대한 결과를 제공한다.

[0055] 한편, 다양한 암 유형의 pcPSN을 클러스터링하여 서로 다른 암 유형 간에 pcPSN의 유사점 또는 차이점이 있는지 분석하였다. 서로 다른 유형의 암의 pcPSN을 공정하게 비교하기 위해서는, 동일한 유전자 집합에서 생성되어야

한다. 따라서 MalaCards에서 획득한 모든 암 관련 유전자를 10가지 유형의 암에 대해 통합하고 단일 유전자 세트를 만든 후에 동일한 유전자 세트를 사용하여 4,321개의 종양 샘플에 대해 pcPSNs를 구성하였다.

[0056] 네트워크 비교를 위해 하기의 수학적식 7과 같은 Jaccard 지수를 서로 다른 pcPSN 사이의 거리 측정값으로 사용하였다.

수학적식 7

$$J(N_i, N_j) = \frac{|N_i \cap N_j|}{|N_i \cup N_j|} = \frac{|N_i \cap N_j|}{|N_i| + |N_j| - |N_i \cap N_j|}$$

[0057]

[0058] 여기서, N_i 및 N_j 는 각각 샘플 i 및 j 의 네트워크에 포함되는 에지 세트를 나타낸다. Jaccard 지수는 모든 에지에 대한 공통 에지의 비율이므로 $[0, 1]$ 범위의 값을 가진다.

[0059] 한편, pcPSN과 cPSN에서 도출한 잠재적인 예후 유전자 쌍은 1~5년내 개별 암 환자의 생존 또는 사망을 예측하는 분류 모델을 개발하는데 사용되었다. 본 실험에서는 로지스틱 회귀(logistic regression), 균형 랜덤 포레스트(balanced random forest) 및 방사형 기저 함수 커널(radial basis function kernel)을 사용하는 서포트 벡터 머신(support vector machine, SVM)의 세 가지 유형의 분류 모델을 구축하였다. 또한, 암의 유형별로 총 15개의 개별 모델(1년, 2년, 3년, 4년, 5년 생존 예측)을 구축하였다. 각 모델에서 사망자와 생존자는 각각 양성 및 음성 사례로 간주되었다. 또한, 기간 내 관측 중단 데이터는 제외되었다. 예를 들어, 5년 생존 예측에 대한 분류 모델에서 생존 기간이 5년 미만인 관측 중단 환자는 제외되었다.

[0060] 특정 유형의 암에 대한 유전자 쌍에서 m 개의 고유 유전자를 가진 n 개의 잠재적인 예후 유전자 쌍이 있다고 가정하면, 분류 모델에 의해 사용되는 특징은 (1) m 개의 유전자의 발현 수준, (2) n 개의 잠재적인 예후 유전자 쌍의 $\Delta Pcorr$ 또는 ΔPCC , (3) 예후 유전자 쌍의 서브네트워크에 있는 각 m 개의 유전자의 인접 유전자 수이다. 따라서 한 명의 환자는 $(2m + n)$ 개의 인자를 가진 특징 벡터로 표현될 수 있다.

[0061] 또한, 그리드 검색 알고리즘을 사용하여 모델의 파라미터 값을 결정하였다. 모델의 성능은 데이터 세트의 크기가 상대적으로 작기 때문에 LOOCV(Leave-One-Out Cross Validation)을 통해 평가되었다. 분류 모델을 검증하기 위해 모델을 훈련하는 데 사용되지 않은 독립 데이터 세트에서 모델을 테스트하였다. 독립 데이터는 ICGC 데이터 포털의 간세포암(LIRI_JP) 데이터 세트에서 획득하였다.

[0062] 상술한 바와 같은 실험을 통해, 10가지 유형의 암 환자 4,321명을 대상으로 pcPSNs와 cPSNs를 구축하였으며, pcPSNs에서 BRCA, KIRC, LIHC, LUAD, PRAD 및 UCEC에 대해 각각 총 279, 51, 132, 117, 73 및 23개의 잠재적인 예후 유전자 쌍을 도출하였다. 나머지 4가지 유형의 암은 유전자 쌍과 관련된 p -값이 의미 없기 때문에 예후 유전자 쌍이 발견되지 않았다.

[0063] 도 4는 본 발명의 실험에 따라 Kaplan-Meier 플롯을 사용하여 시각화한 두 그룹의 생존율을 도시하는 도면으로, 6가지의 암 유형에서 로그 순위 테스트의 p -값이 가장 낮은 잠재적인 예후 유전자 쌍에 대한 두 환자 그룹의 전체 생존율을 나타낸다.

[0064] 또한, 유전자 쌍의 예후력과 해당 유전자 쌍의 단일 유전자의 예후력을 비교하기 위해 잠재적인 예후 유전자 쌍에 관련된 모든 유전자를 대상으로 로그 순위 테스트를 실시하였다.

[0065] 도 5는 비교 실험에 따라 잠재적인 예후 유전자 쌍에 관련된 모든 유전자를 대상으로 로그 순위 테스트를 실시한 결과를 도시하는 도면이다.

[0066] 도 5를 참조하면, KDM5B_TRADD 유전자 쌍의 개별 유전자 KDM5B와 TRADD의 유전자 발현 수준과 메틸화 수준은 모두 예후력이 없는 반면, KDM5B_TRADD 유전자 쌍은 BRCA에서 매우 낮은 p -값($5.80E-08$)을 보임을 알 수 있다.

[0067] 표 1은 BRCA에서 10개의 잠재적인 예후 유전자 쌍과 해당 유전자 쌍에 관련된 개별 유전자의 유전자 발현 수준과 메틸화 수준에 대한 로그 순위 테스트의 p -값을 나타낸다. 유전자 쌍과는 달리, 개별 유전자의 유전자 발현 수준과 메틸화 수준 모두 로그 순위 테스트에서 유의하지 않을 정도로 높은 p -값을 가짐을 알 수 있다.

표 1

gene pair of psPSN		gene expression		DNA methylation	
gene1_gene2	p-value	gene1	gene2	gene1	gene2
KDM5B_TRADD	5.80E-08	-	-	-	-
EFNA3_STC2	3.47E-07	-	2.92E-03	1.41E-03	-
GSN_ZEB1	6.43E-07	-	-	-	-
ESR1_MUC1	1.72E-06	-	-	4.40E-04	-
CTCF_HSP90AA1	1.26E-05	-	1.27E-02	3.30E-02	-
CDKN2C_PIGR	4.78E-05	-	2.71E-03	-	-
VWF_ZEB1	5.74E-05	-	-	-	-
TGFBR3_ZEB1	6.00E-05	-	-	9.99E-03	-
CASP9_ETS1	6.31E-05	8.82E-03	-	1.79E-02	-
PML_TRADD	6.39E-05	-	-	-	-

[0068]

[0069]

도 6은 2명의 KIRC 환자의 pcPSN에서 추출된 서브네트워크를 비교하여 도시하는 도면으로, 도 6의 (a)는 예후 유전자 쌍 BIRC5_IL7의 고 ΔPcorr 그룹에 속하는 환자의 서브네트워크이고, 도 6의 (b)는 동일 예후 유전자 쌍의 저 ΔPcorr 그룹에 속하는 환자의 서브네트워크이다. 고 ΔPcorr 그룹의 환자(a)는 793일 만에 사망하였고, 저 ΔPcorr 그룹의 환자(b)는 2,087일 이상 생존하였다.

[0070]

도 6을 참조하면, (a)의 서브네트워크는 (b)의 서브네트워크에 비해 큰 것을 알 수 있다. 위에서 설명한 바와 같이 예후 유전자 쌍의 서브네트워크는 유전자 쌍의 2개의 유전자와 그 인접 유전자로 구성된다. 또한, 서브네트워크의 유전자를 연결하는 에지는 $\Delta Pcorr \geq$ 모든 종양 샘플의 평균 ΔPcorr인 에지로 제한된다. 여기서, ΔPcorr는 종양 샘플과 정상 샘플의 상관관계의 차이를 반영하므로, 정상 그룹과 차이가 큰 암 환자가 생존시간이 짧고 해당 암 환자의 서브네트워크가 더 크게 나타난다.

[0071]

표 2는 DAVID 기능 주석 분석에서 높은 점수를 받은 4개의 클러스터를 보여주는 것으로, BRCA의 KDM5B_TRADD 예후 유전자 쌍에 대해 두 환자 그룹을 나타낸다.

표 2

high ΔP_{corr}				
Annotation cluster 1	Enrichment score: 6.11	#genes	p-value	FDR
KEGG_PATHWAY	hsa05200:Pathways in cancer	19	1.55E-08	2.00E-06
KEGG_PATHWAY	hsa05215:Prostate cancer	9	1.63E-06	7.54E-05
KEGG_PATHWAY	hsa04151:PI3K-Akt signaling pathway	14	1.85E-05	4.57E-04
Annotation cluster 2	Enrichment score: 6.08			
UP_KEYWORDS	DNA damage	17	8.63E-12	3.34E-10
UP_KEYWORDS	DNA repair	15	1.00E-10	3.10E-09
GOTERM_BP_DIRECT	GO:0006281 DNA repair	10	7.96E-06	6.84E-04
Annotation cluster 3	Enrichment score: 5.62			
UP_KEYWORDS	Transcription regulation	32	4.54E-08	8.79E-07
UP_KEYWORDS	Transcription	32	8.62E-08	1.34E-06
UP_KEYWORDS	DNA-binding	26	6.45E-06	7.15E-05
Annotation cluster 4	Enrichment score: 5.00			
UP_KEYWORDS	Apoptosis	14	1.39E-06	1.79E-05
GOTERM_BP_DIRECT	GO:0006919 activation of cysteine-type endopeptidase activity involved in apoptotic process	7	7.89E-06	6.84E-04
GOTERM_BP_DIRECT	GO:0006915 apoptotic process	13	9.22E-05	6.34E-03
low ΔP_{corr}				
Annotation cluster 1	Enrichment score: 5.06			
UP_KEYWORDS	Kinase	19	1.03E-08	6.59E-07
UP_SEQ_FEATURE	active site:Proton acceptor	18	1.82E-08	5.19E-06
INTERPRO	IPR017441:Protein kinase, ATP binding site	14	6.29E-08	1.82E-05
Annotation cluster 2	Enrichment score: 5.04			
UP_SEQ_FEATURE	signal peptide	40	2.13E-08	5.19E-06
GOTERM_CC_DIRECT	GO:0005576 extracellular region	25	1.78E-06	3.18E-04
UP_KEYWORDS	Disulfide bond	36	4.04E-06	7.00E-05
Annotation cluster 3	Enrichment score: 4.35			
UP_KEYWORDS	Growth factor	10	1.27E-08	6.59E-07
GOTERM_MF_DIRECT	GO:0008083 growth factor activity	11	2.04E-08	2.98E-06
GOTERM_CC_DIRECT	GO:0031093 platelet alpha granule lumen	5	1.93E-04	4.93E-03
Annotation cluster 4	Enrichment score: 4.33			
UP_KEYWORDS	Kinase	14	1.03E-08	6.59E-07
INTERPRO	IPR017441:Protein kinase, ATP binding site	14	6.29E-08	1.82E-05
UP_SEQ_FEATURE	domain:Protein kinase	13	6.31E-08	1.03E-05

[0072]

[0073]

고 ΔP_{corr} 그룹의 주석 클러스터 1은 암과 관련된 경로를 포함한다. PI3K-Akt 신호 경로는 신호 전달, 아포토시스 및 세포 증식에 매우 중요한 역할을 하며, 암 발병에서 돌연변이가 발견되는 대표적인 경로 중 하나이다. 주석 클러스터 2는 DNA 손상/복구 기능의 주석을 포함하고, 주석 클러스터 4는 아포토시스와 관련된 주석을 나타낸다.

[0074]

반면, 저 ΔP_{corr} 그룹에 대해 높은 점수를 받은 클러스터는 단백질 합성, 효소 촉매 작용, 세포 활성화, 신호 전달, 세포 성장 또는 세포 분화에 대한 주석을 포함한다. 이러한 기능의 대부분은 암의 초기 단계에서 발견되는 것으로 알려져 있다.

[0075]

도 7은 10가지 암 유형의 4,321개 pcPSN의 계층적 클러스터링 결과를 도시하는 도면이다.

[0076]

도 7을 참조하면, 동일한 유형의 암의 모든 pcPSN은 다른 유형의 암의 다른 네트워크보다 서로 더 가깝고, 이에 따라 함께 군집화됨을 확인할 수 있다.

[0077]

서로 다른 암 유형에서 샘플의 배치 효과로 인해 완벽한 클러스터링이 발생했는지 여부를 확인하기 위해 유전자 발현 수준과 DNA 메틸화 수준을 이용한 추가 클러스터링을 수행하였다.

[0078]

1,615개의 MalaCards 유전자의 유전자 발현 수준과 DNA 메틸화 수준이 계층적 클러스터링 분석에서 pcPSN을 생성하는데 사용되었다. 클러스터링 정확도는 조정된 랜드 인덱스(adjusted rand index, ARI)에 의해 평가되었으며, ARI의 범위는 -1~1 이다. 여기서 0은 랜덤 라벨링과 비슷하며 1은 완벽한 일치를 나타낸다. 표 3은 클러스터링 결과에서 예측된 암 유형과 실제 암 유형 사이의 ARI를 보여준다.

표 3

feature	distance	linkage	adjusted rand index (ARI)	
			genes from Malacards	random genes
pcPSN	network	complete	1.000	0.977
		average	1.000	0.943
		single	1.000	0.931
gene expression + DNA methylation	euclidean	complete	0.441	0.342
		average	0.131	0.043
		single	0.000	0.000
	manhattan	complete	0.470	0.275
		average	0.131	0.042
		single	0.000	0.000
	cosine	complete	0.470	0.366
		average	0.129	0.043
		single	0.000	0.000
gene expression	euclidean	complete	0.421	0.355
		average	0.133	0.043
		single	0.000	0.000
	manhattan	complete	0.369	0.254
		average	0.133	0.043
		single	0.000	0.000
	cosine	complete	0.651	0.363
		average	0.129	0.043
		single	0.000	0.000
DNA methylation	euclidean	complete	0.093	0.086
		average	0.000	0.005
		single	0.000	0.000
	manhattan	complete	0.076	0.087
		average	0.003	0.009
		single	0.000	0.000
	cosine	complete	0.087	0.102
		average	0.000	0.000
		single	0.000	0.000

[0079]

[0080]

유전자 발현 및/또는 DNA 메틸화 수준을 이용한 클러스터링의 ARI는 pcPSN을 사용한 클러스터링의 ARI보다 훨씬 낮은 것을 알 수 있다. 이러한 결과는 배치 효과만이 완벽한 군집화를 위한 유일한 이유는 아니라는 것을 의미한다. 유전자 발현과 DNA 메틸화의 2차 부분 상관관계에 기초한 본 발명의 네트워크가 이러한 양호한 클러스터링 결과에 기여한 것을 알 수 있다.

[0081]

한편, 개발된 분류 모델의 성능은 하기의 수학적 식 8 내지 수학적 식 13으로 정의되는 민감도(sensitivity, SN), 특이도(specificity, SP), 양의 예측값(positive predictive value, PPV), 음의 예측값(negative predictive value, NPV), 정확도(accuracy, ACC) 및 매튜스 상관계수(Matthews correlation coefficient, MCC)의 6가지 지표를 통해 평가되었다. 여기서, MCC는 불균형 데이터 세트에서 다른 지표에 비해 모델의 성능을 보다 정확하게 평가할 수 있다. MCC는 정확한 벡터와 예측 벡터 사이의 피어슨 상관 계수이므로 [-1, 1] 범위의 값을 갖는다.

수학적 식 8

$$SN = \frac{TP}{TP + FN}$$

[0082]

수학식 9

$$SP = \frac{TN}{TN + FP}$$

[0083]

수학식 10

$$PPV = \frac{TP}{TP + FP}$$

[0084]

수학식 11

$$NPV = \frac{TN}{TN + FN}$$

[0085]

수학식 12

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}$$

[0086]

수학식 13

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

[0087]

[0088]

표 4는 BRCA에서의 SVM 모델의 성능을 보여준다.

표 4

year	dead	alive	TP	TN	FP	FN	sensitivity	specificity	PPV	NPV	accuracy	MCC
pcPSN												
1	15	670	0	667	3	15	0.000	0.996	0.000	0.978	0.974	-0.010
2	30	427	10	397	30	20	0.333	0.930	0.250	0.952	0.891	0.231
3	50	312	26	279	33	24	0.520	0.894	0.441	0.921	0.843	0.387
4	61	233	22	223	10	39	0.361	0.957	0.688	0.851	0.833	0.414
5	72	179	57	157	22	15	0.792	0.877	0.722	0.913	0.853	0.651
cPSN												
1	15	670	0	660	10	15	0.000	0.985	0.000	0.978	0.964	-0.018
2	30	427	18	370	57	12	0.600	0.867	0.240	0.969	0.849	0.312
3	50	312	30	279	33	20	0.600	0.894	0.476	0.933	0.854	0.450
4	61	233	36	198	35	25	0.590	0.850	0.507	0.888	0.796	0.417
5	72	179	45	153	26	27	0.625	0.855	0.634	0.850	0.789	0.482

[0089]

[0090]

SVM 모델은 1년 생존 예측에서 5년 생존 예측까지 해가 갈수록 더 나은 성능을 보였다. pcPSN의 5년 모델은 민감도 0.792, 특이도 0.877, PPV 0.722, NPV 0.913, 정확도 0.853 및 MCC 0.651을 달성하였다. 이에 반해, cPSN의 5년 모델은 민감도 0.625, 특이도 0.855, PPV 0.634, NPV 0.850, 정확도 0.789 및 MCC 0.482를 달성하였다. pcPSN 모델의 성능은 모든 지표에서 cPSN 모델의 성능보다 높은 것을 알 수 있다. 특히, 민감도 및 PPV에서 상대적으로 큰 개선을 보이는데, 이는 pcPSN 모델(FNR = 0.208)이 cPSN 모델(FNR = 0.37)보다 낮은 거짓 음성률(FNR)에 기인하는 것이다.

[0091]

도 8은 BRCA 환자의 5년 생존 예측에서 각 분류 모델의 성능을 도시하는 도면으로, (a)는 pcPSN을 사용한 각 분류 모델의 성능을 나타내고, (b)는 cPSN을 사용한 각 분류 모델의 성능을 나타낸다.

[0092]

도 8을 참조하면, SVM 모델은 cPSN의 민감도와 NPV를 제외한 모든 지표에서 최고의 성능을 보여준다. cPSN을 이용한 SVM 모델의 민감도 및 NPV가 낮은 것은 상술한 바와 같이 FNR이 높기 때문일 수 있다.

[0093]

독립 코호트에서 예측 모델의 검증을 위해 ICGC 데이터 포털에서 간세포암(LIRI) 데이터 세트를 수집하였다. LIRI 데이터 세트에는 DNA 메틸화 데이터가 포함되어 있지 않기 때문에 공정한 비교를 위해 cPSN과 함께 모델을 사용하여 독립 테스트를 수행하였다. LIRI 환자 중 5년 이상 생존한 환자는 2명뿐이어서 5년 생존 예측은 LIRI 데이터 세트로 하지 않았다. 도 9는 LIRI 데이터 세트의 ROC 곡선과 ROC 곡선 아래 면적을 도시하는 도면이다.

[0094]

테스트에서 LIHC 환자의 4년 생존 예측 모델은 민감도 0.884, 특이도 0.750, PPV 0.905, NPV 0.706, 정확도 0.847 및 MCC 0.622를 달성하였다. 독립 데이터 세트를 사용한 검증에서 LOOCV보다 약간 더 낮은 정확도(0.887 대 0.847)와 MCC(0.762 대 0.622)를 달성하였으나, 전체적인 성능은 여전히 높은 검증 수준을 유지하고 있다.

[0095]

예측 모델의 성능에 대한 각 특징의 기여도를 조사하기 위해 각 특징(즉, 유전자의 발현 수준, 예후 유전자 쌍의 ΔPCC 및 인접 유전자의 수)에 대해 별도의 SVM 모델을 구축하여 예측 모델의 성능에 대한 각 특징의 기여도를 평가하였다.

[0096]

표 5는 간암 환자의 4년 생존 예측에 대한 각 특징의 성능을 보여준다.

표 5

feature	SN	SP	PPV	NPV	ACC	MCC
LIHC leave-one-out cross validation						
expression	0.649	0.746	0.822	0.540	0.684	0.378
Δ PCC	0.868	0.603	0.798	0.717	0.774	0.493
#neighbors	0.693	0.857	0.898	0.607	0.751	0.527
total	1.000	0.683	0.851	1.000	0.887	0.762
LIRI_JP independent test						
expression	0.814	0.625	0.854	0.556	0.763	0.424
Δ PCC	0.953	0.375	0.804	0.750	0.797	0.427
#neighbors	0.814	0.750	0.897	0.600	0.797	0.530
total	0.884	0.750	0.905	0.706	0.847	0.622

[0097]

[0098]

표 5와 같이 유전자 쌍의 Δ PCC 및 인접 유전자 수가 유전자 발현 수준에 비해 성능이 높으며, 모든 특징을 사용한 경우 가장 높은 성능을 보였다. 암의 유형에 따라 결과가 다소 상이하나 유전자 상관관계(Δ PCC)와 인접 유전자 수는 유전자 발현 수준에 비해 환자의 생존율을 더 잘 예측함을 알 수 있다.

[0099]

독립 코호트에서 잠재적 예후 유전자 쌍의 검증을 위해 ICGC 데이터 세트에서 간암 환자의 생존 시간을 분석하였다.

[0100]

도 10은 간암에 대한 두 개의 예후 유전자 쌍에 대한 두 개의 데이터 세트에서 간암 환자의 생존율을 도시하는 도면으로, 생존 분석에 사용된 예측 유전자 쌍은 원래 TCGA LIHC 데이터 세트에서 도출되었으며 ICGC LIRI 데이터 세트에서도 유사한 예측력을 보임을 확인하였다.

[0101]

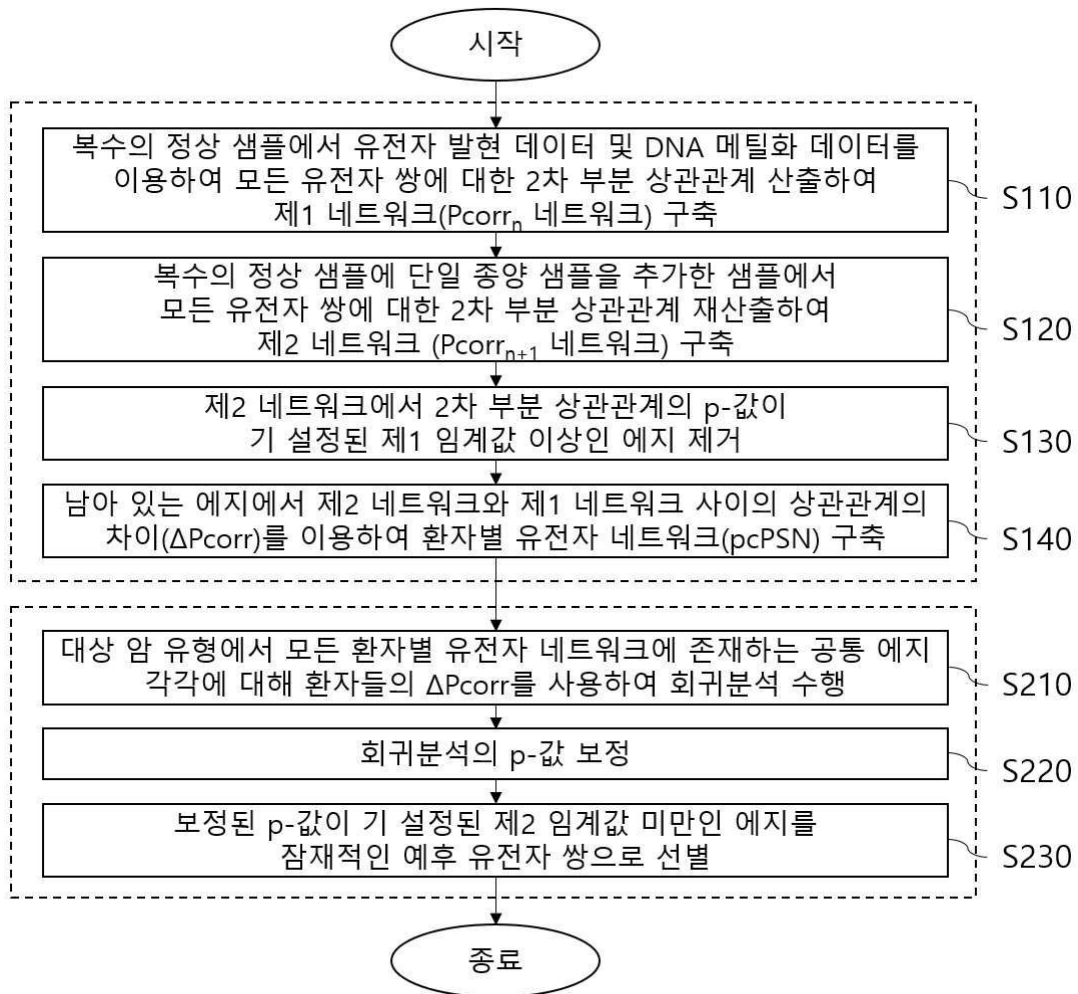
상술한 바와 같은 본 발명의 실시예에 따르면, 부분 상관관계를 통해 유전자 발현에 영향을 미치는 인자를 제어함으로써 환자별 네트워크를 구축할 수 있다.

[0102]

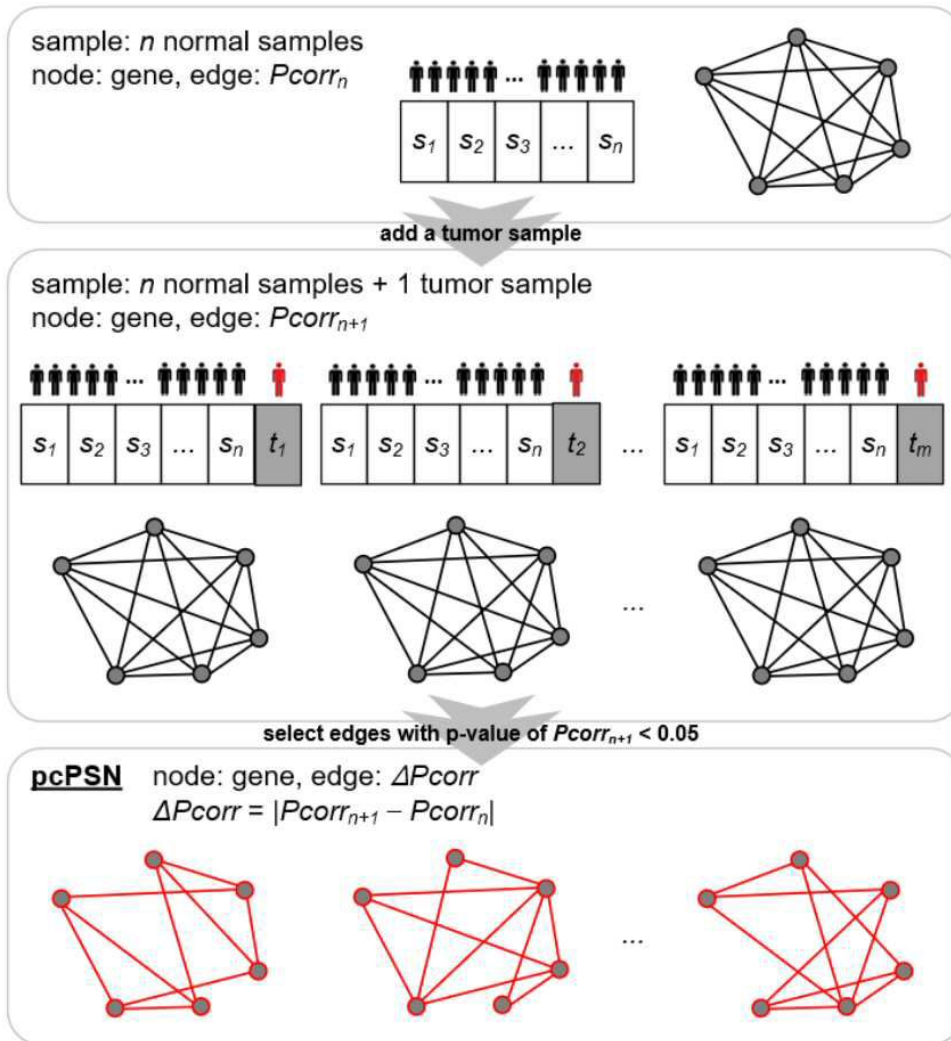
본 발명은 도면에 도시된 실시예를 참고로 하여 설명되었으나 이는 예시적인 것에 불과하며, 당해 기술이 속하는 분야에서 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 따라서 본 발명의 진정한 기술적 보호범위는 아래의 특허청구범위의 기술적 사상에 의하여 정해져야 할 것이다.

도면

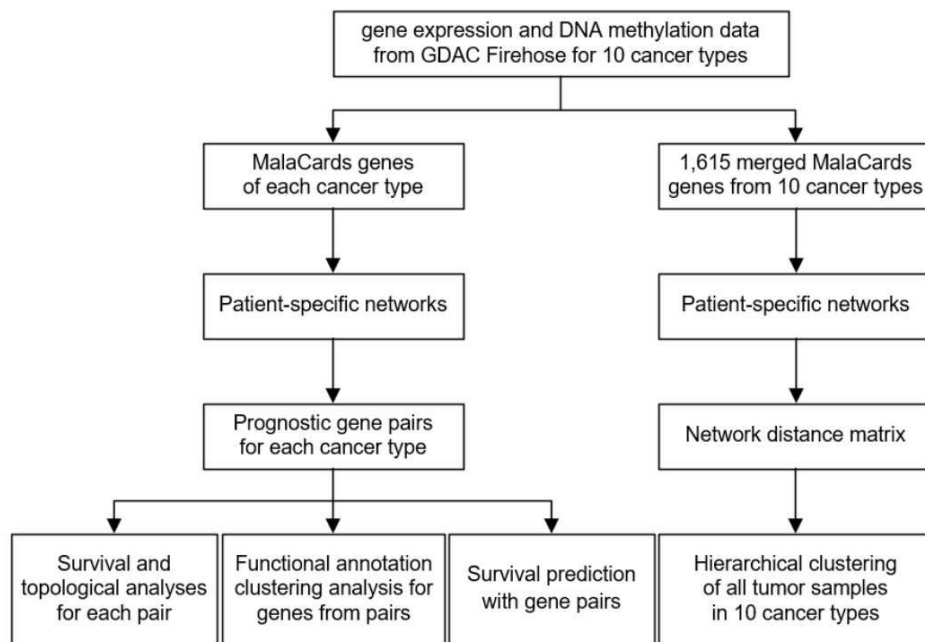
도면1



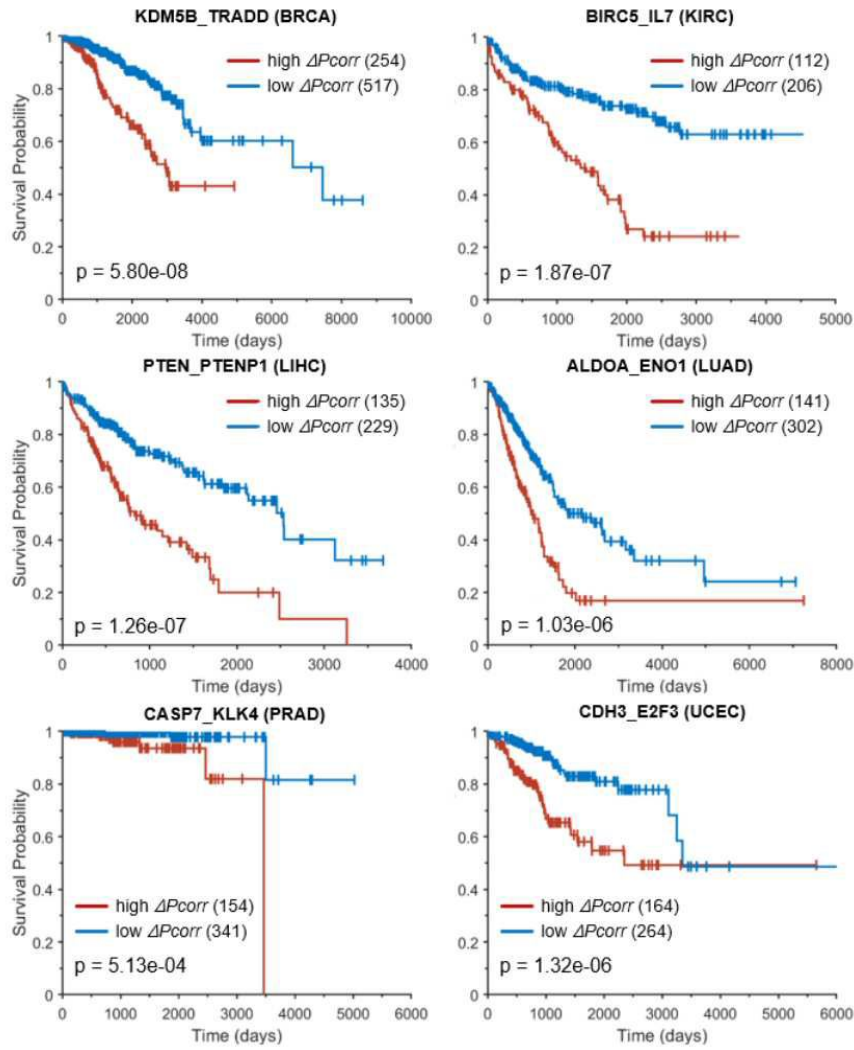
도면2



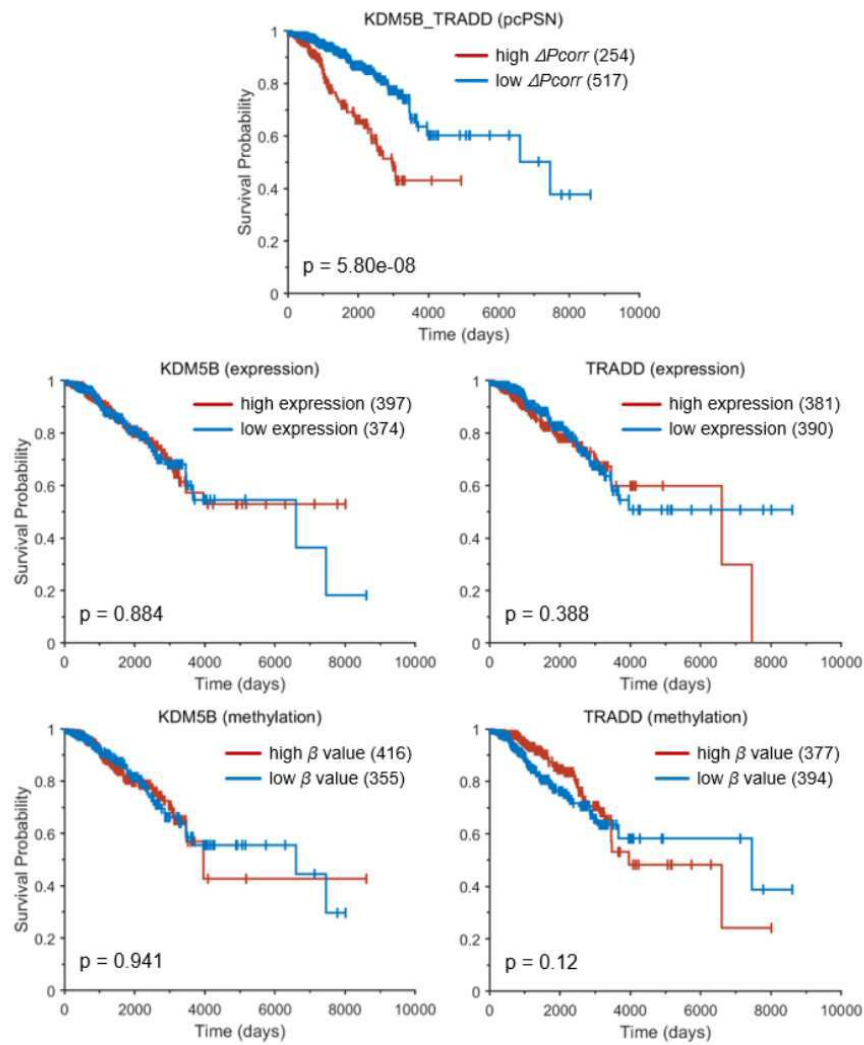
도면3



도면4

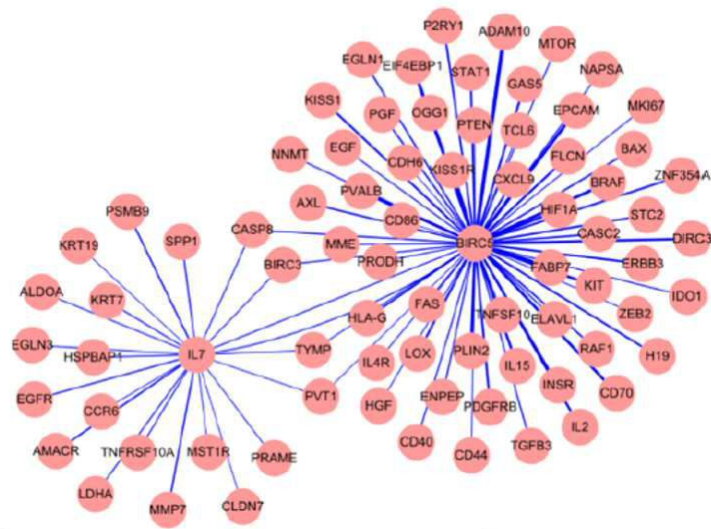


도면5

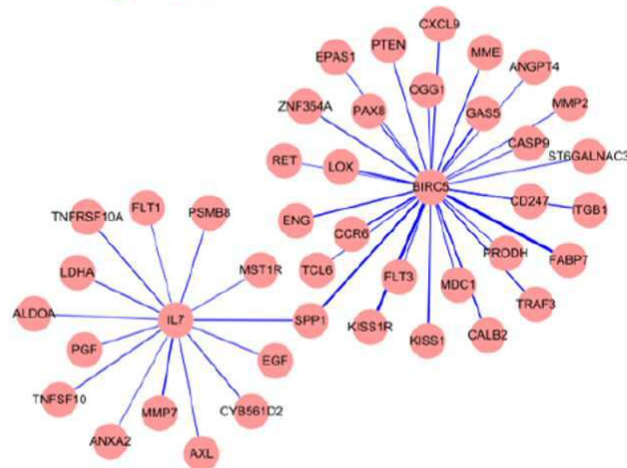


도면6

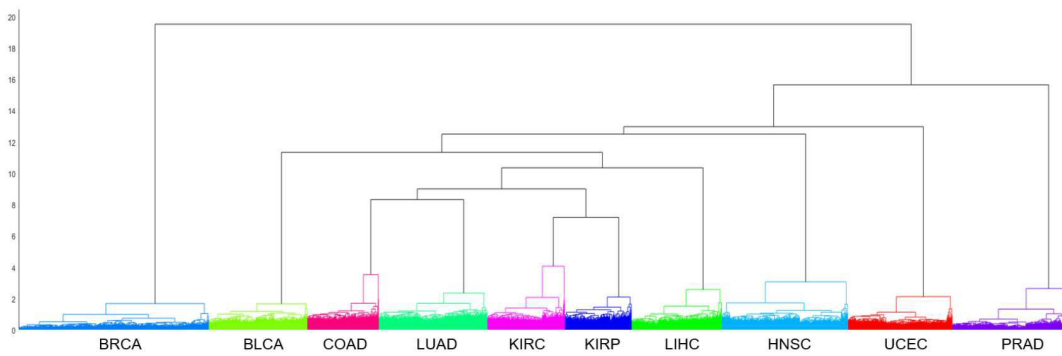
(a)



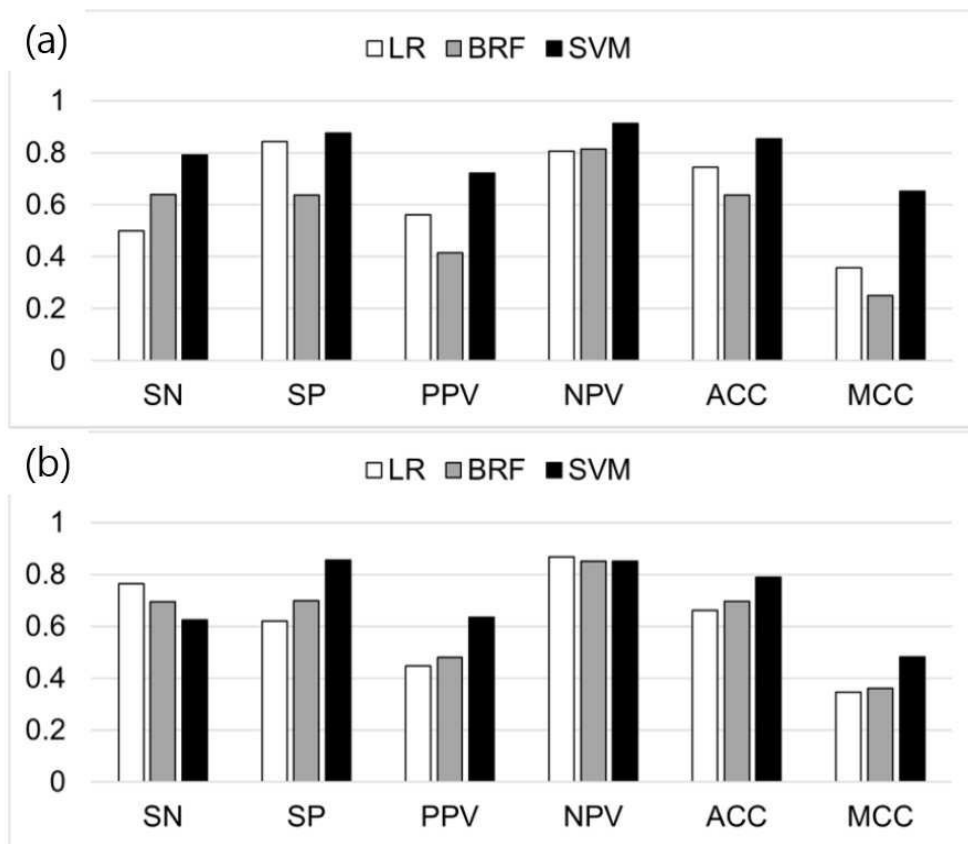
(b)



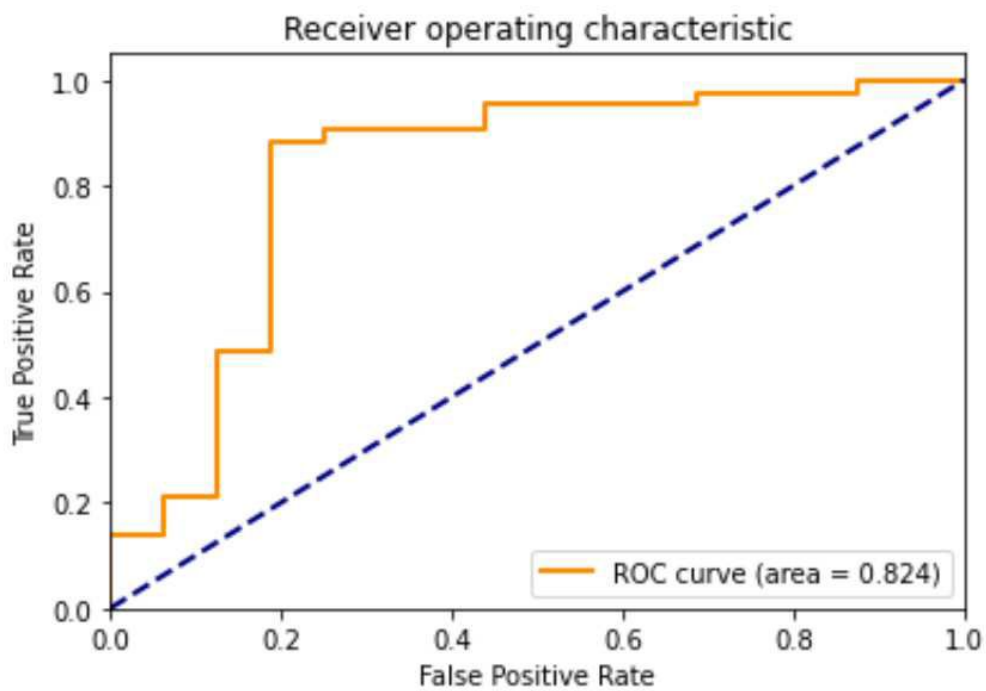
도면7



도면8



도면9



도면10

