(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0047711
(43) 공개일자 2024년04월12일

(51) 국제특허분류(Int. Cl.)
*G16B 5/00* (2019.01) *G06N 20/00* (2019.01)
*G16B 20/00* (2019.01) *G16B 40/20* (2019.01)
(52) CPC특허분류
*G16B 5/00* (2019.02)
*G06N 20/00* (2021.08)
(21) 출원번호 10-2022-0127077
(22) 출원일자 2022년10월05일
심사청구일자 2022년10월05일

(71) 출원인
**인하대학교 산학협력단**
인천광역시 미추홀구 인하로 100(용현동, 인하대학교)
(72) 발명자
**한경숙**
서울특별시 서초구 사임당로 137 , 3동 309호(서초동,신동아아파트)
**이석우**
서울특별시 영등포구 여의동로3길 10, 101동 2502호(여의도동,여의도자이)
*(뒷면에 계속)*
(74) 대리인
**네이트특허법인**

전체 청구항 수 : 총 14 항

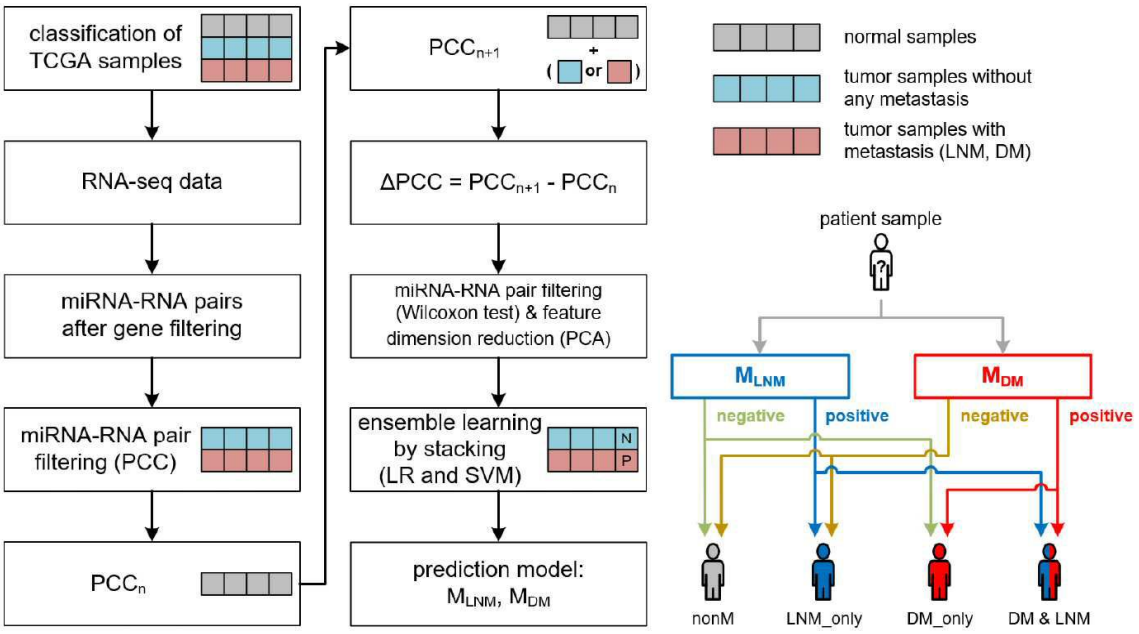
(54) 발명의 명칭 **마이크로 RNA와 표적 RNA의 상관관계를 이용한 암 환자의 림프절 전이 및 원격 전이 예측 모델 구축 방법 및 프로그램을 기록한 컴퓨터로 판독할 수 있는 기록 매체**

## *(57) 요 약*

본 발명은 암 샘플 데이터세트를 전이 여부에 따라 분류하고, 암과 연계된 마이크로 RNA와 표적 RNA를 분류하는 샘플 분류 수단; 정상 샘플 데이터세트에 대하여 마이크로 RNA와 표적 RNA 쌍 사이의 상관관계와, 암 샘플 데이터세트가 추가된 데이터세트에 대하여 마이크로 RNA와 표적 RNA 쌍 사이의 교란된 상관관계를 연산하는 상관관계

(뒷면에 계속)

*대 표 도* - 도2



연산 수단; 정상 샘플 데이터세트의 상관관계와 교란된 상관관계 사이의 차이 값을 이용하여 림프절 전이 및 원격 전이에 특이적인 마이크로 RNA와 표적 RNA 쌍을 선별하는 상관관계 필터링 수단; 및 마이크로 RNA와 표적 RNA 쌍 사이의 상관관계에 기초한 기계 학습을 수행하여 림프절 전이 예측 모델 및 원격 전이 예측 모델을 생성하는 예측 모델 구축 수단을 포함하는 림프절 전이 예측 모델 및 원격 전이 예측 모델을 구축하는 프로그램을 기록한 컴퓨터로 판독할 수 있는 기록 매체 및 이러한 프로그램을 이용한 림프절 전이 예측 모델 및 원격 전이 예측 모델을 구축하는 방법에 관한 것이다. 본 발명에 따른 프로그램 및 방법을 이용하여, 림프절 전이 및 원격 전이가 진행되기 전에 암 샘플로부터 림프절 전이 및 원격 전이를 정밀하게 예측할 수 있으며, 암 환자의 예후 예측과 분석과 같은 의료 분야에 활용될 수 있다.

(52) CPC특허분류
*G16B 20/00* (2019.02)
*G16B 40/20* (2019.02)

(72) 발명자
**조명훈**
대구광역시 달서구 선원로 171, 103동 1705호(이곡동,성서무지개타운)
**이욱**
인천광역시 미추홀구 한나루로463번길 66, 201호(용현동)
**박병규**
인천광역시 남동구 인주대로676번길 22, 1동 408호(구월동,동아아파트)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711156519
과제번호 66777-1
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 중견후속연구
연구과제명 바이오 빅데이터로부터 개인별 분자관계 네트워크 추론과 암의 조기 진단과 예후 예측을 위한 네트워크 바이오마커 발굴
기 여 율 1/1
과제수행기관명 인하대학교
연구기간 2022.03.01 ~ 2023.02.28

# 명 세 서

## 청구범위

### 청구항 1

암 샘플 데이터세트를 전이(metastasis)가 일어나지 않은 샘플 데이터세트, 림프절 전이만 발생한 샘플 데이터세트, 원격 전이만 발생한 샘플 데이터세트, 및 림프절 전이와 원격 전이가 발생한 샘플 데이터세트로 분류하고, 암과 연계된 마이크로 RNA(microRNA, miRNA)와, 상기 마이크로 RNA의 표적 RNA를 분류하는 샘플 분류 수단;

정상 샘플 데이터세트에서 상기 샘플 분류 수단에 의해 분류된 암과 연계된 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 각각에 대한 상관관계와, 상기 정상 샘플 데이터세트에 암 샘플 데이터세트를 추가한 데이터세트에서 상기 샘플 분류 수단에 의해 분류된 암과 연계된 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 각각에 대하여 교란된 상관관계를 연산하는 상관관계 연산 수단;

상기 정상 샘플 데이터세트로부터 연산된 상관관계와, 상기 교란된 상관관계 사이의 차이 값을 이용하여, 상기 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 중에서 각각 림프절 전이에 특이적인 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 및 원격 전이에 특이적인 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍을 선별하는 상관관계 필터링 수단; 및

상기 상관관계 필터링 수단에 의해 선별된 상기 림프절 전이에 특이적인 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 및 상기 원격 전이에 특이적인 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 사이의 상관관계에 기초한 기계 학습을 수행하여, 암 샘플에 대한 림프절 전이 예측 모델 및 원격 전이 예측 모델을 각각 생성하는 예측 모델 구축 수단

을 포함하는 림프절 전이 예측 모델 및 원격 전이 예측 모델을 구축하는 프로그램을 기록한 컴퓨터로 판독할 수 있는 기록 매체.

### 청구항 2

제 1항에 있어서,

상기 정상 샘플 데이터세트로부터 연산된 상관관계 및 상기 교란된 상관관계는 피어슨 상관계수(Pearson correlation coefficient, PCC)를 이용하여 연산되는 림프절 전이 예측 모델 및 원격 전이 예측 모델을 구축하는 프로그램을 기록한 컴퓨터로 판독할 수 있는 기록 매체.

### 청구항 3

제 1항에 있어서,

상기 정상 샘플 데이터세트로부터 연산된 상관관계 및 상기 교란된 상관관계는 하기 식 (1)로 나타낸 피어슨 상관관계를 이용하여 연산되는 림프절 전이 예측 모델 및 원격 전이 예측 모델을 구축하는 프로그램을 기록한 컴퓨터로 판독할 수 있는 기록 매체.

$$PCC(X,Y) = \frac{\sum_{k=1}^{n}(X_k - \bar{X})(Y_k - \bar{Y})}{\sqrt{\sum_{k=1}^{n}(X_k - \bar{X})^2}\sqrt{\sum_{k=1}^{n}(Y_k - \bar{Y})^2}} \quad (1)$$

식 1에서, $n$은 샘플의 개수, $X_k$는 샘플 $k$에서 마이크로 RNA인 X의 counts per million (CPM) 값; $\bar{X}$는 $n$개의 샘플에서 miRNA인 X의 평균 CPM 값; $Y_k$는 샘플 $k$에서 표적 RNA인 Y의 CPM 값; $\bar{Y}$는 $n$개의 샘플에서 표적 RNA인 Y

의 평균 CPM 값이다.

**청구항 4**

제 1항에 있어서,

상기 림프절 전이 예측 모델은 림프절 전이만 있는 샘플, 또는 림프절 전이 및 원격 전이가 발생하는 샘플은 림프절 전이에 대하여 양성으로 분류하고, 전이가 없는 샘플은 림프절 전이에 대하여 음성으로 분류하며,

상기 원격 전이 예측 모델은 원격 전이만 있는 샘플, 또는 림프절 전이 및 원격 전이가 발생하는 샘플은 원격 전이에 대하여 양성으로 분류하고, 전이가 없는 샘플 또는 림프절 전이만 있는 샘플은 원격 전이에 대하여 음성으로 분류하는 림프절 전이 예측 모델 및 원격 전이 예측 모델을 구축하는 프로그램을 기록한 컴퓨터로 판독할 수 있는 기록 매체.

**청구항 5**

제 1항에 있어서,

상기 마이크로 RNA의 표적 RNA는 전령 RNA(mRNA), 긴 비-암호화 RNA (long non-coding RNA, lncRNA), 및 위유전자(pseudogene) 중에서 적어도 하나를 포함하는 림프절 전이 예측 모델 및 원격 전이 예측 모델을 구축하는 프로그램을 기록한 컴퓨터로 판독할 수 있는 기록 매체.

**청구항 6**

제 1항에 있어서,

상기 상관관계 필터링 수단은 상기 선별된 림프절 전이에 특이적인 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 및 상기 선별된 원격 전이에 특이적인 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍에 대한 차원 감소 분석을 더욱 수행하는 림프절 전이 예측 모델 및 원격 전이 예측 모델을 구축하는 프로그램을 기록한 컴퓨터로 판독할 수 있는 기록 매체.

**청구항 7**

제 1항에 있어서,

상기 기계 학습은 서포터 벡터 머신(support vector machine) 및 로지스틱 회귀(logistic regression) 중에서 적어도 하나를 기본 학습자로 하여 수행되는 림프절 전이 예측 모델 및 원격 전이 예측 모델을 구축하는 프로그램을 기록한 컴퓨터로 판독할 수 있는 기록 매체.

**청구항 8**

샘플 분류 수단에 의하여, 암 샘플 데이터세트를 전이(metastasis)가 일어나지 않은 샘플 데이터세트, 림프절 전이만 발생한 샘플 데이터세트, 원격 전이만 발생한 샘플 데이터세트, 및 림프절 전이와 원격 전이가 발생한 샘플 데이터세트로 분류하고, 암과 연계된 마이크로 RNA(microRNA, miRNA)와, 상기 마이크로 RNA의 표적 RNA를 분류하는 단계;

상관관계 연산 수단에 의하여, 정상 샘플 데이터세트에서 상기 샘플 분류 수단에 의해 분류된 암과 연계된 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 각각에 대한 상관관계와, 상기 정상 샘플 데이터세트에 암 샘플 데이터세트를 추가한 데이터세트에서 상기 샘플 분류 수단에 의해 분류된 암과 연계된 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 각각에 대하여 교란된 상관관계를 연산하는 단계;

상관관계 필터링 수단에 의하여, 상기 정상 샘플 데이터세트로부터 연산된 상관관계와, 상기 교란된 상관관계

사이의 차이 값을 이용하여, 상기 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 중에서 각각 림프절 전이에 특이적인 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 및 원격 전이에 특이적인 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍을 선별하는 단계; 및

예측 모델 구축 수단에 의하여, 상기 상관관계 필터링 수단에 의해 선별된 상기 림프절 전이에 특이적인 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 및 상기 원격 전이에 특이적인 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 사이의 상관관계에 기초한 기계 학습을 수행하여, 암 샘플에 대한 림프절 전이 예측 모델 및 원격 전이 예측 모델을 각각 생성하는 단계

를 림프절 전이 예측 모델 및 원격 전이 예측 모델을 구축하는 방법.

**청구항 9**

제 8항에 있어서,

상기 정상 샘플 데이터세트로부터 연산된 상관관계 및 상기 교란된 상관관계는 피어슨 상관계수(Pearson correlation coefficient, PCC)를 이용하여 연산되는 림프절 전이 예측 모델 및 원격 전이 예측 모델을 구축하는 방법.

**청구항 10**

제 8항에 있어서,

상기 정상 샘플 데이터세트로부터 연산된 상관관계 및 상기 교란된 상관관계는 하기 식 (1)로 나타낸 피어슨 상관관계를 이용하여 연산되는 림프절 전이 예측 모델 및 원격 전이 예측 모델을 구축하는 방법.

$$PCC(X,Y) = \frac{\sum_{k=1}^{n}(X_k - \bar{X})(Y_k - \bar{Y})}{\sqrt{\sum_{k=1}^{n}(X_k - \bar{X})^2}\sqrt{\sum_{k=1}^{n}(Y_k - \bar{Y})^2}} \quad (1)$$

식 1에서, $n$은 샘플의 개수, $X_k$는 샘플 $k$에서 마이크로 RNA인 X의 counts per million (CPM) 값; $\bar{X}$는 $n$개의 샘플에서 miRNA인 X의 평균 CPM 값; $Y_k$는 샘플 $k$에서 표적 RNA인 Y의 CPM 값; $\bar{Y}$는 $n$개의 샘플에서 표적 RNA인 Y의 평균 CPM 값이다.

**청구항 11**

제 8항에 있어서,

상기 림프절 전이 예측 모델은 림프절 전이만 있는 샘플, 또는 림프절 전이 및 원격 전이가 발생하는 샘플은 림프절 전이에 대하여 양성으로 분류하고, 전이가 없는 샘플은 림프절 전이에 대하여 음성으로 분류하며,

상기 원격 전이 예측 모델은 원격 전이만 있는 샘플, 또는 림프절 전이 및 원격 전이가 발생하는 샘플은 원격 전이에 대하여 양성으로 분류하고, 전이가 없는 샘플 또는 림프절 전이만 있는 샘플은 원격 전이에 대하여 음성으로 분류하는 림프절 전이 예측 모델 및 원격 전이 예측 모델을 구축하는 방법.

**청구항 12**

제 8항에 있어서,

상기 마이크로 RNA의 표적 RNA는 전령 RNA(mRNA), 긴 비-암호화 RNA (long non-coding RNA, lncRNA), 및 위유전자(pseudogene) 중에서 적어도 하나를 포함하는 림프절 전이 예측 모델 및 원격 전이 예측 모델을 구축하는

방법.

**청구항 13**

제 8항에 있어서,

상기 상관관계 필터링 수단은 상기 선별된 림프절 전이에 특이적인 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 및 상기 선별된 원격 전이에 특이적인 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍에 대한 차원 감소 분석을 더욱 수행하는 림프절 전이 예측 모델 및 원격 전이 예측 모델을 구축하는 방법.

**청구항 14**

제 8항에 있어서,

상기 기계 학습은 서포터 벡터 머신(support vector machine) 및 로지스틱 회귀(logistic regression) 중에서 적어도 하나를 기본 학습자로 하여 수행되는 림프절 전이 예측 모델 및 원격 전이 예측 모델을 구축하는 방법.

## *발명의 설명*

### *기 술 분 야*

[0001] 본 특허출원은 대한민국 정부 과학기술정보통신부의 "중견후속연구" 연구사업의 일환으로서 "바이오 빅데이터로부터 개인별 분자관계 네트워크 추론과 암의 조기 진단과 예후 예측을 위한 네트워크 바이오마커 발굴"(주관기관: 인하대학교; 과제고유번호: 1711156519) 과제의 수행 결과물에 관한 것이다.

[0002] 본 발명은 암 환자의 전이 예측 모듈을 구축하는 방법에 관한 것으로, 더욱 상세하게는 마이크로 RNA와 표적 RNA의 상관 관계를 이용하여 암 환자의 림프절 전이 예측 모델 및 원격 전이 예측 모델 구축 방법 및 프로그램을 기록한 컴퓨터로 판독할 수 있는 기록 매체에 관한 것이다.

### *배 경 기 술*

[0003] 암(cancer) 및 종양(tumor)를 치료, 예방하기 위한 많은 연구가 수행되었으나, 여전히 암은 세계적으로 주요한 사망 원인으로서, 매년 1000만 명 가량, 즉 6명 중의 1명은 암으로 사망하고 있다. 특히 전이는 암으로 인한 이환 및 사망의 주요 원인으로서, 암으로 인한 사망 중에서 대략 90%는 암의 전이에 기인한다. 전이(metastasis)는 암세포가 최초 종양 부위로부터 통상 림프 시스템이나 혈관을 통해 신체의 다른 부분으로 전파되는 것이다.

[0004] 암의 TNM 시기 결정(원발성 종양인 T, 국소 림프절 N, 전이 M)에 따르면, 암의 전이는 크게 림프절 전이(lymph node metastasis)와 원격 전이(distant metastasis)로 구분된다. 림프절 전이에서, 암세포는 인근의 림프절로 이동하지만, 원격 전이에서 암세포는 최초 종양 부위에서 멀리 떨어진 기관까지 퍼지며, 림프절 전이는 흔히 원격 전이로 진행된다. 전이가 있는 환자는 전이가 일어나지 않은 환자보다 재발의 위험은 크고 생존율은 낮아진다.

[0005] 암의 전이 여부에 대한 진단은 주로 임상적 검사 및 영상 장치를 활용하여 이루어진다. 임상적 검사 및 영상 검사는 전이를 진단하기 위한 가장 직접적이고 정확한 방법이지만, 이러한 진단은 전이가 발생한 이후에만 가능하다는 단점이 있다.

[0006] 따라서, 암에 대한 전이가 발생하기 이전에 신속하고 정확하게 림프절 전이 및 원격 전이를 예측할 수 있는 방법에 대한 필요가 있다.

## *발명의 내용*

### *해결하려는 과제*

[0007] 본 발명은 전술한 종래 기술의 문제점을 해소하기 위하여 제안된 것으로, 본 발명의 목적은 암 환자의 샘플 데이터에 기초하여 림프절 전이 및 원격 전이 여부를 신속하고 정확하게 예측할 수 있는 림프절 전이 및 원격 전이 예측 모델을 구축하는 프로그램을 기록한 컴퓨터로 판독할 수 있는 기록 매체 및 이를 이용하여 암의 림프절

전이 및 원격 전이 예측 모델을 구축하는 방법을 제공하고자 하는 것이다.

## 과제의 해결 수단

[0008] 일 측면에 따르면, 본 발명은 암 샘플 데이터세트를 전이(metastasis)가 일어나지 않은 샘플 데이터세트, 림프절 전이만 발생한 샘플 데이터세트, 원격 전이만 발생한 샘플 데이터세트, 및 림프절 전이와 원격 전이가 발생한 샘플 데이터세트로 분류하고, 암과 연계된 마이크로 RNA(microRNA, miRNA)와, 상기 마이크로 RNA의 표적 RNA를 분류하는 샘플 분류 수단; 정상 샘플 데이터세트에서 상기 샘플 분류 수단에 의해 분류된 암과 연계된 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 각각에 대한 상관관계와, 상기 정상 샘플 데이터세트에 암 샘플 데이터세트를 추가한 데이터세트에서 상기 샘플 분류 수단에 의해 분류된 암과 연계된 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 각각에 대하여 교란된 상관관계를 연산하는 상관관계 연산 수단; 상기 정상 샘플 데이터세트로부터 연산된 상관관계와, 상기 교란된 상관관계 사이의 차이 값을 이용하여, 상기 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 중에서 각각 림프절 전이에 특이적인 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 및 원격 전이에 특이적인 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍을 선별하는 상관관계 필터링 수단; 및 상기 상관관계 필터링 수단에 의해 선별된 상기 림프절 전이에 특이적인 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 및 상기 원격 전이에 특이적인 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 사이의 상관관계에 기초한 기계 학습을 수행하여, 암 샘플에 대한 림프절 전이 예측 모델 및 원격 전이 예측 모델을 각각 생성하는 예측 모델 구축 수단을 포함하는 림프절 전이 예측 모델 및 원격 전이 예측 모델을 구축하는 프로그램을 기록한 컴퓨터로 판독할 수 있는 기록 매체를 제공한다.

[0009] 상기 정상 샘플 데이터세트로부터 연산된 상관관계 및 상기 교란된 상관관계는 피어슨 상관계수(Pearson correlation coefficient, PCC)를 이용하여 연산될 수 있다.

[0010] 일례로, 상기 정상 샘플 데이터세트로부터 연산된 상관관계 및 상기 교란된 상관관계는 하기 식 (1)로 나타낸 피어슨 상관관계를 이용하여 연산될 수 있다.

[0011]
$$PCC(X,Y) = \frac{\sum_{k=1}^{n}(X_k - \bar{X})(Y_k - \bar{Y})}{\sqrt{\sum_{k=1}^{n}(X_k - \bar{X})^2}\sqrt{\sum_{k=1}^{n}(Y_k - \bar{Y})^2}} \quad (1)$$

[0012] 식 1에서, $n$은 샘플의 개수, $X_k$는 샘플 $k$에서 마이크로 RNA인 X의 counts per million (CPM) 값; $\bar{X}$는 $n$개의 샘플에서 miRNA인 X의 평균 CPM 값; $Y_k$는 샘플 $k$에서 표적 RNA인 Y의 CPM 값; $\bar{Y}$는 $n$개의 샘플에서 표적 RNA인 Y의 평균 CPM 값이다.

[0013] 상기 림프절 전이 예측 모델은 림프절 전이만 있는 샘플, 또는 림프절 전이 및 원격 전이가 발생하는 샘플은 림프절 전이에 대하여 양성으로 분류하고, 전이가 없는 샘플은 림프절 전이에 대하여 음성으로 분류하며, 상기 원격 전이 예측 모델은 원격 전이만 있는 샘플, 또는 림프절 전이 및 원격 전이가 발생하는 샘플은 원격 전이에 대하여 양성으로 분류하고, 전이가 없는 샘플 또는 림프절 전이만 있는 샘플은 원격 전이에 대하여 음성으로 분류할 수 있다.

[0014] 상기 마이크로 RNA의 표적 RNA는 전령 RNA(mRNA), 긴 비-암호화 RNA (long non-coding RNA, lncRNA), 및 위유전자(pseudogene) 중에서 적어도 하나, 예를 들어 mRNA, lncRNA 및 위유전자를 모두 포함할 수 있다.

[0015] 상기 상관관계 필터링 수단은 상기 선별된 림프절 전이에 특이적인 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 및 상기 선별된 원격 전이에 특이적인 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍에 대한 차원 감소 분석을 더욱 수행할 수 있다.

[0016] 일례로, 상기 기계 학습은 서포터 벡터 머신(support vector machine) 및 로지스틱 회귀(logistic regression) 중에서 적어도 하나를 기본 학습자로 하여, 예를 들어 2개를 모두 이용한 앙상블 학습자로서, 수행될 수 있다.

[0017] 다른 측면에서, 본 발명은 샘플 분류 수단에 의하여, 암 샘플 데이터세트를 전이(metastasis)가 일어나지 않은 샘플 데이터세트, 림프절 전이만 발생한 샘플 데이터세트, 원격 전이만 발생한 샘플 데이터세트, 및 림프절 전이와 원격 전이가 발생한 샘플 데이터세트로 분류하고, 암과 연계된 마이크로 RNA(microRNA, miRNA)와, 상기 마이크로 RNA의 표적 RNA를 분류하는 단계; 상관관계 연산 수단에 의하여, 정상 샘플 데이터세트에서 상기 샘플

분류 수단에 의해 분류된 암과 연계된 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 각각에 대한 상관관계와, 상기 정상 샘플 데이터세트에 암 샘플 데이터세트를 추가한 데이터세트에서 상기 샘플 분류 수단에 의해 분류된 암과 연계된 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 각각에 대하여 교란된 상관관계를 연산하는 단계; 상관관계 필터링 수단에 의하여, 상기 정상 샘플 데이터세트로부터 연산된 상관관계와, 상기 교란된 상관관계 사이의 차이 값을 이용하여, 상기 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 중에서 각각 림프절 전이에 특이적인 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 및 원격 전이에 특이적인 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍을 선별하는 단계; 및 예측 모델 구축 수단에 의하여, 상기 상관관계 필터링 수단에 의해 선별된 상기 림프절 전이에 특이적인 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 및 상기 원격 전이에 특이적인 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 사이의 상관관계에 기초한 기계 학습을 수행하여, 암 샘플에 대한 림프절 전이 예측 모델 및 원격 전이 예측 모델을 각각 생성하는 단계를 림프절 전이 예측 모델 및 원격 전이 예측 모델을 구축하는 방법을 제공한다.

## *발명의 효과*

[0018] 본 발명의 프로그램 및 방법에 따르면, 암 환자 샘플로부터 림프절 전이 및 원격 전이 샘플을 각각 분류, 학습하고, 암과 연계된 마이크로 RNA 및 상기 마이크로 RNA의 표적 RNA 쌍 사이의 상관계수를 토대로 예측 모델을 구축하여, 암 환자 샘플로부터 림프절 전이 및 원격 전이 여부를 예측할 수 있다.

[0019] 다시 말하면, 본 발명의 프로그램 및 방법을 이용하여, 암과 연계된 마이크로 RNA-표적 RNA 쌍 사이의 상관계수에 기초하여, 암 환자 샘플로부터 림프절 전이 및 원격 전이 여부를 정밀하게 예측할 수 있다.

[0020] 본 발명을 활용하여 암 환자의 림프절 전이 및 원격 전이 여부를 검출, 진단하는 의학적 용도로 활용될 수 있으며, 암 환자의 예후 예측과 분석에도 적용될 수 있을 것으로 기대된다.

## *도면의 간단한 설명*

[0021] 도 1은 본 발명의 예시적인 실시형태에 따라, 암 환자 샘플 데이터세트에 기초하여 림프절 전이 및 원격 전이를 예측하기 위한 모델을 구축, 생성하는 프로그램이 탑재된 컴퓨터와, 프로그램이 기록된 컴퓨터로 판독 가능한 기록 매체의 구성을 개략적으로 도시한 도면이다.

도 2는 본 발명의 예시적인 실시형태에 따라, 암 환자 샘플 데이터세트에 기초하여 림프절 전이 및 원격 전이를 예측하는 모델을 구축하는 전체 프레임워크를 개략적으로 나타낸 모식도이다.

도 3은 본 발명의 예시적인 실시형태에 따라 암 환자 샘플 데이터세트로부터 림프절 전이 및 원격 전이 모델을 구축하고, 이를 토대로 림프절 전이 및 원격 전이 여부를 예측하는 방법을 개략적으로 나타낸 플로 차트이다.

도 4는 본 발명의 예시적인 실시형태에 따라 암 환자 샘플로부터 전이를 예측하는 것을 개략적으로 나타낸 모식도이다.

도 5는 본 발명의 예시적인 실시형태에 따라 구축된 림프절 전이 예측 모델을 이용하여 림프절 전이의 예측 성능을 나타낸 그래프이다.

도 6은 본 발명의 예시적인 실시형태에 따라 구축된 원격 전이 예측 모델을 이용하여 원격 전이의 예측 성능을 나타낸 그래프이다.

## *발명을 실시하기 위한 구체적인 내용*

[0022] 이하, 필요한 경우에 첨부하는 도면을 참조하면서 본 발명을 설명한다.

[0023] 암 환자로부터 얻은 샘플에 기초하여 림프절 전이는 물론이고 원격 전이 여부를 정확하고 신속하게 예측할 수 있는 모델을 구축할 필요가 있다. 본 발명에서는 암 세포의 림프절 전이는 물론이고 원격 전이를 모두 예측할 수 있다. 암은 복잡하고 이질적인 질병이기 때문에, 개별 유전자의 비정상적 발현만으로는 암의 발생 및 전이를 완전히 설명할 수 있다. 따라서, 개별 유전자의 발현보다는 유전자 상호작용의 이상조절(dysregulation)에 의해 암의 발생 및 전이가 보다 잘 설명될 수 있다.

[0024] 본 발명에서는 마이크로 RNA와 마이크로 RNA의 표적 RNA 쌍 사이의 상호작용 메커니즘을 활용하여 암 샘플 데이터세트로부터 암 샘플의 림프절 전이 및 원격 전이를 효과적이고 신속하게 예측할 수 있는 예측 모델을 구축하는 프로그램 및 방법을 제공한다.

[0025] 도 1은 본 발명의 예시적인 실시형태에 따라, 암 환자 샘플 데이터세트에 기초하여 림프절 전이 및 원격 전이를 예측하기 위한 모델을 구축, 생성하는 프로그램이 탑재된 컴퓨터와, 프로그램이 기록된 컴퓨터로 판독 가능한 기록 매체의 구성을 개략적으로 도시한 도면이다. 도 2는 본 발명의 예시적인 실시형태에 따라, 암 환자 샘플 데이터세트에 기초하여 림프절 전이 및 원격 전이를 예측하는 모델을 구축하는 전체 프레임워크를 개략적으로 나타낸 모식도이다. 도 3은 본 발명의 예시적인 실시형태에 따라 암 환자 샘플 데이터세트로부터 림프절 전이 및 원격 전이 모델을 구축하고, 이를 토대로 림프절 전이 및 원격 전이 여부를 예측하는 방법을 개략적으로 나타낸 플로 차트이다.

[0026] 도 1에 나타내 바와 같이, 암 샘플 데이터로부터 암의 림프절 전이 및 원격 전이를 모두 예측할 수 있는 프로그램(210)이 기록된 매체(200)는 적절한 컴퓨터(100)에 탑재될 수 있다. 컴퓨터(100)는 최소한의 데이터 처리(data processing) 능력을 가지는 데이터 처리 장치이다.

[0027] 예를 들어, 컴퓨터(100)는 데스크톱, 노트북과 같은 컴퓨터 단말기이고/이거나, 스마트폰, 태블릿 PC, PDA 등과 같은 모바일 단말기일 수 있다. 필요한 경우, 컴퓨터(100)는 인터넷이나, 3G, LTE 및/또는 5G 등과 같은 데이터 통신망과 연결될 수 있다. 예를 들어, 컴퓨터(100)는 서버(도시하지 않음)와 통신망을 통하여 연결될 수도 있고, 서버에 로컬(local) 형태로 연결될 수도 있지만, 본 발명에 따른 컴퓨터가 반드시 통신망에 연결되는 것은 아니다.

[0028] 컴퓨터(100)는 다양한 하드웨어(110 내지 140 및 160)와 이러한 하드웨어를 실행시키기 위한 구동/응용 프로그램(150)과 같은 소프트웨어인 다양한 컴포넌트를 포함할 수 있다.

[0029] 예시적인 실시형태에서, 컴퓨터(100)는 키보드, 마우스 및/또는 터치 패널과 같이 데이터 또는 정보, 예를 들어 응용 프로그램의 일종인 림프절 전이 예측 모델 및 원격 전이 예측 모델을 구축하는 프로그램을 이용하여 적절한 후보 압타머를 훈련, 학습하기 위한 초기 입력 값을 작성할 수 있는 입력부(110)를 갖는다. 컴퓨터(100)는 또한 입력 또는 다운로드 된 데이터 또는 정보를 인쇄하기 위한 프린터와 같은 출력부(120)와, 이들 데이터를 화면으로 표시하기 위한 모니터와 같은 디스플레이(130)를 포함한다. 아울러, 컴퓨터(100)는 다양한 데이터 및 정보(예를 들어, 데이터베이스로부터 취득한 암 샘플 데이터세트, 이러한 데이터세트를 토대로 훈련되거나 구축된 암의 림프절 전이 예측 모델 및 원격 전이 예측 모델 등)를 저장하고 있는 RAM, ROM, 하드디스크 등과 같은 메모리(140)를 가지고 있으며, 컴퓨터(100)에서의 다양한 구현 작업을 위하여 구동/응용 프로그램(150)을 탑재하고 있다.

[0030] 컴퓨터(100)에서 구현되는 구동/응용 프로그램(150)의 예로서 WINDOWS 등과 같은 구동 소프트웨어는 물론이고, 문서, 이미지, 영상 데이터 등의 콘텐츠를 작성, 편집할 수 있는 콘텐츠 편집 소프트웨어, 인터넷과 같은 통신망에서의 데이터나 정보를 검색하기 위한 브라우저, 인공 신경망을 이용한 훈련, 학습 모델을 구축하기 위한 소프트웨어 등을 들 수 있다. 하지만, 그 외에도 다양한 구동/응용 프로그램이 컴퓨터(100)에 탑재될 수 있을 것이다.

[0031] 또한, 컴퓨터(100)는 이들 하드웨어와 소프트웨어의 작업을 적절히 제어할 수 있도록 CPU와 같은 제어부(160)를 포함한다. 특히 컴퓨터(100) 중의 메모리(140)와 제어부(160)는 기록 매체(200)에 기록된 프로그램(210) 명령에 따른 일련의 작업을 실현한다는 점에서, 일종의 프로그램 실행부를 구성한다.

[0032] 한편, 컴퓨터(100)가 판독할 수 있도록 컴퓨터(100)에 탑재될 수 있는 기록 매체(200)는 본 발명에 따라 암 샘플 데이터세트로부터 암과 연계된 마이크로 RNA - 표적 RNA 쌍 사이의 상관계수에 기초하여, 암 환자 샘플로부터 림프절 전이 및 원격 전이를 예측하는 프로그램(210)이 기록된다. 후술하는 바와 같이, 본 발명에 따른 프로그램(210)을 사용하여 암 환자 샘플 데이터세트로부터 암의 림프절 전이 및 원격 전이를 정확하고 효율적으로 예측할 수 있는 모델이 생성, 구축될 수 있다. 예를 들어, 본 발명에 따라 컴퓨터로 판독 가능한 기록 매체(200)는 암 환자 샘플 데이터세트로부터 암의 림프절 전이 및 원격 전이를 예측하기 위한 일련의 프로그램 명령은 물론이고, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합한 것을 포함할 수 있다.

[0033] 본 발명에 따라 암 환자 샘플로부터 림프절 전이 및 원격 전이를 예측하는 프로그램(210) 명령이 기록된 컴퓨터로 판독 가능한 기록 매체(200)의 예는 컴퓨터 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록 장치를 포함한다. 컴퓨터(100)가 판독할 수 있는 기록 매체(200)의 예로는 자기 테이프, 하드 디스크, 플로피 디스크와 같은 자기 매체, CD-ROM, DVD와 같은 광기록 매체, 플롭티컬(floptical) 디스크와 같은 자기-광 매체, 및 ROM, RAM, 플래시 메모리 등과 같은 프로그램 명령을 저장, 수행하도록 특별히 구성된 하드웨어 장치가 포함된다.

[0034] 또한, 캐리어 웨이브(예를 들어, 인터넷을 통해 전송)의 형태로 구현되는 것도 포함되는데, 이러한 기록 매체(200)는 프로그램 명령, 데이터 구조 등을 지정하는 신호를 전송하는 반송파를 포함하는 광 또는 금속선, 도파관 등의 전송 매체일 수도 있다. 이러한 컴퓨터(100)가 판독 가능한 기록 매체(200)는 통신망으로 연결된 컴퓨터 시스템에 분산되어, 분산 방식으로 컴퓨터(100)가 읽을 수 있는 코드가 저장되고 실행될 수 있다. 한편, 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 아래에서 예시적으로 기술하고 있는, 본 발명을 구현하기 위한 기능적인 프로그램, 코드 및 코드 세그먼트들은 본 발명이 속하는 기술분야에서 통상의 프로그래머들에 의해 용이하게 추론될 수 있을 것이다.

[0035] 기록 매체(200)에 기록된 암 환자 샘플로부터 림프절 전이 예측 모델 및 원격 전이 예측 모듈을 구축하는 프로그램(이하, '림프절 전이 및 원격 전이 예측 모듈 구축 프로그램'이라고 약칭하는 경우가 있다. 210)은 암 샘플 데이터세트를 림프절 전이 및 원격 전이 유무에 따라 분류하는 샘플 분류 수단(220), 샘플 분류 수단(220)에 따라 분류된 암 샘플 데이터세트로부터 암과 연계된 마이크로 RNA-표적 RNA 쌍의 상관관계를 연산하는 상관관계 연산 수단(230), 상관관계 연산 수단(230)에서 연산된 상관관계 사이의 값을 이용하여, 신뢰성 있는 마이크로 RNA-표적 RNA 쌍을 선별하는 상관관계 필터링 수단(240), 선별된 상기 마이크로 RNA-표적 RNA 쌍에 기초한 기계학습을 수행하여, 암 샘플에 대한 림프절 전이 예측 모델 및 원격 전이 예측 모델을 각각 생성하는 예측 모델 구축 수단(250)을 포함할 수 있다. 본 명세서에서 암(cancer)과 종양(tumor)은 상호 교환 가능하게 사용될 수 있다.

[0036] 샘플 분류 수단(220)은 암 샘플 데이터세트를 전이(metastasis)가 일어나지 않은 샘플 데이터세트, 림프절 전이만 발생한 샘플 데이터세트, 원격 전이만 발생한 샘플 데이터세트, 및 림프절 전이와 원격 전이가 모두 발생한 샘플 데이터세트로 분류한다. 예를 들어, 암 샘플에 대하여 종양(Tumor), 림프절(Node, N) 및 전이(Metastasis, M) 상태에 따른 TNM 단계 임상 보완 데이터를 토대로 암 샘플 데이터세트를 전술한 4개의 데이터세트로 분류할 수 있다.

[0037] 예를 들어, 암 샘플 데이터세트는 암유전체지도(The Cancer Genome Atlas, TCGA)와 같이 공중이 접근할 수 있는 사이트를 이용하여 입수할 수 있으나, 이에 한정되는 것은 아니다.

[0038] 또한, 샘플 분류 수단(220)은 암과 연계된 마이크로 RNA와, 상기 마이크로 RNA의 표적 RNA를 더욱 분류할 수 있다. 후술하는 바와 같이, 본 발명에서는 전이된 암 세포에서 마이크로 RNA(microRNA, miRNA)와 표적 RNAs 사이의 차별화된 상관관계에 기초하여 전이를 예측한다.

[0039] 마이크로RNA는 생물의 유전자 발현을 제어하는 RNA로서 통상적으로 20~25개의 뉴클레오타이드(nucleotide)로 구성된다. 대부분의 생명체는 DNA 형태의 유전자로부터 메신저 RNA(messenger RNA, mRNA)로 1차 전사(transcription)되고, mRNA가 단백질 또는 펩타이드로 2차 번역(translation)되어 최종적인 단백질을 발현(expression)한다.

[0040] 유전자가 DNA로부터 mRNA로 전사된 뒤, 마이크로 RNA가 mRNA와 상보적(complementary) 결합하여, RNA silencing 및 전사 이후의 유전자 발현 조절과 같은 세포 내 유전자 발현 과정에서 중요한 조절 인자로 작용한다. 마이크로 RNA와 mRNA가 상보적으로 결합하면서, mRNA 가닥이 2개의 조각으로 분리되거나, mRNA 말단의 폴리아데노신 테일(polyadenosine tail)이 축소되면서 mRNA를 불안정화하거나, 리보솜에 의한 mRNA 번역의 효율을 감소시킨다.

[0041] 마이크로 RNA는 스템-루프(stem-loop) 구조를 포함하는 짧은 헤어핀 구조를 형성하기 위해 스스로 접히는 RNA 전사체(transcript) 영역으로부터 타온 마이크로 RNA를 제외하면 간섭 RNA(RNA interference, RNAi) 경로의 작은 간섭 RNA(short interference RNA, siRNA)와 유사하다. 특히 동물의 마이크로 RNA는 5' 말단에 있는 6~8개 정도의 적은 뉴클레오타이드 서열을 통해 그들의 표적을 인식할 수 있는 것으로 알려져 있다. 통상 하나의 마이크로 RNA는 수백 가지의 표적 mRNA의 발현을 조절하는 반면, 하나의 표적 RNA는 다수의 마이크로 RNA에 의해 조절된다.

[0042] 암과 연계된 마이크로 RNA로서, AKT1 RNA에 결합하는 miR-138은 혀의 편평상피세포암(tongue squamous cell carcinoma)에서 AKT1의 발현을 조절하고, miR-519d는 구강 편평상피세포암(oral squamous cell carcinoma) 및 유방암에서 MMP3를 제어하여 림프절 전이를 억제하는 것으로 알려져 있다.

[0043] 예시적인 측면에서, 암과 연계된 마이크로 RNA 및 마이크로 RNA의 표적 RNA는 TCGA의 Genomic Data Commons (GDC)에서 공개한 데이터를 이용하여 추출, 필터링(filtering)될 수 있다. 예를 들어, 마이크로 RNA의 표적 RNA

는 단백질 또는 펩타이드로 발현(expression)되는 전령 RNA(messenger RNA, mRNA), 단백질로 발현되지 않으며 통상 200개 이상의 뉴클레오타이드로 이루어지는 긴 비-암호화 RNA(long non-coding RNA, lncRNA) 및 위유전자(pseudogene) 중에서 적어도 하나를 포함할 수 있다. 일례로, 마이크로 RNA의 표적 RNA는 mRNA, lncRNA 및 위유전자를 모두 포함할 수 있다. 긴 비-암호화 RNA는 intergenic lncRNA, intronic lncRNA, Sense lncRNA, Antisense lncRNA, Bidirectional lncRNA 및/또는 Enhancer lncRNA를 포함할 것이다.

[0044] 상관관계 연산 수단(230)은 샘플 분류 수단(220)에 의해 분류된 암과 연계된 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 각각에 대한 상관관계를 연산한다. 구체적으로, 상관관계 연산 수단(230)은 암이 발생하지 않은 정상 샘플 데이터세트에서 샘플 분류 수단(220)에 의해 분류된 암과 연계된 마이크로 RNA와 마이크로 RNA의 표적 RNA 쌍(miRNA-RNA 쌍) 각각에 대한 상관관계를 연산한다. 또한, 상관관계 연산 수단(230)은 정상 샘플 데이터세트에 암 샘플 데이터세트를 추가한 데이터세트에서 샘플 분류 수단(230)에 의해 분류된 암과 연계된 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 상 각각에 대하여 교란된 상관관계를 연산할 수 있다.

[0045] 예시적인 실시형태에서, 정상 샘플 데이터세트에서 연산된 상관관계 및 암 샘플 데이터세트가 추가된 데이터세트에서 연산된 상관관계는 각각 피어슨 상관계수(Pearson correlation coefficient, PCC)를 이용하여 연산될 수 있다. 도 2를 참조하면, 정상 샘플 데이터세트에서 암과 연계된 마이크로 RNA와 마이크로 RNA의 표적 RNA 쌍에 대한 피어슨 상관계수를 $PCC_n$으로 표시하였고, 적어도 하나의 암 샘플 데이터세트가 추가된 데이터세트에서 암과 연계된 마이크로 RNA와 마이크로 RNA의 표적 RNA 쌍에 대한 피어슨 상관계수를 $PCC_{n+1}$로 표시하였다.

[0046] 본 발명에서는 정상 암 세포 샘플에 포함된 식별자(identifiers)를 이용하여 암과 연계된 마이크로 RNA와 표적 RNA 쌍들의 상관관계를 연산한다. 예를 들어, 마이크로 RNA와 표적 RNA 쌍들의 상관관계는 피어슨 상관계수(Pearson correlation coefficient, PCC) 방법에 따라 연산될 수 있다. 마이크로 RNA (X)와, 표적 RNA (Y) 사이의 피어슨 상관계수는 하기 식 (1)에 따라 연산될 수 있다.

$$PCC(X,Y) = \frac{\sum_{k=1}^{n}(X_k - \bar{X})(Y_k - \bar{Y})}{\sqrt{\sum_{k=1}^{n}(X_k - \bar{X})^2}\sqrt{\sum_{k=1}^{n}(Y_k - \bar{Y})^2}} \quad (1)$$

[0047]

[0048] 식 1에서, $n$은 샘플의 개수, $X_k$는 샘플 $k$에서 마이크로 RNA인 X의 counts per million (CPM) 값; $\bar{X}$는 $n$개의 샘플에서 miRNA인 X의 평균 CPM 값; $Y_k$는 샘플 $k$에서 표적 RNA인 Y의 CPM 값; $\bar{Y}$는 $n$개의 샘플에서 표적 RNA인 Y의 평균 CPM 값이다.

[0049] 예를 들어, 상관관계 연산 수단(230)은 정상 샘플 데이터세트에서 피어슨 상관계수 및 암 샘플 데이터세트가 추가된 데이터세트에서 교란된 피어슨 상관계수를 각각 식 (1)을 이용하여 연산할 수 있다. 상관관계 연산 수단(230)은, 샘플 분류 수단(220)에 의해 분류된 전이가 없는 샘플 데이터세트, 림프절 전이만 발생한 샘플 데이터세트, 원격 전이만 발생한 샘플 데이터세트 및 림프절 전이와 원격 전이가 모두 발생한 샘플 데이터세트를 각각 별개로 정상 샘플 데이터세트에 추가하여, 마이크로 RNA와 표적 RNA 쌍 사이의 교란된 상관관계, 예를 들어 식 (1)을 이용한 피어슨 상관계수($PCC_{n+1}$)를 연산할 수 있다.

[0050] 상관관계 필터링 수단(240)은 상관관계 연산 수단(230)에 의해 연산된, 정상 샘플 데이터세트에서 연산된 상관관계, 예를 들어 피어슨 상관계수($PCC_n$)와, 정상 샘플 데이터세트에 적어도 하나의 암 샘플 데이터세트가 추가된 데이터세트에서 연산된 교란된 상관관계, 예를 들어 교란된 피어슨 상관계수($PCC_{n+1}$) 사이의 차이 값을 이용하여, 전이 예측 모듈을 학습시키기에 적절한 마이크로 RNA-표적 RNA 쌍만을 선별한다.

[0051] 각각의 종양 샘플에서 마이크로 RNA와 표적 RNA 쌍에 대한 차이 값인 ΔPCC는 암 샘플 데이터세트가 추가된 데이터세트에서 연산된 교란된 상관관계와, 정상 샘플 데이터세트에서 연산된 상관관계 사이의 차이를 의미한다. 예를 들어, 상관관계 연산 수단(210)에서 정상 샘플 데이터세트 및 암 샘플이 추가된 데이터세트에서 각각 피어슨 상관계수를 이용하여, 마이크로 RNA-표적 RNA 쌍의 상관관계를 연산한 경우, 교란된 상관관계와 정상 샘플 데이터세트에서의 상관관계의 차이(ΔPCC)는 식 (2)를 이용하여 연산될 수 있다.

$$\Delta PCC(X,Y) = PCC_{n+1}(X,Y) - PCC_n(X,Y) \quad (2)$$

[0052]

[0053] 식 (2)에서, $PCC_n$은 n개의 정상 샘플을 사용하여 연산된 PCC이고, $PCC_{n+1}$은 1개의 종양 샘플에 의해 교란된(perturbed) n개의 정상 샘플을 사용하여 연산된 PCC를 나타낸다.

[0054] 상관관계 필터링 수단(240)은, 정상 샘플 데이터세트로부터 연산된 상관관계와, 교란된 상관관계 사이의 차이(ΔPCC)를 이용하여, 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 중에서 각각 림프절 전이에 특이적인 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍 및 원격 전이에 특이적인 마이크로 RNA와 상기 마이크로 RNA의 표적 RNA 쌍을 선별한다. 상이한 암 샘플 데이터세트 및 전이 유무에 따라 상이하게 분류되는 마이크로 RNA-표적 RNA 사이의 상호작용을 활용하여, 림프절 전이에 특이적인 선별된 마이크로 RNA-표적 RNA 쌍과, 원격 전이에 특이적인 선별된 마이클 RNA-표적 RNA 쌍이 달라진다.

[0055] 필요한 경우, 상관관계 필터링 수단(240)은 ΔPCC 값에 기초하여, 샘플의 양성 및 음성 그룹 사이에서 Wilcoxon 검정을 수행하고 신뢰성 있는 마이크로 RNA-표적 RNA 쌍만을 선택할 수 있다(예를 들어, p-value가 0.1 미만, 0.05 미만 또는 0.01 미만). 아울러, 선별된 림프절 전이에 특이적인 마이크로 RNA-표적 RNA 쌍 및 선별된 원격 전이에 특이적인 마이크로 RNA-표적 RNA 쌍에 대한 차원 감소 분석, 예를 들어 주성분분석(principal component analysis, PCA)을 수행하여 특징인 마이크로 RNA-표적 RNA 상의 개수를 줄일 수 있다.

[0056] 선택적으로, 상관관계 필터링 수단(240)은 양성 데이터세트 및 음성 데이터세트에 공통적으로 존재하는 마이크로 RNA-표적 RNA 쌍을 제거할 수도 있다. 다른 실시형태에서, |PCC(miRNA, RNA)| 값이 너무 낮은 경우(예를 들어, 0.5 미만, 0.4 미만, 또는 0.3 미만)인 마이크로 RNA-표적 RNA 쌍은 상관관계 필터링 수단(240)에 의해 제거될 수 있다.

[0057] 예측 모델 구축 수단(250)은 상관관계 필터링 수단(240)에 의해 선별된, 림프절 전이에 특이적인 마이크로 RNA-표적 RNA 쌍과, 원격 전이에 특이적인 마이크로 RNA-표적 RNA 쌍 사이의 상관관계에 기초한 기계 학습을 수행하여, 각각 암 샘플에 대한 림프절 전이 예측 모델($M_{LNM}$) 및 원격 전이 예측 모델($M_{DM}$)을 생성한다.

[0058] 림프절 전이 예측 모델($M_{LMN}$) 및 원격 전이 예측 모델($M_{DM}$)을 생성, 구축하기 위하여, 선별된 마이크로 RNA-표적 RNA 쌍으로부터 기계 학습을 수행하기 위한 기본 학습자는 특별히 한정되지 않는다. 일례로, 서포터 벡터 머신(support vector machine, SVM) 및 로지스틱 회귀(logistic regression, LR) 중에서 적어도 하나를 기본 학습자로 기계 학습자로 사용할 수 있다. 예시적인 실시형태에서, SVM 및 LR을 모두 기본 학습자로 활용한 스태킹(stacking)에 의한 앙상블(ensemble) 학습을 수행하여, 암 전이 예측 성능을 향상시킬 수 있다.

[0059] 림프절 전이에서 관찰되는 유전자 발현은 흔히 원격 전이에서 관찰되는 유전자 발현과 다르기 때문에, 단일 모델에 의하여 림프절 전이와 원격 전이를 모두 예측하기는 어렵다. 따라서, 본 발명은 림프절 전이 예측 모델(도 3에서 $M_{LNM}$)과 원격 전이 예측 모델(도 3에서 $M_{DM}$)을 사용한다. 림프절 전이 예측 모델과 원격 예측 모델을 구축한 뒤, 입력 값으로서 암 샘플을 사용하여 림프절 전이 예측 모델($M_{LNM}$)과 원격 전이 예측 모델($M_{DM}$)을 실행시킨다.

[0060] 림프절 전이 예측 모델($M_{LNM}$)과 원격 전이 예측 모델($M_{DM}$)에서 해당 암 샘플이 모두 음성(negative)으로 분류되면, 해당 샘플은 전이가 일어나지 않은 것으로 분류된다. 해당 암 샘플이 림프절 전이 예측 모델($M_{LMN}$)에서는 양성(positive)이고, 원격 전이 모델($M_{DM}$)에서는 음성이라고 분류된다면, 해당 암 샘플에 대해서는 림프절 전이가 있는 것으로 예측된다. 마찬가지로, 해당 암 샘플이 림프절 전이 예측 모델($M_{LMN}$)에서는 음성이고, 원격 전이 예측 모델($M_{DM}$)에서는 양성으로 분류된다면, 해당 암 샘플에 대해서는 원격 전이만이 예상된다. 2개의 모델이 모두 해당 샘플을 모두 양성이라고 분류하면, 해당 샘플에 대해서는 림프절 전이 및 원격 전이가 예상될 것이다.

[0061] 다른 실시형태에서, 림프절 전이 예측 모델($M_{LMN}$)은 림프절 전이만 있는 샘플, 또는 림프절 전이 및 원격 전이가 발생하는 샘플은 림프절 전이에 대하여 양성으로 분류하고, 전이가 없는 샘플은 림프절 전이에 대하여 음성으로 분류할 수 있다. 원격 전이 예측 모델($M_{DM}$)은 원격 전이만 있는 샘플, 또는 림프절 전이 및 원격 전이가 발생하는 샘플은 원격 전이에 대하여 양성으로 분류하고, 전이가 없는 샘플 또는 림프절 전이만 있는 샘플은 원격 전이에 대하여 음성으로 분류할 수 있다.

[0062] 도 3을 참조하면, 림프절 전이 예측 모델 및 원격 전이 예측 모델을 구축하는 방법은, 암 샘플 데이터세트를 전이 유무에 따라 분류하고, 암과 연계된 마이크로 RNA와 마이크로 RNA의 표적 RNA를 분류하는 단계(S210 단계),

암과 연계된 마이크로 RNA-표적 RNA 쌍 사이의 상관관계를 연산하는 단계(S220 단계), 정상 샘플 데이터세트에서 연산된 상관관계와, 암 샘플이 추가된 데이터세트에서 연산된 교란된 상관관계 사이의 차이 값을 활용하여 림프절 전이에 특이적인 마이크로 RNA-표적 RNA 쌍과, 원격 전이에 특이적인 마이크로 RNA-표적 RNA 쌍을 선별하는 단계(S230 단계) 및 선별된 림프절 전이 및 원격 전이에 특이적인 마이크로 RNA-표적 RNA 쌍 사이의 상관관계에 기초한 기계 학습을 수행하여 암 샘플에 대한 림프절 전이 예측 모델 및 원격 전이 예측 모델을 각각 생성하는 단계(S240 단계)를 포함할 수 있다.

[0063] 암 환자의 샘플을 분류하는 단계(S210 단계)는 샘플 분류 수단(220)에 의해 수행될 수 있다. 암 환자의 샘플을 분류하는 단계(S210)에서, 암 샘플 데이터세트를 전이 여부에 따라 전이가 일어나지 않은 샘플 데이터세트, 림프절 전이만 발생한 샘플 데이터세트, 원격 전이만 발생한 샘플 데이터세트, 및 림프절 전이와 원격 전이가 모두 발생한 샘플 데이터세트로 분류하고, 암과 연계된 마이크로 RNA 및 마이크로 RNA의 표적 RNA(예를 들어, mRNA, lncRNA, 위유전자 중에서 적어도 하나)를 분류한다.

[0064] 마이크로 RNA-표적 RNA 쌍의 상관관계를 연산하는 단계(S220)는 상관관계 연산 수단(230)에 의해 수행될 수 있다. 마이크로 RNA-표적 RNA 쌍의 상관관계를 연산하는 단계(S220)에서 정상 샘플 데이터세트에서 샘플 분류 단계(S210 단계)에서 샘플 분류 수단(220)에 의해 분류된 암과 연계된 마이크로 RNA-표적 RNA 쌍 각각에 대한 상관관계와, 정상 샘플 데이터세트에 암 샘플 데이터세트를 추가한 데이터세트에서 샘플 분류 수단(220)에 의해 분류된 암과 연계된 마이크로 RNA-표적 RNA 쌍 각각에 대한 교란된 상관관계가 연산된다. 일례로, 마이크로 RNA-표적 RNA 쌍의 상관관계는 식 (1)로 연산되는 피어슨 상관관계일 수 있다.

[0065] 마이크로 RNA-표적 RNA 쌍을 필터링(선별)하는 단계(S230 단계)는 상관관계 필터링 수단(240)에 의해 수행될 수 있다. 마이크로 RNA-표적 RNA 쌍을 선별하는 단계(S230 단계)에서, 정상 샘플 데이터세로부터 연산된 상관관계와, 교란된 상관관계 사이의 차이 값이 연산된다. 예를 들어, 이들 상관관계 사이의 차이 값은 식 (2)를 이용하여 연산될 수 있다. 이어서, 마이크로 RNA-표적 RNA 쌍을 선별하는 단계(S230 단계)에서는 이러한 상관관계 사이의 차이 값을 이용하여 림프절 전이에 특이적인 마이크로 RNA-표적 RNA 쌍과, 원격 전이에 특이적인 마이크로 RNA-표적 RNA 쌍을 선별한다. 필요한 경우, 선별하는 단계(S230 단계)에서 선별된 마이크로 RNA-표적 RNA 쌍에 대한 차원 감소 분석이 더욱 수행되고/되거나, 신뢰성이 없는 마이크로 RNA-표적 RNA 쌍이 제거될 수 있다.

[0066] 암 전이 예측 모듈을 구축하는 단계(S240 단계)는 예측 모델 구축 수단(250)에 의해 수행될 수 있다. 암 전이 예측 모듈을 구축하는 단계(S240 단계)에서, 선별 단계(S230 단계)에서 선별된 림프절 전이에 특이적인 마이크로 RNA-표적 RNA 쌍 및 원격 전이에 특이적인 마이크로 RNA-표적 RNA 쌍 사이의 상관관계에 기초한 기계 학습을 수행하여, 암 샘플에 대한 림프절 전이 예측 모델 및 원격 전이 예측 모델을 각각 생성한다. 일례로, 기계 학습과 관련해서, SVM 및/또는 LR이 기본 학습자로 활용될 수 있으나, 이에 한정되지 않는다.

[0067] 이하, 예시적인 실시형태를 통하여 본 발명을 설명하지만, 본 발명이 하기 실시형태에 기재된 기술사상으로 한정되지 않는다.

[0068] 실시형태

[0069] [데이터 수집 및 샘플 분류]

[0070] 암유전체지도(The Cancer Genome Atlas, TCGA)는 총 33가지 형태의 암 데이터세트가 있다. 이 중에서 최소 10개의 정상 샘플을 가지고 림프절 전이가 일어나 최소한 50개의 종양 샘플을 가지는 데이터세트를 선택하였다. 33가지 종류의 암 데이터세트 중에서, 방광 요로상피세포암(urothelial bladder carcinoma, BLCA), 침습 유방암(breast invasive carcinoma, BRCA), 대장선암(colon adenocarcinoma, COAD), 식도암(esophageal carcinoma, ESCA), 두경부 편평세포암(head-neck squamous cell carcinoma, HNSC), 폐선암(lung adenocarcinoma, LUAD), 폐 편평세포암 (lung squamous cell carcinoma, LUSC), 전립선암(prostate adenocarcinoma, PRAD), 위선암(stomach adenocarcinoma, STAD), 갑상선암(thyroid carcinoma, THCA)인 10 종류의 암 데이터세트가 기준을 충족하였다.

[0071] 이어서, 임상 보완 데이터의 종양(Tumor), 림프절(Node, N) 및 전이(Metastasis, M) (TNM) 단계 인덱스에 기초하여 종양 샘플을 아래 4가지로 분류하였다.

[0072] (1) 전이가 없는 샘플 (nonM): 1-4 단계의 T, 0 단계의 N, 0 단계의 M;

[0073] (2) 림프절 전이만 있는 샘플(LNM_only): 1-4 단계의 T, 1-3 단계의 N, 0 단계의 M;

[0074] (3) 원격 전이만 있는 샘플(DM_only): 1-4 단계의 T, 0 단계의 N, 1 단계의 M;

[0075] (4) 림프절 전이 및 원격 전이 모두 있는 샘플(LNM&DM): 1-4 단계의 T, 1-3 단계의 N, 1 단계의 M.

[0076] 하기 표 1은 10가지 형태의 암 형태 중에서 각각의 암 타입의 정상 샘플과 종양 샘플의 수를 나타낸다.

표 1

[0077] 암 타입에서 정상 샘플과 종양 샘플 개수

| 암 | 정상 | nonM | LNM_only | DM_only | LNM&DM |
|---|---|---|---|---|---|
| BLCA | 19 | 118 | 44 | 0 | 7 |
| BRCA | 113 | 450 | 449 | 1 | 18 |
| COAD | 41 | 228 | 105 | 9 | 55 |
| ESCA | 11 | 56 | 64 | 2 | 6 |
| HNSC | 44 | 81 | 98 | 0 | 1 |
| LUAD | 59 | 219 | 124 | 11 | 11 |
| LUSC | 49 | 258 | 149 | 3 | 3 |
| PRAD | 52 | 316 | 75 | 1 | 1 |
| STAD | 32 | 103 | 210 | 2 | 23 |
| THCA | 59 | 145 | 127 | 3 | 4 |

[0078] [마이크로 RNA 및 표적 RNA 추출 및 분류]

[0079] 유전자 발현 데이터 및 임상 보완 데이터로서, TCGA의 Genomic Data Commons (GDC) 데이터 포털(https://portal.gdc.cancer.gov/)로부터 얻은 최신 데이터(Data release 33.1 - 2022년 5월 31일)를 사용하여, RNA-seq 발현 데이터로부터 유전자의 raw read counts를 추출하였다. 평균 raw read count가 1 미만인 유전자를 제거하여 낮게 발현되는 유전자를 필터링하였다. 잔류하는 유전자들은 Ensemble project로부터 얻어진 annotation file을 사용하여 4개의 바이오타입(miRNAs, lncRNAs, mRNA, pseudogenes)으로 분류하였다. 하기 표 2는 각각의 암 타입에서 miRNA, lncRNA, mRNA 및 pseudogene의 개수를 나타낸다. 이어서, R package edgeR에서 TMM(trimmed mean of M values) 방법을 사용하여, 유전자의 raw read counts를 CPM(counts per million)으로 정규화하였다.

표 2

[0080] 각각의 암 타입에서 암 연계 miRNA, lncRNA, mRNA 및 pseudogene의 개수

| 암 | miRNAs | lncRNAs | mRNA | pseudogenes |
|---|---|---|---|---|
| BLCA | 143 | 9,612 | 18,038 | 4,994 |
| BRCA | 150 | 10,070 | 18,035 | 5,380 |
| COAD | 144 | 8,477 | 17,515 | 5,102 |
| ESCA | 418 | 12,588 | 18,658 | 8,713 |
| HNSC | 88 | 8,563 | 17,912 | 4,493 |
| LUAD | 182 | 10,291 | 18,037 | 5,845 |
| LUSC | 147 | 10,206 | 18,152 | 5,507 |
| PRAD | 126 | 8,764 | 17,731 | 4,686 |
| STAD | 345 | 12,472 | 18,657 | 8,700 |
| THCA | 140 | 8,256 | 17,487 | 4,610 |

[0081] [miRNA-RNA 상호작용 연산 및 특징 선별]

[0082] miRNA와 표적 RNAs (lncRNAs, mRNAs, pseudogenes) 사이의 초기 쌍을 얻기 위하여, 모든 miRNA (X)와 표적 RNA (Y) 쌍에 대한 피어슨 상관계수(PCC)를 식 (1)에 따라 연산하였다.

[0083] 전술한 바와 같이, 본 발명은 암 세포의 림프절 전이 예측 모델($M_{LNM}$)과 원격 전이 예측 모델($M_{DM}$)로 구성될 수 있다. 2개의 모델은 모두 2진 분류자(binary classifier)이며, 입력 값으로 종양 샘플을 취한다. 하지만, 2개의 모델은 다음과 같이 상이한 양성 및 음성 데이터세트를 가지도록 구축될 수 있다.

[0084] 림프절 전이 예측 모델($M_{LNM}$)에서 림프절 전이만 있는 샘플(LNM_only) 또는 림프절 전이 및 원격 전이 모두 있는 샘플(LNM&DM)은 양성으로 분류되고, 전이가 없는 샘플(nonM)은 음성으로 분류된다. 반면, 원격 전이 예측 모델($M_{DM}$)에서는 원격 전이만 있는 샘플(DM_only) 또는 림프절 전이 및 원격 전이 모두 있는 샘플(LNM&DM)은 모두 양성으로 분류되지만, 전이가 없는 샘플(nonM) 또는 림프절 전이만 있는 샘플(LNM_only)은 음성으로 분류된다. 즉, 본 발명의 예시적인 실시형태에 따르면, 림프절 전이를 예측할 때, 양성 데이터세트 내에 림프절 전이 및 원격 전이 모두 있는 샘플(LNM&DM)을 포함시킨다. 림프절 전이 및 원격 전이 모두 있는 샘플(LNM&DM)은 또한 원격 전이를 예측하는 경우에도 양성 데이터세트에 포함된다.

[0085] 2가지 예측 모델에 대하여 miRNA-표적 RNA 쌍은 다르게 선택되었다. 필요에 따라, 양성 데이터세트 및 음성 데이터세트에 공통적인 miRNA-표적 RNA 쌍을 제거하고, 양성 데이터세트와 음성 데이터세트에서 miRNA-표적 RNA 쌍에 대한 PCCs를 별도로 연산하였다. 한편, |PCC(miRNA, RNA)|이 0.4 미만인 miRNA-RNA 쌍인 경우의 상관관계는 예측에 사용하기에 충분히 크지 않기 때문에 모두 제거하였다.

[0086] 각각의 종양 샘플에서 miRNA X와 표적 RNA Y 쌍에 대한 차이 값인 ΔPCC는 식 (2)에 따라 연산하였다. 결국, 각각의 종양 샘플에 대하여, 단일 종양 샘플에 의해 교란된 PCC($PCC_{n+1}$)와, 정상 샘플에서 연산된 PCC($PCC_n$) 사이에서 PCC 값의 차이를 얻을 수 있다.

[0087] 이어서, 종양 샘플의 양성 및 음성 그룹 중에서, ΔPCC 값에 기초하여, 2개의 그룹 사이에서 Wilcoxon 검정(Wilcoxon test)를 수행하고, Wilcoxon 검정에서 p-value가 0.01 미만인 miRNA-표적 RNA 쌍만을 선택하였다. 주성분분석(principal component analysis, PCA)를 통하여 특징인 miRNA-표적 RNA 쌍의 개수를 더욱 줄였다. 각각의 필터링 과정 후에 잔존하는 특징의 개수를 표 3에 나타낸다.

*표 3*

[0088] 필터링 과정 후에 남은 특징(miRNA-표적 RNA 쌍)의 개수

| 암 | PCC 필터링 후 | Wilcoxon 검정 후 | PCA 후 |
|---|---|---|---|
| 림프절 전이 예측 모델 | | | |
| BLCA | 169,439 | 9,501 | 45 |
| BRCA | 170,673 | 3,619 | 166 |
| COAD | 312,968 | 3,970 | 162 |
| ESCA | 706,722 | 27,312 | 65 |
| HNSC | 82,959 | 2,281 | 58 |
| LUAD | 320,323 | 13,891 | 137 |
| LUSC | 43,340 | 1,296 | 83 |
| PRAD | 78,722 | 5,036 | 150 |
| STAD | 230,038 | 8,136 | 120 |
| THCA | 124,722 | 12,738 | 102 |
| 원격 전이 예측 모델 | | | |
| BRCA | 572,862 | 19,634 | 134 |
| COAD | 273,660 | 4,968 | 112 |
| LUAD | 863,846 | 20,632 | 55 |
| STAD | 1,222,396 | 43,240 | 58 |

[0089] [예측 모델 구축]

[0090] 특징으로서 miRNA-표적 RNA 쌍의 ΔPCC 값을 사용해서, 기계 학습을 통하여 림프절 전이 예측 모델 및 원격 전이 예측 모델을 다음과 같이 구축하였다. 2개의 모델은 기본 학습자로서 지도형 기계 학습 모델(알고리즘)인 서포트 벡터 머신(support vector machine, SVM)과 로지스틱 회귀(logistic regression, LR)를 갖는 앙상블 학습자(ensemble learner)이다. 제 2의 학습자로서 LR을 사용한 스태킹(stacking)에 의하여 기본 학습자를 조합하면, 예측의 정확성을 높일 수 있다.

[0091] 데이터세트들은 무작위로 7:3의 비율로 훈련 및 테스트 데이터세트로 각각 분할되었다(partitioned). 5-fold 교차 검증을 사용한 격자 검색(grid search)에 의한 예측 모델의 하이퍼파라미터(hyperparameters)를 최적화하기 위해 훈련 데이터를 사용하였다. 무작위 데이터 분할과, TCGA 데이터세트가 작고 불균형 하다는 점에 기인하여,

모델을 평가할 때 데이터 분할, 훈련 및 테스트의 전체 과정을 10회 반복하였다.

[0092] [결과 및 평가]

[0093] A. 림프절 전이 및 원격 전이 예측

[0094] 표 4는 림프절 전이 예측 모델 $M_{LNM}$의 성능 중에서 민감도(sensitivity), 특이도(specificity), 정확도(accuracy), 양성예측도(positive predictive value, PPV), 음성예측도(negative predictive value, NPV), AUC (Area Under the ROC Curve)와 같은 성능을 나타낸다. 림프절 전이 예측 모델 $M_{LNM}$의 AUC 값은 5-fold 교차 검증에서 0.754 내지 0.983, 독립 테스트에서 0.732 내지 0.961이었으며, ESCA에서 림프절 전이 예측 모델 $M_{LNM}$의 성능은 최고였다.

표 4

[0095] 5-fold 교차 검증 및 독립 테스트에서 $M_{LNM}$의 성능

| 암 | 평가 | 민감도 | 특이도 | 정확도 | PPV | NPV | AUC |
|---|---|---|---|---|---|---|---|
| BLCA | 교차 검증 | 0.941 | 0.971 | 0.962 | 0.937 | 0.974 | 0.956 |
| | 독립 테스트 | 0.906 | 0.969 | 0.951 | 0.933 | 0.962 | 0.938 |
| BRCA | 교차 검증 | 0.763 | 0.744 | 0.754 | 0.756 | 0.752 | 0.754 |
| | 독립 테스트 | 0.745 | 0.720 | 0.732 | 0.736 | 0.730 | 0.732 |
| COAD | 교차 검증 | 0.937 | 0.938 | 0.938 | 0.915 | 0.955 | 0.938 |
| | 독립 테스트 | 0.910 | 0.962 | 0.941 | 0.945 | 0.939 | 0.936 |
| ESCA | 교차 검증 | 0.975 | 0.992 | 0.982 | 0.993 | 0.970 | 0.983 |
| | 독립 테스트 | 0.952 | 0.970 | 0.960 | 0.978 | 0.948 | 0.961 |
| HNSC | 교차 검증 | 0.957 | 0.936 | 0.948 | 0.948 | 0.948 | 0.947 |
| | 독립 테스트 | 0.940 | 0.908 | 0.925 | 0.930 | 0.926 | 0.924 |
| LUAD | 교차 검증 | 0.767 | 0.901 | 0.850 | 0.827 | 0.863 | 0.834 |
| | 독립 테스트 | 0.678 | 0.896 | 0.813 | 0.804 | 0.818 | 0.787 |
| LUSC | 교차 검증 | 0.785 | 0.915 | 0.867 | 0.845 | 0.879 | 0.850 |
| | 독립 테스트 | 0.767 | 0.914 | 0.859 | 0.843 | 0.868 | 0.840 |
| PRAD | 교차 검증 | 0.711 | 0.963 | 0.914 | 0.824 | 0.933 | 0.837 |
| | 독립 테스트 | 0.660 | 0.969 | 0.909 | 0.854 | 0.922 | 0.815 |
| STAD | 교차 검증 | 0.960 | 0.922 | 0.948 | 0.965 | 0.912 | 0.941 |
| | 독립 테스트 | 0.927 | 0.867 | 0.908 | 0.941 | 0.849 | 0.897 |
| THCA | 교차 검증 | 0.832 | 0.837 | 0.835 | 0.824 | 0.846 | 0.835 |
| | 독립 테스트 | 0.782 | 0.822 | 0.803 | 0.800 | 0.812 | 0.802 |

[0096] 원격 전이 샘플 데이터세트 개수의 한계로 인하여, 원격 전이 예측 모델은 4 종류의 암에 대해서만 수행하였다. BRCA에 대해서는 19개의 원격 전이 샘플, COAD에 대해서는 64개의 원격 전이 샘플, LUAD에 대해서는 22개의 원격 전이 샘플, STAD에 대해서는 25개의 원격 전이 샘플만 있었다. 표 5는 원격 전이 예측 모델 $M_{DM}$의 민감도, 특이도, 정확도, PPV, NPV, AUC를 나타낸다. 원격 전이 예측 모델 $M_{DM}$의 AUC 값은 5-fold 교차 검증에서 0.722 내지 0.949, 독립 테스트에서 0.733 내지 0.907이었다. 원격 전이 예측 모델의 성능은 림프절 전이 예측 모델에 비해 약간 낮았으나, 여전히 높은 수준을 유지하였다. 원격 전이를 예측하기 어렵다는 점을 고려할 때, 본 실시 형태에서 원격 전이 예측 모델의 성능은 여전히 우수하였다.

표 5

[0097] 5-fold 교차 검증 및 독립 테스트에서 $M_{DM}$의 성능

| 암 | 평가 | 민감도 | 특이도 | 정확도 | PPV | NPV | AUC |
|---|---|---|---|---|---|---|---|
| BRCA | 교차 검증 | 0.900 | 0.998 | 0.996 | 0.916 | 0.997 | 0.949 |
| | 독립 테스트 | 0.816 | 0.997 | 0.993 | 0.881 | 0.995 | 0.907 |
| COAD | 교차 검증 | 0.882 | 0.982 | 0.966 | 0.908 | 0.977 | 0.932 |
| | 독립 테스트 | 0.810 | 0.968 | 0.943 | 0.842 | 0.965 | 0.889 |
| LUAD | 교차 검증 | 0.446 | 0.998 | 0.966 | 0.961 | 0.967 | 0.722 |
| | 독립 테스트 | 0.471 | 0.996 | 0.962 | 0.905 | 0.966 | 0.733 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| STAD | 교차 검증 | 0.694 | 0.997 | 0.975 | 0.965 | 0.977 | 0.845 |
| | 독립 테스트 | 0.712 | 0.994 | 0.972 | 0.924 | 0.977 | 0.853 |

[0098] 전이 여부가 알려지지 않은 환자의 샘플 데이터세트로부터, 본 실시형태에 따른 예측 모델은 림프절 전이 및/또는 원격 전이를 예측할 수 있다. 일례로, 도 4는 COAD 샘플에 대한 3가지 케이스를 나타낸다. COAD 샘플 TCGA-CA-5254는 림프절 전이 예측 모델 $M_{LNM}$ 및 원격 전이 예측 모델 $M_{DM}$에서 음성으로 예측되었으며, 실제로 이 샘플에서 전이는 발견되지 않았다. 2번째 COAD 샘플 TCGA-AA-3994는 림프절 전이 예측 모델에서는 양성으로, 원격 전이 예측 모델에서는 음성으로 분류되었으며, 실제로 이 샘플에서는 림프절 전이만이 발생하였다. 3번째 COAD 샘플 TCGA-AA-3684는 림프절 전이 예측 모델 및 원격 전이 예측 모델에서 모두 양성으로 분류되었으며, 해당 샘플에서는 림프절 전이 및 원격 전이가 모두 발생하였다.

[0099] B. 다른 방법과의 비교

[0100] 전술한 바와 같이 본 실시형태에서는 특징으로서 miRNA-표적 RNA 쌍 사이의 ΔPCC를 사용하였다. 전이를 예측하기 위한 대부분의 학습 기반 방법은 특징으로서 유전자 발현을 이용한다. 따라서, 본 실시형태에 따라 miRNA-표적 RNA 쌍의 ΔPCC를 특징으로 하는 예측 방법(ΔPCC); miRNA-표적 RNA 쌍과 관련된 유전자 발현을 특징으로 하는 예측 방법(expression); 191가지의 전이-예측 유전자의 발현을 특징으로 하는 예측 방법(expression191)을 각각 비교하였다.

[0101] 공정한 평가를 위하여, 동일한 방식으로 상이한 특징을 갖는 3개의 모델을 평가하였다. 즉, 데이터세트를 무작위로 7:3의 비율로 훈련 데이터세트와 테스트 데이터세트로 분할하고, 5-fold 교차 검증을 사용한 격자 검색에 의하여 예측 모델의 하이퍼파라미터를 최적화하기 위해 훈련 데이터를 훈련 데이터세트를 사용하였다. 모델 평가를 위하여 데이터 분할, 훈련 및 테스트의 전체 과정을 10회 반복하였다.

[0102] 도 5는 림프절 전이의 예측에서 3가지 모델의 평균 AUC 값을 비교한 그래프이다. 교차 검증 및 독립 테스트 모두에서 10가지 암 타입 모두에서 본 실시형태에 따라 생성한 miRNA-표적 RNA 쌍의 ΔPCC 방법이 다른 특징을 이용한 방법보다 우수하였다. 또한, miRNA-표적 RNA 쌍과 연계된 유전자 발현을 이용하면, 191개의 전이-예측 유전자를 이용하는 것보다 성능이 우수하였다.

[0103] 동일한 결과가 원격 전이 예측에서 3가지 모델의 평균 AUC 값을 비교한 도 6의 그래프에서도 관찰되었다. 교차 검증 및 독립 테스트 모두에서 4가지 암 타입 모두에서 본 실시형태에 따라 사용한 miRNA-표적 RNA 쌍의 ΔPCC 방법이 다른 2가지 모델보다 우수하였다. 흥미롭게도 191가지 전이-예측 유전자는 림프절 전이 및 원격 전이를 예측하지 못하였다. 결국, 림프절 전이와 원격 전이를 예측할 때, 유전자 발현보다는 본 실시형태에서 따라 miRNA-표적 RNA 쌍 상호작용의 ΔPCC가 보다 강력하다는 것을 보여준다.

[0104] 상기에서는 본 발명의 예시적인 실시형태 및 실시예에 기초하여 본 발명을 설명하였으나, 본 발명이 상기 실시형태 및 실시예에 기재된 기술사상으로 한정되는 것은 아니다. 오히려 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자라면 전술한 실시형태 및 실시예를 토대로 다양한 변형과 변경을 용이하게 추고할 수 있다. 하지만, 이러한 변형과 변경은 모두 본 발명의 권리범위에 속한다는 점은, 첨부하는 청구범위에서 분명하다.

## *부호의 설명*

[0105] 100: 컴퓨터 110: 입력부

120: 출력부 130: 디스플레이

140: 메모리 150: 구동/응용 프로그램

160: 제어부

200: 기록 매체

210: 림프절 전이 및 원격 전이 예측 프로그램

220: 샘플 분류 수단 230: 상관관계 연산 수단

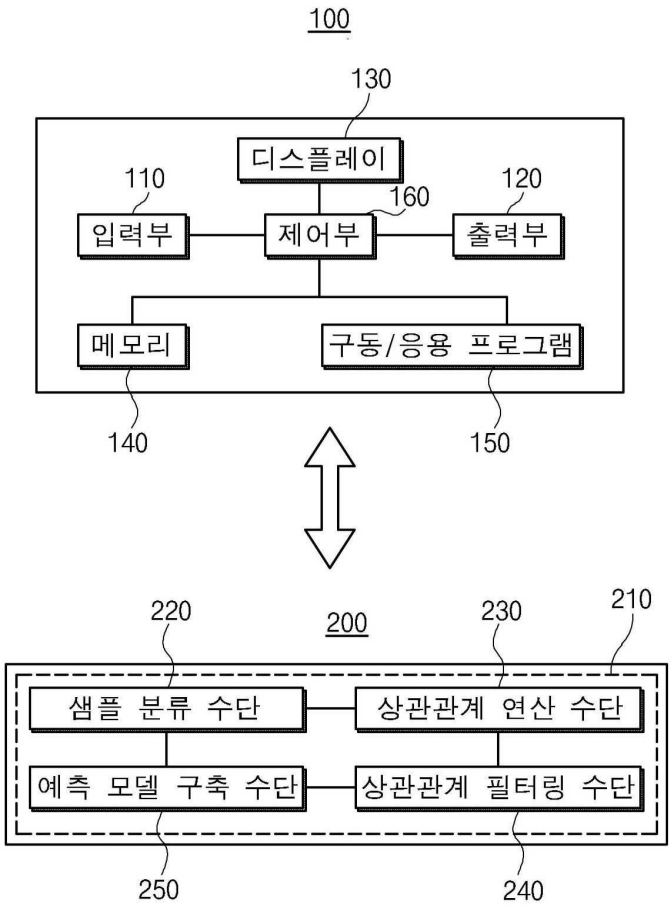
240: 상관관계 필터링 수단 250: 예측 모델 구축 수단

$M_{LNM}$: 림프절 전이 예측 모델　　$M_{DM}$: 원격 전이 예측 모델

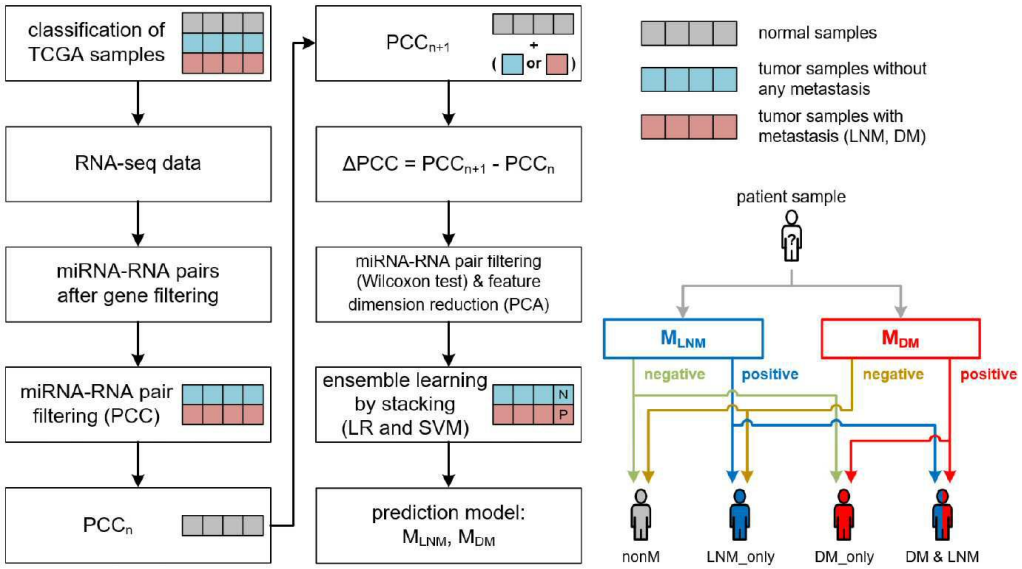
TCGA: 암유전체지도　　PCC: 피어슨 상관계수

## 도면

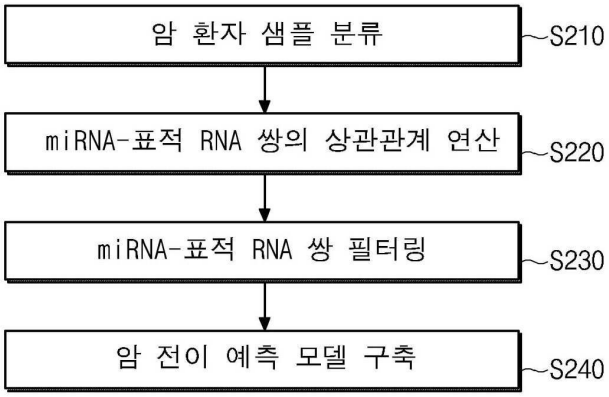
### 도면1



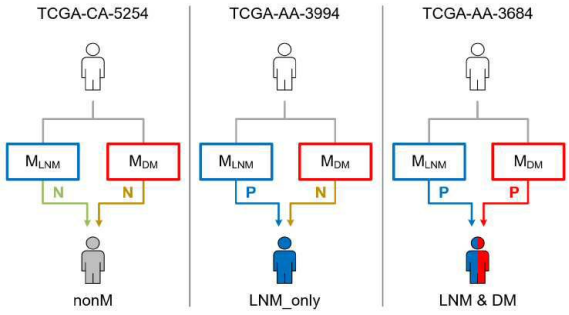

### 도면2



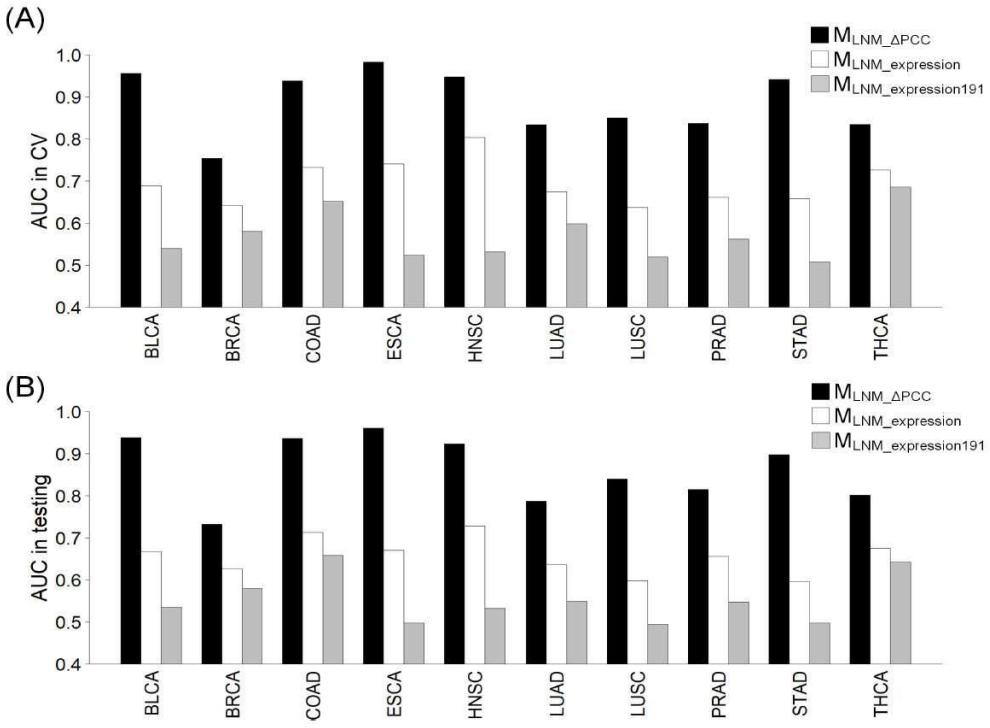
도면3


암 환자 샘플 분류
S210
miRNA-표적 RNA 쌍의 상관관계 연산
S220
miRNA-표적 RNA 쌍 필터링
S230
암 전이 예측 모델 구축
S240

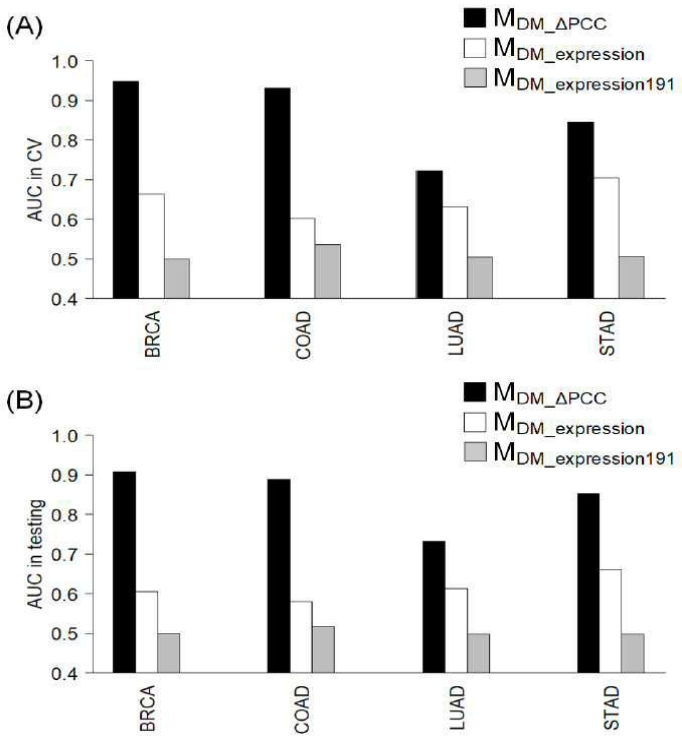
도면4


TCGA-CA-5254
TCGA-AA-3994
TCGA-AA-3684
MLNM
MDM
N
P
nonM
LNM_only
LNM & DM

도면5


(A)
(B)
AUC in CV
AUC in testing
MLNM_ΔPCC
MLNM_expression
MLNM_expression191
1.0
0.9
0.8
0.7
0.6
0.5
0.4
BLCA
BRCA
COAD
ESCA
HNSC
LUAD
LUSC
PRAD
STAD
THCA

도면6


(A)
$M_{DM_\Delta PCC}$
$M_{DM_expression}$
$M_{DM_expression191}$
AUC in CV
1.0
0.9
0.8
0.7
0.6
0.5
0.4
BRCA
COAD
LUAD
STAD
(B)
$M_{DM_\Delta PCC}$
$M_{DM_expression}$
$M_{DM_expression191}$
AUC in testing
1.0
0.9
0.8
0.7
0.6
0.5
0.4
BRCA
COAD
LUAD
STAD