(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0034351
(43) 공개일자 2022년03월18일

(51) 국제특허분류(Int. Cl.)
G16B 40/20 (2019.01) *G16B 20/20* (2019.01)
G16B 30/10 (2019.01) *G16B 50/00* (2019.01)
(52) CPC특허분류
G16B 40/20 (2019.02)
G16B 20/20 (2019.02)
(21) 출원번호 10-2020-0116606
(22) 출원일자 2020년09월11일
심사청구일자 2020년09월11일

(71) 출원인
인천대학교 산학협력단
인천광역시 연수구 아카데미로 119 (송도동)
(72) 발명자
안재균
경기도 과천시 별양로 66-11 래미안슈르 347동 1201호
정희원
인천광역시 연수구 솔샘로43번길 15 성호아파트 103동 605호
(74) 대리인
김효성

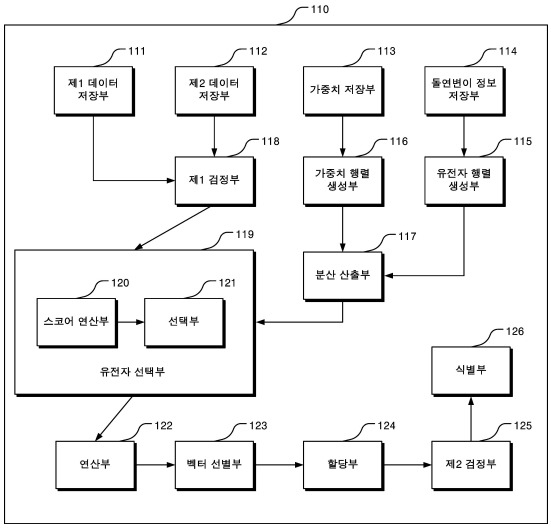
전체 청구항 수 : 총 8 항

(54) 발명의 명칭 **암 유발 유전자의 식별을 위한 전자 장치 및 그 동작 방법**

(57) 요 약

암 유발 유전자의 식별을 위한 전자 장치 및 그 동작 방법이 개시된다. 본 발명은 암환자와 정상 환자로부터 수집된 유전자 데이터를 기초로 사람의 유전자들 중에서 암의 발생에 영향을 미치는 특정 유전자들만을 식별할 수 있는 전자 장치 및 그 동작 방법에 대한 것이다.

대 표 도 - 도1



(52) CPC특허분류
G16B 30/10 (2019.02)
G16B 50/00 (2019.02)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711112452
과제번호 2019R1A2C3005212
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 개인기초연구(과기정통부)(R&D)
연구과제명 딥러닝을 이용한 간암 표적항암제 내성기전 규명 및 이를 극복할 새로운 표적항암제 탐색
기 여 율 1/1
과제수행기관명 인천대학교
연구기간 2020.03.01 ~ 2021.02.28

명 세 서

청구범위

청구항 1

복수의 암환자들로부터 사전 수집된 n(n은 6이상의 자연수)개의 유전자들 각각의 발현량 데이터인 제1 발현량 데이터가 저장되어 있는 제1 데이터 저장부;

복수의 정상 환자들로부터 사전 수집된 상기 n개의 유전자들 각각의 발현량 데이터인 제2 발현량 데이터가 저장되어 있는 제2 데이터 저장부;

상기 복수의 암환자들 중 어느 한 명인 제1 암환자로부터 사전 측정된 상기 n개의 유전자들 상호 간의 영향력 정도를 표상하는 가중치 - 상기 가중치는 그 값이 클수록 유전자 상호 간의 영향력이 큼을 의미함 - 가 저장되어 있는 가중치 저장부;

상기 n개의 유전자들 중 상기 복수의 암환자들과 상기 복수의 정상 환자들에서 돌연변이가 적어도 1회 이상 발생한 것으로 사전 확인된 적어도 하나의 돌연변이 유전자에 대한 정보가 저장되어 있는 돌연변이 정보 저장부;

상기 n개의 유전자들 각각에 대응되는 행과 상기 n개의 유전자들 각각에 대응되는 열을 갖는 n x n 크기의 행렬에서 상기 적어도 하나의 돌연변이 유전자에 대응되는 행과 열이 만나는 지점의 성분을 1로 할당하고, 나머지 지점의 성분을 0으로 할당함으로써, n x n 크기의 유전자 행렬을 생성하는 유전자 행렬 생성부;

상기 가중치 저장부를 참조하여, 상기 n x n 크기의 행렬에서 상기 n개의 유전자들 상호 간의 행과 열이 만나는 지점에 상기 n개의 유전자들 상호 간의 가중치를 성분으로 할당함으로써, n x n 크기의 가중치 행렬을 생성하는 가중치 행렬 생성부;

상기 유전자 행렬에 대해 상기 가중치 행렬을 t(t는 2이상의 자연수)회 곱하여 n x n 크기의 제1 연산 행렬을 생성하고, 상기 제1 연산 행렬을 구성하는 n개의 행벡터들 각각을 상기 n개의 유전자들 각각에 대응되는 제1 벡터로 지정한 후 상기 n개의 유전자들 각각에 대해, 각 유전자에 대응되는 제1 벡터를 구성하는 n개의 성분들의 분산을 연산함으로써, 상기 n개의 유전자들 각각에 대응되는 분산을 산출하는 분산 산출부;

상기 n개의 유전자들 각각에 대해, 상기 제1 암환자로부터 수집된 제1 발현량 데이터와 상기 복수의 정상 환자들로부터 수집된 제2 발현량 데이터 간의 단일 표본 t-검정(one sample t-test)을 수행함으로써, 상기 n개의 유전자들 각각에 대응되는 제1 t-검정 값을 생성하는 제1 검정부; 및

상기 n개의 유전자들 각각에 대응되는 분산과 제1 t-검정 값을 기초로 상기 n개의 유전자들 중 암에 대한 영향과 돌연변이 유전자에 대한 영향을 크게 받는 것으로 확인되는 k(k는 2이상의 자연수)개의 상위 유전자들과 암에 대한 영향과 돌연변이 유전자에 대한 영향을 작게 받는 것으로 확인되는 k개의 하위 유전자들을 선택하는 유전자 선택부

를 포함하는 암 유발 유전자의 식별을 위한 전자 장치.

청구항 2

제1항에 있어서,

상기 유전자 선택부는

상기 n개의 유전자들 각각에 대응되는 분산과 제1 t-검정 값을 서로 곱하여 상기 n개의 유전자들 각각에 대한 스코어(score)를 연산하는 스코어 연산부; 및

상기 n개의 유전자들 중 상기 스코어가 높은 순으로 상기 k개의 상위 유전자들을 선택하고, 상기 n개의 유전자들 중 상기 스코어가 낮은 순으로 상기 k개의 하위 유전자들을 선택하는 선택부

를 포함하는 암 유발 유전자의 식별을 위한 전자 장치.

청구항 3

제2항에 있어서,

상기 k개의 상위 유전자들과 상기 k개의 하위 유전자들이 선택되면, 상기 n x n 크기의 행렬에서 대각 성분을 모두 1로 할당하여 n x n 크기의 초기 행렬을 생성하고, 상기 초기 행렬에 대해 상기 가중치 행렬을 s(s는 2이상의 자연수)회 곱하여 n x n 크기의 제2 연산 행렬을 생성하는 연산부;

상기 제2 연산 행렬을 구성하는 n개의 행벡터들 각각을 상기 n개의 유전자들 각각에 대응되는 제2 벡터로 지정한 후 상기 n개의 유전자들 각각에 대응되는 제2 벡터 중 상기 k개의 상위 유전자들에 대응되는 제2 벡터와 상기 k개의 하위 유전자들에 대응되는 제2 벡터를 선별하는 벡터 선별부;

상기 k개의 상위 유전자들에 대응되는 제2 벡터에 포함된 각 성분을 상기 n개의 유전자들 각각에 대응되는 상위 영향력 값으로 하나씩 할당함으로써, 상기 n개의 유전자들 각각에 대한 k개의 상위 영향력 값들을 생성하고, 상기 k개의 하위 유전자들에 대응되는 제2 벡터에 포함된 각 성분을 상기 n개의 유전자들 각각에 대응되는 하위 영향력 값으로 하나씩 할당함으로써, 상기 n개의 유전자들 각각에 대한 k개의 하위 영향력 값들을 생성하는 할당부;

상기 n개의 유전자들 각각에 대해, 상기 k개의 상위 영향력 값들과 상기 k개의 하위 영향력 값들 간의 독립 표본 t-검정(independent two sample t-test)을 수행함으로써, 상기 n개의 유전자들 각각에 대응되는 제2 t-검정 값을 생성하는 제2 검정부; 및

상기 n개의 유전자들 중 상기 제2 t-검정 값이 높은 순으로 p(p는 2 이상의 자연수)개의 유전자들을 선택한 후 상기 선택된 p개의 유전자들을 상기 제1 암환자에 대한 암 유발 유전자로 식별하는 식별부

를 더 포함하는 암 유발 유전자의 식별을 위한 전자 장치.

청구항 4

복수의 암환자들로부터 사전 수집된 n(n은 6이상의 자연수)개의 유전자들 각각의 발현량 데이터인 제1 발현량 데이터가 저장되어 있는 제1 데이터 저장부를 유지하는 단계;

복수의 정상 환자들로부터 사전 수집된 상기 n개의 유전자들 각각의 발현량 데이터인 제2 발현량 데이터가 저장되어 있는 제2 데이터 저장부를 유지하는 단계;

상기 복수의 암환자들 중 어느 한 명인 제1 암환자로부터 사전 측정된 상기 n개의 유전자들 상호 간의 영향력 정도를 표상하는 가중치 - 상기 가중치는 그 값이 클수록 유전자 상호 간의 영향력이 큼을 의미함 - 가 저장되어 있는 가중치 저장부를 유지하는 단계;

상기 n개의 유전자들 중 상기 복수의 암환자들과 상기 복수의 정상 환자들에서 돌연변이가 적어도 1회 이상 발생한 것으로 사전 확인된 적어도 하나의 돌연변이 유전자에 대한 정보가 저장되어 있는 돌연변이 정보 저장부를 유지하는 단계;

상기 n개의 유전자들 각각에 대응되는 행과 상기 n개의 유전자들 각각에 대응되는 열을 갖는 n x n 크기의 행렬에서 상기 적어도 하나의 돌연변이 유전자에 대응되는 행과 열이 만나는 지점의 성분을 1로 할당하고, 나머지 지점의 성분을 0으로 할당함으로써, n x n 크기의 유전자 행렬을 생성하는 단계;

상기 가중치 저장부를 참조하여, 상기 n x n 크기의 행렬에서 상기 n개의 유전자들 상호 간의 행과 열이 만나는 지점에 상기 n개의 유전자들 상호 간의 가중치를 성분으로 할당함으로써, n x n 크기의 가중치 행렬을 생성하는 단계;

상기 유전자 행렬에 대해 상기 가중치 행렬을 t(t는 2이상의 자연수)회 곱하여 n x n 크기의 제1 연산 행렬을 생성하고, 상기 제1 연산 행렬을 구성하는 n개의 행벡터들 각각을 상기 n개의 유전자들 각각에 대응되는 제1 벡터로 지정한 후 상기 n개의 유전자들 각각에 대해, 각 유전자에 대응되는 제1 벡터를 구성하는 n개의 성분들의 분산을 연산함으로써, 상기 n개의 유전자들 각각에 대응되는 분산을 산출하는 단계;

상기 n개의 유전자들 각각에 대해, 상기 제1 암환자로부터 수집된 제1 발현량 데이터와 상기 복수의 정상 환자들로부터 수집된 제2 발현량 데이터 간의 단일 표본 t-검정(one sample t-test)을 수행함으로써, 상기 n개의 유전자들 각각에 대응되는 제1 t-검정 값을 생성하는 단계; 및

상기 n개의 유전자들 각각에 대응되는 분산과 제1 t-검정 값을 기초로 상기 n개의 유전자들 중 암에 대한 영향과 돌연변이 유전자에 대한 영향을 크게 받는 것으로 확인되는 k(k는 2이상의 자연수)개의 상위 유전자들과 암

에 대한 영향과 돌연변이 유전자에 대한 영향을 작게 받는 것으로 확인되는 k개의 하위 유전자들을 선택하는 단계

를 포함하는 암 유발 유전자의 식별을 위한 전자 장치의 동작 방법.

청구항 5

제4항에 있어서,

상기 선택하는 단계는

상기 n개의 유전자들 각각에 대응되는 분산과 제1 t-검정 값을 서로 곱하여 상기 n개의 유전자들 각각에 대한 스코어(score)를 연산하는 단계; 및

상기 n개의 유전자들 중 상기 스코어가 높은 순으로 상기 k개의 상위 유전자들을 선택하고, 상기 n개의 유전자들 중 상기 스코어가 낮은 순으로 상기 k개의 하위 유전자들을 선택하는 단계

를 포함하는 암 유발 유전자의 식별을 위한 전자 장치의 동작 방법.

청구항 6

제5항에 있어서,

상기 k개의 상위 유전자들과 상기 k개의 하위 유전자들이 선택되면, 상기 n x n 크기의 행렬에서 대각 성분을 모두 1로 할당하여 n x n 크기의 초기 행렬을 생성하고, 상기 초기 행렬에 대해 상기 가중치 행렬을 s(s는 2이상의 자연수)회 곱하여 n x n 크기의 제2 연산 행렬을 생성하는 단계;

상기 제2 연산 행렬을 구성하는 n개의 행벡터들 각각을 상기 n개의 유전자들 각각에 대응되는 제2 벡터로 지정한 후 상기 n개의 유전자들 각각에 대응되는 제2 벡터 중 상기 k개의 상위 유전자들에 대응되는 제2 벡터와 상기 k개의 하위 유전자들에 대응되는 제2 벡터를 선별하는 단계;

상기 k개의 상위 유전자들에 대응되는 제2 벡터에 포함된 각 성분을 상기 n개의 유전자들 각각에 대응되는 상위 영향력 값으로 하나씩 할당함으로써, 상기 n개의 유전자들 각각에 대한 k개의 상위 영향력 값들을 생성하고, 상기 k개의 하위 유전자들에 대응되는 제2 벡터에 포함된 각 성분을 상기 n개의 유전자들 각각에 대응되는 하위 영향력 값으로 하나씩 할당함으로써, 상기 n개의 유전자들 각각에 대한 k개의 하위 영향력 값들을 생성하는 단계;

상기 n개의 유전자들 각각에 대해, 상기 k개의 상위 영향력 값들과 상기 k개의 하위 영향력 값들 간의 독립 표본 t-검정(independent two sample t-test)을 수행함으로써, 상기 n개의 유전자들 각각에 대응되는 제2 t-검정 값을 생성하는 단계; 및

상기 n개의 유전자들 중 상기 제2 t-검정 값이 높은 순으로 p(p는 2 이상의 자연수)개의 유전자들을 선택한 후 상기 선택된 p개의 유전자들을 상기 제1 암환자에 대한 암 유발 유전자로 식별하는 단계

를 더 포함하는 암 유발 유전자의 식별을 위한 전자 장치의 동작 방법.

청구항 7

제4항 내지 제6항 중 어느 한 항의 방법을 컴퓨터와의 결합을 통해 실행시키기 위한 컴퓨터 프로그램을 기록한 컴퓨터 판독 가능 기록 매체.

청구항 8

제4항 내지 제6항 중 어느 한 항의 방법을 컴퓨터와의 결합을 통해 실행시키기 위한 저장매체에 저장된 컴퓨터 프로그램.

발명의 설명

기 술 분 야

[0001] 본 발명은 암 유발 유전자의 식별을 위한 전자 장치 및 그 동작 방법에 대한 것이다.

배 경 기 술

[0002] 최근, 식생활의 서구화 등으로 인해 암환자가 증가함에 따라 암의 치료를 위한 다양한 방법이 강구되고 있다.

[0003] 암을 치료하는데 있어서, 암을 유발하는 유전자만을 식별할 수 있다면, 해당 유전자의 특성을 기반으로 암환자에게 적절한 진단 및 치료 방법을 적용함으로써, 암 치료의 효과를 극대화할 수 있을 것이다.

[0004] 하지만, 사람의 유전자는 그 종류가 너무 많기 때문에 암을 유발하는 유전자를 식별하는 것이 쉽지 않은 실정이다.

[0005] 이와 관련해서, 암환자와 정상 환자로부터 유전자에 대한 데이터를 수집한 후 수집된 데이터를 기초로 유전자와 암 발생 간의 연관도를 분석함으로써, 사람의 유전자들 중 암을 유발하는데 영향을 미치는 유전자를 식별하는 기술의 도입을 고려할 수 있다.

[0006] 따라서, 암환자와 정상 환자로부터 수집된 유전자 데이터를 기초로 사람의 유전자들 중에서 암의 발생에 영향을 미치는 특정 유전자들만을 식별함으로써, 암의 진단과 치료에 도움을 줄 수 있는 데이터 분석 기술에 대한 연구가 필요하다.

발명의 내용

해결하려는 과제

[0007] 본 발명은 암환자와 정상 환자로부터 수집된 유전자 데이터를 기초로 사람의 유전자들 중에서 암의 발생에 영향을 미치는 특정 유전자들만을 식별할 수 있는 전자 장치 및 그 동작 방법을 제시함으로써, 암의 진단과 치료에 도움이 될 수 있도록 지원하고자 한다.

과제의 해결 수단

[0008] 본 발명의 일실시예에 따른 암 유발 유전자의 식별을 위한 전자 장치는 복수의 암환자들로부터 사전 수집된 n(n은 6이상의 자연수)개의 유전자들 각각의 발현량 데이터인 제1 발현량 데이터가 저장되어 있는 제1 데이터 저장부, 복수의 정상 환자들로부터 사전 수집된 상기 n개의 유전자들 각각의 발현량 데이터인 제2 발현량 데이터가 저장되어 있는 제2 데이터 저장부, 상기 복수의 암환자들 중 어느 한 명인 제1 암환자로부터 사전 측정된 상기 n개의 유전자들 상호 간의 영향력 정도를 표상하는 가중치 - 상기 가중치는 그 값이 클수록 유전자 상호 간의 영향력이 큼을 의미함 - 가 저장되어 있는 가중치 저장부, 상기 n개의 유전자들 중 상기 복수의 암환자들과 상기 복수의 정상 환자들에서 돌연변이가 적어도 1회 이상 발생한 것으로 사전 확인된 적어도 하나의 돌연변이 유전자에 대한 정보가 저장되어 있는 돌연변이 정보 저장부, 상기 n개의 유전자들 각각에 대응되는 행과 상기 n개의 유전자들 각각에 대응되는 열을 갖는 n x n 크기의 행렬에서 상기 적어도 하나의 돌연변이 유전자에 대응되는 행과 열이 만나는 지점의 성분을 1로 할당하고, 나머지 지점의 성분을 0으로 할당함으로써, n x n 크기의 유전자 행렬을 생성하는 유전자 행렬 생성부, 상기 가중치 저장부를 참조하여, 상기 n x n 크기의 행렬에서 상기 n개의 유전자들 상호 간의 행과 열이 만나는 지점에 상기 n개의 유전자들 상호 간의 가중치를 성분으로 할당함으로써, n x n 크기의 가중치 행렬을 생성하는 가중치 행렬 생성부, 상기 유전자 행렬에 대해 상기 가중치 행렬을 t(t는 2이상의 자연수)회 곱하여 n x n 크기의 제1 연산 행렬을 생성하고, 상기 제1 연산 행렬을 구성하는 n개의 행벡터들 각각을 상기 n개의 유전자들 각각에 대응되는 제1 벡터로 지정한 후 상기 n개의 유전자들 각각에 대해, 각 유전자에 대응되는 제1 벡터를 구성하는 n개의 성분들의 분산을 연산함으로써, 상기 n개의 유전자들 각각에 대응되는 분산을 산출하는 분산 산출부, 상기 n개의 유전자들 각각에 대해, 상기 제1 암환자로부터 수집된 제1 발현량 데이터와 상기 복수의 정상 환자들로부터 수집된 제2 발현량 데이터 간의 단일 표본 t-검정(one sample t-test)을 수행함으로써, 상기 n개의 유전자들 각각에 대응되는 제1 t-검정 값을 생성하는 제1 검정부 및 상기 n개의 유전자들 각각에 대응되는 분산과 제1 t-검정 값을 기초로 상기 n개의 유전자들 중 암에 대한 영향과 돌연변이 유전자에 대한 영향을 크게 받는 것으로 확인되는 k(k는 2이상의 자연수)개의 상위 유전자들과 암에 대한 영향과 돌연변이 유전자에 대한 영향을 작게 받는 것으로 확인되는 k개의 하위 유전자들을 선택하는 유전자 선택부를 포함한다.

[0009] 또한, 본 발명의 일실시예에 따른 암 유발 유전자의 식별을 위한 전자 장치의 동작 방법은 복수의 암환자들로부터 사전 수집된 n(n은 6이상의 자연수)개의 유전자들 각각의 발현량 데이터인 제1 발현량 데이터가 저장되어 있는 제1 데이터 저장부를 유지하는 단계, 복수의 정상 환자들로부터 사전 수집된 상기 n개의 유전자들 각각의 발현량 데이터인 제2 발현량 데이터가 저장되어 있는 제2 데이터 저장부를 유지하는 단계, 상기 복수의 암환자들

중 어느 한 명인 제1 암환자로부터 사전 측정된 상기 n개의 유전자들 상호 간의 영향력 정도를 표상하는 가중치 - 상기 가중치는 그 값이 클수록 유전자 상호 간의 영향력이 큼을 의미함 - 가 저장되어 있는 가중치 저장부를 유지하는 단계, 상기 n개의 유전자들 중 상기 복수의 암환자들과 상기 복수의 정상 환자들에서 돌연변이가 적어도 1회 이상 발생한 것으로 사전 확인된 적어도 하나의 돌연변이 유전자에 대한 정보가 저장되어 있는 돌연변이 정보 저장부를 유지하는 단계, 상기 n개의 유전자들 각각에 대응되는 행과 상기 n개의 유전자들 각각에 대응되는 열을 갖는 n x n 크기의 행렬에서 상기 적어도 하나의 돌연변이 유전자에 대응되는 행과 열이 만나는 지점의 성분을 1로 할당하고, 나머지 지점의 성분을 0으로 할당함으로써, n x n 크기의 유전자 행렬을 생성하는 단계, 상기 가중치 저장부를 참조하여, 상기 n x n 크기의 행렬에서 상기 n개의 유전자들 상호 간의 행과 열이 만나는 지점에 상기 n개의 유전자들 상호 간의 가중치를 성분으로 할당함으로써, n x n 크기의 가중치 행렬을 생성하는 단계, 상기 유전자 행렬에 대해 상기 가중치 행렬을 t(t는 2이상의 자연수)회 곱하여 n x n 크기의 제1 연산 행렬을 생성하고, 상기 제1 연산 행렬을 구성하는 n개의 행벡터들 각각을 상기 n개의 유전자들 각각에 대응되는 제1 벡터로 지정한 후 상기 n개의 유전자들 각각에 대해, 각 유전자에 대응되는 제1 벡터를 구성하는 n개의 성분들의 분산을 연산함으로써, 상기 n개의 유전자들 각각에 대응되는 분산을 산출하는 단계, 상기 n개의 유전자들 각각에 대해, 상기 제1 암환자로부터 수집된 제1 발현량 데이터와 상기 복수의 정상 환자들로부터 수집된 제2 발현량 데이터 간의 단일 표본 t-검정을 수행함으로써, 상기 n개의 유전자들 각각에 대응되는 제1 t-검정 값을 생성하는 단계 및 상기 n개의 유전자들 각각에 대응되는 분산과 제1 t-검정 값을 기초로 상기 n개의 유전자들 중 암에 대한 영향과 돌연변이 유전자에 대한 영향을 크게 받는 것으로 확인되는 k(k는 2이상의 자연수)개의 상위 유전자들과 암에 대한 영향과 돌연변이 유전자에 대한 영향을 작게 받는 것으로 확인되는 k개의 하위 유전자들을 선택하는 단계를 포함한다.

발명의 효과

[0010] 본 발명은 암환자와 정상 환자로부터 수집된 유전자 데이터를 기초로 사람의 유전자들 중에서 암의 발생에 영향을 미치는 특정 유전자들만을 식별할 수 있는 전자 장치 및 그 동작 방법을 제시함으로써, 암의 진단과 치료에 도움이 될 수 있다.

도면의 간단한 설명

[0011] 도 1은 본 발명의 일실시예에 따른 암 유발 유전자의 식별을 위한 전자 장치의 구조를 도시한 도면이다.

도 2는 본 발명의 일실시예에 따른 암 유발 유전자의 식별을 위한 전자 장치의 동작을 설명하기 위한 도면이다.

도 3은 본 발명의 일실시예에 따른 암 유발 유전자의 식별을 위한 전자 장치의 동작 방법을 도시한 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0012] 이하에서는 본 발명에 따른 실시예들을 첨부된 도면을 참조하여 상세하게 설명하기로 한다. 이러한 설명은 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다. 각 도면을 설명하면서 유사한 참조부호를 유사한 구성요소에 대해 사용하였으며, 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 본 명세서 상에서 사용되는 모든 용어들은 본 발명이 속하는 기술분야에서 통상의 지식을 가진 사람에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다.

[0013] 본 문서에서, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있다는 것을 의미한다. 또한, 본 발명의 다양한 실시예들에 있어서, 각 구성요소들, 기능 블록들 또는 수단들은 하나 또는 그 이상의 하부 구성요소로 구성될 수 있고, 각 구성요소들이 수행하는 전기, 전자, 기계적 기능들은 전자회로, 집적회로, ASIC(Application Specific Integrated Circuit) 등 공지된 다양한 소자들 또는 기계적 요소들로 구현될 수 있으며, 각각 별개로 구현되거나 2 이상이 하나로 통합되어 구현될 수도 있다.

[0014] 한편, 첨부된 블록도의 블록들이나 흐름도의 단계들은 범용 컴퓨터, 특수용 컴퓨터, 휴대용 노트북 컴퓨터, 네트워크 컴퓨터 등 데이터 프로세싱이 가능한 장비의 프로세서나 메모리에 탑재되어 지정된 기능들을 수행하는 컴퓨터 프로그램 명령들(instructions)을 의미하는 것으로 해석될 수 있다. 이들 컴퓨터 프로그램 명령들은 컴퓨터 장치에 구비된 메모리 또는 컴퓨터에서 판독 가능한 메모리에 저장될 수 있기 때문에, 블록도의 블록들 또는 흐름도의 단계들에서 설명된 기능들은 이를 수행하는 명령 수단을 내포하는 제조물로 생산될 수도 있다. 아울러, 각 블록 또는 각 단계는 특정된 논리적 기능(들)을 실행하기 위한 하나 이상의 실행 가능한 명령들을 포

함하는 모듈, 세그먼트 또는 코드의 일부를 나타낼 수 있다. 또, 몇 가지 대체 가능한 실시예들에서는 블록들 또는 단계들에서 언급된 기능들이 정해진 순서와 달리 실행되는 것도 가능함을 주목해야 한다. 예컨대, 잇달아 도시되어 있는 두 개의 블록들 또는 단계들은 실질적으로 동시에 수행되거나, 역순으로 수행될 수 있으며, 경우에 따라 일부 블록들 또는 단계들이 생략된 채로 수행될 수도 있다.

[0015] 도 1은 본 발명의 일실시예에 따른 암 유발 유전자의 식별을 위한 전자 장치의 구조를 도시한 도면이다.

[0016] 도 1을 참조하면, 본 발명에 따른 전자 장치(110)는 제1 데이터 저장부(111), 제2 데이터 저장부(112), 가중치 저장부(113), 돌연변이 정보 저장부(114), 유전자 행렬 생성부(115), 가중치 행렬 생성부(116), 분산 산출부(117), 제1 검정부(118) 및 유전자 선택부(119)를 포함한다.

[0017] 제1 데이터 저장부(111)에는 복수의 암환자들로부터 사전 수집된 n(n은 6이상의 자연수)개의 유전자들 각각의 발현량 데이터인 제1 발현량 데이터가 저장되어 있다.

[0018] 예컨대, 상기 n개의 유전자들이 'g1, g2, g3, ..., gn'이라고 하는 경우, 제1 데이터 저장부(111)에는 하기의 표 1과 같이, 복수의 암환자들로부터 사전 수집된 상기 n개의 유전자들 각각의 발현량 데이터인 상기 제1 발현량 데이터가 저장되어 있을 수 있다.

표 1

[0020]

	암환자 1	암환자 2	암환자 3	암환자 4	...	암환자 N
g1	데이터A1	데이터A2	데이터A3	데이터A4	...	데이터A5
g2	데이터A6	데이터A7	데이터A8	데이터A9	...	데이터A10
g3	데이터A11	데이터A12	데이터A13	데이터A14	...	데이터A15
g4	데이터A16	데이터A17	데이터A18	데이터A19	...	데이터A20
...	...	...	...	...	...	...
gn	데이터A21	데이터A22	데이터A23	데이터A24	..	데이터A25

[0022] 제2 데이터 저장부(112)에는 복수의 정상 환자들로부터 사전 수집된 상기 n개의 유전자들 각각의 발현량 데이터인 제2 발현량 데이터가 저장되어 있다.

[0023] 예컨대, 제2 데이터 저장부(112)에는 하기의 표 2와 같이, 복수의 정상 환자들로부터 사전 수집된 상기 n개의 유전자들 각각의 발현량 데이터인 상기 제2 발현량 데이터가 저장되어 있을 수 있다.

표 2

[0025]

	정상 환자 1	정상 환자 2	정상 환자 3	정상 환자 4	...	정상 환자 N
g1	데이터B1	데이터B2	데이터B3	데이터B4	...	데이터B5
g2	데이터B6	데이터B7	데이터B8	데이터B9	...	데이터B10
g3	데이터B11	데이터B12	데이터B13	데이터B14	...	데이터B15
g4	데이터B16	데이터B17	데이터B18	데이터B19	...	데이터B20
...	...	...	...	...	...	...
gn	데이터B21	데이터B22	데이터B23	데이터B24	..	데이터B25

[0027] 가중치 저장부(113)에는 상기 복수의 암환자들 중 어느 한 명인 제1 암환자로부터 사전 측정된 상기 n개의 유전자들 상호 간의 영향력 정도를 표상하는 가중치가 저장되어 있다.

[0028] 여기서, 상기 가중치는 그 값이 클수록 유전자 상호 간의 영향력이 큰 것을 의미하는 척도로서, 개발자에 의해 실험 등을 통해 사전 지정되는 값을 의미한다.

[0029] 예컨대, 상기 n개의 유전자들이 'g1, g2, g3, ..., gn'이라고 하는 경우, 가중치 저장부(113)에는 하기의 표 3과 같이, 상기 n개의 유전자들 상호 간의 영향력 정도를 표상하는 가중치가 저장되어 있을 수 있다.

표 3

[0031]

유전자	g1	g2	g3	g4	...	gn
g1	0	0.45	0	0.73	...	0
g2	0.28	0	0.96	0	...	0.56
g3	0.21	0	0	0.009	...	0.71
g4	0	0	0	0	...	0
...	...	...	...	...	...	...
gn	0	0.06	0.31	0	...	0

[0033] 이때, 본 발명의 일실시예에 따르면, 가중치 저장부(113)에 저장되어 있는 상기 n개의 유전자들 상호 간의 가중치는 하기의 수학식 1의 연산에 기초하여 개발자에 의해 사전 설정될 수 있다.

수학식 1

[0035]
$$W_{g_i,g_j}^{k} = \left|PCC\left(T_{g_i},T_{g_j}\right)\right| \circ \left|PCC\left(T_{g_i},T_{g_j}\right) - PCC\left(N_{g_i},N_{g_j}\right)\right| \circ \frac{err_{g_i,g_j}^{max} - err_{g_i,g_j}^{k}}{err_{g_i,g_j}^{max}} \circ A$$

[0037] 상기 수학식 1에서, gi, gj는 상기 n개의 유전자들 중 각각 i번째, j번째 유전자를 의미하고, W_{g_i,g_j}^{k}는 상기 제1 암환자를 상기 복수의 암환자들 중 k번째 암환자라고 하는 경우, 상기 제1 암환자에 대한 gi 유전자와 gj 유전자 간의 가중치를 의미한다.

[0038] 그리고, $\circ$ 은 아다마르곱을 의미하고, PCC는 피어슨 상관계수를 의미하며, T_{gi}는 상기 복수의 암환자들로부터 수집된 gi 유전자에 대한 제1 발현량 데이터, N_{gi}는 상기 복수의 정상 환자들로부터 수집된 gi 유전자에 대한 제2 발현량 데이터를 의미한다.

[0039] err은 상기 복수의 암환자들과 상기 복수의 정상 환자들로부터 수집된 gi와 gj의 발현량 데이터를 각각 x, y좌표 값으로 구성하여 좌표 평면에 점으로 나타내고, 이로부터 나는 점들을 통해 그 특성을 나타낼 수 있는 직선을 찾았을 때, 각 점들과 직선 사이의 거리를 의미한다. 이때, err^{max}는 이러한 err 중 이상치(outlier)를 제거한 후의 err 최대 값을 의미하고, err_{g_i,g_j}^{k}는 이러한 err 중 k번째 암환자와 k번째 정상 환자로부터 수집된 gi 유전자와 gj 유전자의 발현량 데이터로 표현한 점에 대한 err을 의미한다.

[0040] 그리고, A는 사전 설정된 유전자 네트워크에서 상기 n개의 유전자들 상호 간에 링크가 설정되어 있는지 여부를 나타내는 행렬로서, 상기 유전자 네트워크에서 서로 링크가 설정되어 있는 유전자 간의 성분은 1로, 그렇지 않은 성분은 0으로 할당된 행렬을 의미한다.

[0041] 여기서, 유전자 네트워크란 서로 다른 종류의 복수의 유전자들 중 서로 영향을 미치는 유전자 간에 링크가 설정되어 있는 네트워크를 의미하는 것으로, 이러한 유전자 네트워크는 생물학적 경로, protein-protein interaction(PPI), Gene Ontology(GO) 데이터 등에 의해서 구축될 수 있다.

[0042] 이때, 개발자는 상기 수학식 1의 연산에 기초하여, 상기 복수의 암환자들과 상기 복수의 정상 환자들로부터 수

집된 상기 n개의 유전자들에 대한 발현량 데이터를 이용하여 상기 제1 암환자인 k번째 암환자에 대한 gi 유전자와 gj 유전자 간의 가중치인 $W_{gi,gj}^{k}$ 를 연산하는 방식으로, 상기 제1 암환자에 대한 상기 n개의 유전자들 상호 간의 가중치를 연산한 후 가중치 저장부(113) 상에 저장해 둘 수 있다.

[0043] 돌연변이 정보 저장부(114)에는 상기 n개의 유전자들 중 상기 복수의 암환자들과 상기 복수의 정상 환자들에서 돌연변이가 적어도 1회 이상 발생한 것으로 사전 확인된 적어도 하나의 돌연변이 유전자에 대한 정보가 저장되어 있다.

[0044] 예컨대, 상기 n개의 유전자들 중 'g1, g3,'이라는 유전자에 돌연변이가 적어도 1회 이상 발생한 것으로 확인되었다면, 돌연변이 정보 저장부(114)에는 'g1, g3'이라는 유전자에 대한 정보가 저장되어 있을 수 있다.

[0045] 유전자 행렬 생성부(115)는 상기 n개의 유전자들 각각에 대응되는 행과 상기 n개의 유전자들 각각에 대응되는 열을 갖는 n x n 크기의 행렬에서 상기 적어도 하나의 돌연변이 유전자에 대응되는 행과 열이 만나는 지점의 성분을 1로 할당하고, 나머지 지점의 성분을 0으로 할당함으로써, n x n 크기의 유전자 행렬을 생성한다.

[0046] 관련해서, 상기 n x n 크기의 행렬(210)은 도 2에 도시된 그림과 같이, 상기 n개의 유전자들 각각에 대응되는 행과 상기 n개의 유전자들 각각에 대응되는 열로 구성된 행렬을 의미한다. 이때, 유전자 행렬 생성부(115)는 상기 n x n 크기의 행렬(210)에서 상기 적어도 하나의 돌연변이 유전자에 대응되는 행과 열이 만나는 지점의 성분을 1로 할당하고, 나머지 지점의 성분을 0으로 할당함으로써, n x n 크기의 유전자 행렬을 생성할 수 있다.

[0047] 예컨대, 상기 적어도 하나의 돌연변이 유전자가 'g1, g3'이라고 하는 경우, 유전자 행렬 생성부(115)는 상기 n x n 크기의 행렬(210)에서 'g1, g3' 각각에 대응되는 행과 열이 만나는 지점의 성분을 1로 할당하고, 나머지 지점의 성분을 0으로 할당함으로써, '$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}$'와 같은 n x n 크기의 유전자 행렬을 생성할 수 있다.

[0048] 가중치 행렬 생성부(116)는 가중치 저장부(113)를 참조하여, 상기 n x n 크기의 행렬에서 상기 n개의 유전자들 상호 간의 행과 열이 만나는 지점에 상기 n개의 유전자들 상호 간의 가중치를 성분으로 할당함으로써, n x n 크기의 가중치 행렬을 생성한다.

[0049] 예컨대, 가중치 저장부(113)에 저장되어 있는 상기 n개의 유전자들 상호 간의 가중치가 상기 표 3과 같다고 하는 경우, 가중치 행렬 생성부(116)는 '$\begin{bmatrix} 0 & 0.45 & 0 & 0.73 & \dots & 0 \\ 0.28 & 0 & 0.96 & 0 & \dots & 0.56 \\ 0.21 & 0 & 0 & 0.009 & \dots & 0.71 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0.06 & 0.31 & 0 & \dots & 0 \end{bmatrix}$'와 같은 가중치 행렬을 생성할 수 있다.

[0050] 분산 산출부(117)는 상기 유전자 행렬에 대해 상기 가중치 행렬을 t(t는 2이상의 자연수)회 곱하여 n x n 크기의 제1 연산 행렬을 생성하고, 상기 제1 연산 행렬을 구성하는 n개의 행벡터들 각각을 상기 n개의 유전자들 각각에 대응되는 제1 벡터로 지정한 후 상기 n개의 유전자들 각각에 대해, 각 유전자에 대응되는 제1 벡터를 구성하는 n개의 성분들의 분산을 연산함으로써, 상기 n개의 유전자들 각각에 대응되는 분산을 산출한다.

[0051] 관련해서, t를 '7'이라고 하는 경우, 분산 산출부(117)는 상기 유전자 행렬에 상기 가중치 행렬을 '7회' 연속해서 곱하여 상기 제1 연산 행렬을 생성할 수 있다. 즉, 분산 산출부(117)는 상기 유전자 행렬에 상기 가중치 행렬을 곱한 후 산출되는 행렬에 상기 가중치 행렬을 다시 곱하는 행위를 총 '7회' 수행함으로써, 상기 제1 연산 행렬을 생성할 수 있다.

[0052] 그러고 나서, 분산 산출부(117)는 상기 제1 연산 행렬을 구성하는 n개의 행벡터들 각각을 상기 n개의 유전자들 각각에 대응되는 제1 벡터로 지정할 수 있다.

[0053] 그 이후, 분산 산출부(117)는 상기 n개의 유전자들 각각에 대해, 각 유전자에 대응되는 제1 벡터를 구성하는 n

개의 성분들의 분산을 연산함으로써, 상기 n개의 유전자들 각각에 대응되는 분산을 산출할 수 있다.

[0054] 관련해서, n이 '100'이라고 하는 경우, 분산 산출부(117)는 100개의 유전자들 각각에 대응되는 제1 벡터를 구성하는 100개의 성분들에 대한 분산(variance)을 연산할 수 있다.

[0055] 제1 검정부(118)는 상기 n개의 유전자들 각각에 대해, 상기 제1 암환자로부터 수집된 제1 발현량 데이터와 상기 복수의 정상 환자들로부터 수집된 제2 발현량 데이터 간의 단일 표본 t-검정(one sample t-test)을 수행함으로써, 상기 n개의 유전자들 각각에 대응되는 제1 t-검정 값을 생성한다.

[0056] 관련해서, 제1 데이터 저장부(111)에 저장되어 있는 데이터가 상기 표 1과 같다고 하고, 제2 데이터 저장부(112)에 저장되어 있는 데이터가 상기 표 2와 같다고 하며, 상기 제1 암환자를 '암환자 1'이라고 하는 경우, 제1 검정부(118)는 상기 n개의 유전자들 각각에 대해, 상기 제1 암환자인 '암환자 1'로부터 수집된 제1 발현량 데이터와 상기 복수의 정상 환자들로부터 수집된 제2 발현량 데이터 간의 단일 표본 t-검정을 수행함으로써, 상기 n개의 유전자들 각각에 대응되는 제1 t-검정 값을 생성할 수 있다.

[0057] 관련해서, 제1 검정부(118)는 'g1' 유전자에 대해서, '암환자 1'의 제1 발현량 데이터인 '데이터A1'과 상기 복수의 정상 환자들의 제1 발현량 데이터인 '데이터B1~데이터B5' 간의 단일 표본 t-검정을 수행함으로써, 'g1' 유전자에 대한 제1 t-검정 값을 생성할 수 있다. 이러한 방식으로, 제1 검정부(118)는 나머지 유전자들 각각에 대해서도 상기 제1 t-검정 값을 생성할 수 있다.

[0058] 유전자 선택부(119)는 상기 n개의 유전자들 각각에 대응되는 분산과 제1 t-검정 값을 기초로 상기 n개의 유전자들 중 암에 대한 영향과 돌연변이 유전자에 대한 영향을 크게 받는 것으로 확인되는 k(k는 2이상의 자연수)개의 상위 유전자들과 암에 대한 영향과 돌연변이 유전자에 대한 영향을 작게 받는 것으로 확인되는 k개의 하위 유전자들을 선택한다.

[0059] 이때, 본 발명의 일실시예에 따르면, 유전자 선택부(119)는 스코어 연산부(120) 및 선택부(121)를 포함할 수 있다.

[0060] 스코어 연산부(120)는 상기 n개의 유전자들 각각에 대응되는 분산과 제1 t-검정 값을 서로 곱하여 상기 n개의 유전자들 각각에 대한 스코어(score)를 연산한다.

[0061] 선택부(121)는 상기 n개의 유전자들 중 상기 스코어가 높은 순으로 상기 k개의 상위 유전자들을 선택하고, 상기 n개의 유전자들 중 상기 스코어가 낮은 순으로 상기 k개의 하위 유전자들을 선택한다.

[0062] 예컨대, n이 '100'이라고 하고, k가 '20'이라고 하는 경우, 스코어 연산부(120)는 100개의 유전자들 각각에 대해서 생성된 분산과 제1 t-검정 값을 각각 서로 곱함으로써, 그 결과 값을 100개의 유전자들 각각에 대한 스코어로 연산할 수 있다.

[0063] 그러고 나서, 선택부(121)는 100개의 유전자들 중 상기 스코어가 높은 순으로 20개의 상위 유전자들을 선택하고, 상기 스코어가 낮은 순으로 20개의 하위 유전자들을 선택할 수 있다.

[0064] 여기서, 상기 상위 유전자들은 상기 스코어가 높은 유전자들로서, 상기 제1 t-검정 값이 큰 값을 갖는 유전자이기 때문에, 암환자의 유전자와 정상 환자의 유전자 간의 발현량의 차이가 유의미한 차이를 갖는 유전자임과 동시에, 상기 유전자 행렬과 상기 가중치 행렬이 곱해진 연산 행렬 상에서 상기 분산이 큰 값을 갖는 유전자이기 때문에, 특정 돌연변이 유전자에 의해 영향을 많이 받는 유전자라고 볼 수 있다. 따라서, 상기 상위 유전자들은 암과의 상관도가 높은 유전자라고 볼 수 있다. 반면에, 상기 하위 유전자들은 상기 스코어가 낮은 유전자들로서, 암환자의 유전자와 정상 환자의 유전자와 간의 발현량의 차이가 작고, 특정 돌연변이 유전자에 의한 영향을 적게 받는 유전자라고 볼 수 있기 때문에 암과의 상관도가 낮은 유전자라고 볼 수 있다.

[0065] 이때, 본 발명의 일실시예에 따르면, 전자 장치(110)는 연산부(122), 벡터 선별부(123), 할당부(124), 제2 검정부(125) 및 식별부(126)를 더 포함할 수 있다.

[0066] 연산부(122)는 상기 k개의 상위 유전자들과 상기 k개의 하위 유전자들이 선택되면, 상기 n x n 크기의 행렬(210)에서 대각 성분을 모두 1로 할당하여 n x n 크기의 초기 행렬을 생성하고, 상기 초기 행렬에 대해 상기 가중치 행렬을 s(s는 2이상의 자연수)회 곱하여 n x n 크기의 제2 연산 행렬을 생성한다.

[0067] 관련해서, 연산부(122)는 상기 n x n 크기의 행렬(210)에서 대각 성분을 모두 1로 할당하여 '$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$', 이라는 초기 행렬을 생성한 후 상기 초기 행렬에 대해 '$\begin{bmatrix} 0 & 0.45 & 0 & 0.73 & \dots & 0 \\ 0.28 & 0 & 0.96 & 0 & \dots & 0.56 \\ 0.21 & 0 & 0 & 0.009 & \dots & 0.71 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0.06 & 0.31 & 0 & \dots & 0 \end{bmatrix}$'와 같은 가중치 행렬을 s회 연속해서 곱함으로써, n x n 크기의 제2 연산 행렬을 생성할 수 있다.

[0068] 벡터 선별부(123)는 상기 제2 연산 행렬을 구성하는 n개의 행벡터들 각각을 상기 n개의 유전자들 각각에 대응되는 제2 벡터로 지정한 후 상기 n개의 유전자들 각각에 대응되는 제2 벡터 중 상기 k개의 상위 유전자들에 대응되는 제2 벡터와 상기 k개의 하위 유전자들에 대응되는 제2 벡터를 선별한다.

[0069] 예컨대, 전술한 예와 같이, n을 '100'이라고 하고, k를 '20'이라고 하는 경우, 벡터 선별부(123)는 상기 제2 연산 행렬을 구성하는 100개의 행벡터들 각각을 100개의 유전자들 각각에 대응되는 제2 벡터로 지정한 후 상기 100개의 유전자들 각각에 대응되는 제2 벡터 중 20개의 상위 유전자들에 대응되는 제2 벡터와 20개의 하위 유전자들에 대응되는 제2 벡터를 선별할 수 있다.

[0070] 할당부(124)는 상기 k개의 상위 유전자들에 대응되는 제2 벡터에 포함된 각 성분을 상기 n개의 유전자들 각각에 대응되는 상위 영향력 값으로 하나씩 할당함으로써, 상기 n개의 유전자들 각각에 대한 k개의 상위 영향력 값들을 생성하고, 상기 k개의 하위 유전자들에 대응되는 제2 벡터에 포함된 각 성분을 상기 n개의 유전자들 각각에 대응되는 하위 영향력 값으로 하나씩 할당함으로써, 상기 n개의 유전자들 각각에 대한 k개의 하위 영향력 값들을 생성한다.

[0071] 관련해서, 앞서 설명한 예와 같이, 20개의 상위 유전자들과 20개의 하위 유전자들에 대응되는 제2 벡터가 선별되었다고 하는 경우, 할당부(124)는 상기 20개의 상위 유전자들에 대응되는 제2 벡터에 포함된 각 성분을 상기 100개의 유전자들 각각에 대응되는 상위 영향력 값으로 하나씩 할당할 수 있다.

[0072] 예컨대, 상기 20개의 상위 유전자들 중 하나가 'g50'이라고 하는 경우, 할당부(124)는 'g50'에 대응되는 제2 벡터에 포함된 100개의 성분들 각각을 상기 100개의 유전자들 각각에 대응되는 상위 영향력 값으로 할당할 수 있다. 즉, 할당부(124)는 'g50'에 대응되는 제2 벡터에 포함된 100개의 성분들 중 첫 번째 성분의 값을 'g1'이라는 유전자에 대한 상위 영향력 값으로 할당하고, 두 번째 성분의 값을 'g2'라는 유전자에 대한 상위 영향력 값으로 할당할 수 있으며, 세 번째 성분의 값을 'g3'이라는 유전자에 대한 상위 영향력 값으로 할당할 수 있다. 이렇게, 할당부(124)는 순차적으로 각 성분을 각 유전자에 대한 상위 영향력 값으로 할당함으로써, 최종적으로 백 번째 성분의 값을 'g100'이라는 유전자에 대한 상위 영향력 값으로 할당할 수 있다. 이러한 방식으로, 할당부(124)는 상기 20개의 상위 유전자들 중 'g50'을 제외한 나머지 19개의 상위 유전자들 각각에 대해서도 제2 벡터를 구성하는 100개의 성분들을 상기 100개의 유전자들 각각에 대한 상위 영향력 값으로 할당할 수 있다.

[0073] 이렇게, 상기 100개의 유전자들 각각에 상위 영향력 값의 할당이 완료되면, 상기 100개의 유전자들 각각에는 총 20개의 상위 유전자에 대한 제2 벡터를 기초로 상위 영향력 값의 할당이 진행되었기 때문에, 총 20개씩 상위 영향력 값이 할당될 수 있다.

[0074] 이렇게, 할당부(124)는 상기 100개의 유전자들 각각에 대해 20개의 상위 영향력 값을 할당하는 방식과 동일한 방식으로, 상기 100개의 유전자들 각각에 대해, 상기 20개의 하위 유전자들 각각에 대응되는 제2 벡터에 포함된 각 성분을 하위 영향력 값으로 하나씩 할당함으로써, 상기 100개의 유전자들 각각에 대한 20개의 하위 영향력 값들도 생성할 수 있다.

[0075] 제2 검정부(125)는 상기 n개의 유전자들 각각에 대해, 상기 k개의 상위 영향력 값들과 상기 k개의 하위 영향력 값들 간의 독립 표본 t-검정(independent two sample t-test)을 수행함으로써, 상기 n개의 유전자들 각각에 대응되는 제2 t-검정 값을 생성한다.

[0076] 관련해서, 앞서 설명한 예와 같이, 상기 100개의 유전자들 각각에 대한 20개의 상위 영향력 값들과 20개의 하위 영향력 값들이 생성되었다고 하는 경우, 제2 검정부(125)는 상기 100개의 유전자들 각각에 대해, 상기 20개의 상위 영향력 값들과 상기 20개의 하위 영향력 값들 간의 독립 표본 t-검정을 수행함으로써, 상기 100개의 유전자들 각각에 대응되는 제2 t-검정 값을 생성할 수 있다. 예컨대, 상기 100개의 유전자들 중 'g1'이라는 유전자가 있다고 하는 경우, 제2 검정부(125)는 'g1'이라는 유전자에 대한 20개의 상위 영향력 값들과 20개의 하위 영향력 값들 간의 독립 표본 t-검정을 수행함으로써, 'g1'이라는 유전자에 대한 제2 t-검정 값을 생성할 수 있다. 이러한 방식으로, 제2 검정부(125)는 상기 100개의 유전자들 각각에 대해서 20개의 상위 영향력 값들과 20개의 하위 영향력 값들 간의 독립 표본 t-검정을 수행함으로써, 제2 t-검정 값을 생성할 수 있다.

[0077] 식별부(126)는 상기 n개의 유전자들 중 상기 제2 t-검정 값이 높은 순으로 p(p는 2 이상의 자연수)개의 유전자들을 선택한 후 상기 선택된 p개의 유전자들을 상기 제1 암환자에 대한 암 유발 유전자로 식별한다.

[0078] 관련해서, 앞서 설명한 예와 같이, 상기 100개의 유전자들에 대한 제2 t-검정 값이 생성되었다고 하고, p를 '30'이라고 하는 경우, 식별부(126)는 상기 100개의 유전자들 중 상기 제2 t-검정 값이 높은 순으로 30개의 유전자들을 선택한 후 상기 선택된 30개의 유전자들을 상기 제1 암환자에 대한 암 유발 유전자로 식별할 수 있다.

[0079] 여기서, 상기 제2 t-검정 값이 높은 것으로 선택된 상기 30개의 유전자들은 20개의 상위 영향력 값들과 20개의 하위 영향력 값들 간에 유의미한 차이가 있는 유전자들로 볼 수 있기 때문에, 식별부(126)는 상기 30개의 유전자들을 상기 제1 암환자에 대한 암 유발 유전자로 식별할 수 있다.

[0080] 도 3은 본 발명의 일실시예에 따른 암 유발 유전자의 식별을 위한 전자 장치의 동작 방법을 도시한 순서도이다.

[0081] 단계(S310)에서는 복수의 암환자들로부터 사전 수집된 n(n은 6이상의 자연수)개의 유전자들 각각의 발현량 데이터인 제1 발현량 데이터가 저장되어 있는 제1 데이터 저장부를 유지한다.

[0082] 단계(S320)에서는 복수의 정상 환자들로부터 사전 수집된 상기 n개의 유전자들 각각의 발현량 데이터인 제2 발현량 데이터가 저장되어 있는 제2 데이터 저장부를 유지한다.

[0083] 단계(S330)에서는 상기 복수의 암환자들 중 어느 한 명인 제1 암환자로부터 사전 측정된 상기 n개의 유전자들 상호 간의 영향력 정도를 표상하는 가중치(상기 가중치는 그 값이 클수록 유전자 상호 간의 영향력이 큼을 의미함)가 저장되어 있는 가중치 저장부를 유지한다.

[0084] 단계(S340)에서는 상기 n개의 유전자들 중 상기 복수의 암환자들과 상기 복수의 정상 환자들에서 돌연변이가 적어도 1회 이상 발생한 것으로 사전 확인된 적어도 하나의 돌연변이 유전자에 대한 정보가 저장되어 있는 돌연변이 정보 저장부를 유지한다.

[0085] 단계(S350)에서는 상기 n개의 유전자들 각각에 대응되는 행과 상기 n개의 유전자들 각각에 대응되는 열을 갖는 n x n 크기의 행렬에서 상기 적어도 하나의 돌연변이 유전자에 대응되는 행과 열이 만나는 지점의 성분을 1로 할당하고, 나머지 지점의 성분을 0으로 할당함으로써, n x n 크기의 유전자 행렬을 생성한다.

[0086] 단계(S360)에서는 상기 가중치 저장부를 참조하여, 상기 n x n 크기의 행렬에서 상기 n개의 유전자들 상호 간의 행과 열이 만나는 지점에 상기 n개의 유전자들 상호 간의 가중치를 성분으로 할당함으로써, n x n 크기의 가중치 행렬을 생성한다.

[0087] 단계(S370)에서는 상기 유전자 행렬에 대해 상기 가중치 행렬을 t(t는 2이상의 자연수)회 곱하여 n x n 크기의 제1 연산 행렬을 생성하고, 상기 제1 연산 행렬을 구성하는 n개의 행벡터들 각각을 상기 n개의 유전자들 각각에 대응되는 제1 벡터로 지정한 후 상기 n개의 유전자들 각각에 대해, 각 유전자에 대응되는 제1 벡터를 구성하는 n개의 성분들의 분산을 연산함으로써, 상기 n개의 유전자들 각각에 대응되는 분산을 산출한다.

[0088] 단계(S380)에서는 상기 n개의 유전자들 각각에 대해, 상기 제1 암환자로부터 수집된 제1 발현량 데이터와 상기 복수의 정상 환자들로부터 수집된 제2 발현량 데이터 간의 단일 표본 t-검정을 수행함으로써, 상기 n개의 유전자들 각각에 대응되는 제1 t-검정 값을 생성한다.

[0089] 단계(S390)에서는 상기 n개의 유전자들 각각에 대응되는 분산과 제1 t-검정 값을 기초로 상기 n개의 유전자들 중 암에 대한 영향과 돌연변이 유전자에 대한 영향을 크게 받는 것으로 확인되는 k(k는 2이상의 자연수)개의 상위 유전자들과 암에 대한 영향과 돌연변이 유전자에 대한 영향을 작게 받는 것으로 확인되는 k개의 하위 유전자들을 선택한다.

[0090] 이때, 본 발명의 일실시예에 따르면, 단계(S390)에서는 상기 n개의 유전자들 각각에 대응되는 분산과 제1 t-검

정 값을 서로 곱하여 상기 n개의 유전자들 각각에 대한 스코어를 연산하는 단계 및 상기 n개의 유전자들 중 상기 스코어가 높은 순으로 상기 k개의 상위 유전자들을 선택하고, 상기 n개의 유전자들 중 상기 스코어가 낮은 순으로 상기 k개의 하위 유전자들을 선택하는 단계를 포함할 수 있다.

[0091] 이때, 본 발명의 일실시예에 따르면, 상기 전자 장치의 동작 방법은 상기 k개의 상위 유전자들과 상기 k개의 하위 유전자들이 선택되면, 상기 n x n 크기의 행렬에서 대각 성분을 모두 1로 할당하여 n x n 크기의 초기 행렬을 생성하고, 상기 초기 행렬에 대해 상기 가중치 행렬을 s(s는 2이상의 자연수)회 곱하여 n x n 크기의 제2 연산 행렬을 생성하는 단계, 상기 제2 연산 행렬을 구성하는 n개의 행벡터들 각각을 상기 n개의 유전자들 각각에 대응되는 제2 벡터로 지정한 후 상기 n개의 유전자들 각각에 대응되는 제2 벡터 중 상기 k개의 상위 유전자들에 대응되는 제2 벡터와 상기 k개의 하위 유전자들에 대응되는 제2 벡터를 선별하는 단계, 상기 k개의 상위 유전자들에 대응되는 제2 벡터에 포함된 각 성분을 상기 n개의 유전자들 각각에 대응되는 상위 영향력 값으로 하나씩 할당함으로써, 상기 n개의 유전자들 각각에 대한 k개의 상위 영향력 값들을 생성하고, 상기 k개의 하위 유전자들에 대응되는 제2 벡터에 포함된 각 성분을 상기 n개의 유전자들 각각에 대응되는 하위 영향력 값으로 하나씩 할당함으로써, 상기 n개의 유전자들 각각에 대한 k개의 하위 영향력 값들을 생성하는 단계, 상기 n개의 유전자들 각각에 대해, 상기 k개의 상위 영향력 값들과 상기 k개의 하위 영향력 값들 간의 독립 표본 t-검정을 수행함으로써, 상기 n개의 유전자들 각각에 대응되는 제2 t-검정 값을 생성하는 단계 및 상기 n개의 유전자들 중 상기 제2 t-검정 값이 높은 순으로 p(p는 2 이상의 자연수)개의 유전자들을 선택한 후 상기 선택된 p개의 유전자들을 상기 제1 암환자에 대한 암 유발 유전자로 식별하는 단계를 더 포함할 수 있다.

[0092] 이상, 도 3을 참조하여 본 발명의 일실시예에 따른 암 유발 유전자의 식별을 위한 전자 장치의 동작 방법에 대해 설명하였다. 여기서, 본 발명의 일실시예에 따른 암 유발 유전자의 식별을 위한 전자 장치의 동작 방법은 도 1과 도 2를 이용하여 설명한 암 유발 유전자의 식별을 위한 전자 장치(110)의 동작에 대한 구성과 대응될 수 있으므로, 이에 대한 보다 상세한 설명은 생략하기로 한다.

[0093] 본 발명의 일실시예에 따른 암 유발 유전자의 식별을 위한 전자 장치의 동작 방법은 컴퓨터와의 결합을 통해 실행시키기 위한 저장매체에 저장된 컴퓨터 프로그램으로 구현될 수 있다.

[0094] 또한, 본 발명의 일실시예에 따른 암 유발 유전자의 식별을 위한 전자 장치의 동작 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

[0095] 이상과 같이 본 발명에서는 구체적인 구성 요소 등과 같은 특정 사항들과 한정된 실시예 및 도면에 의해 설명되었으나 이는 본 발명의 보다 전반적인 이해를 돕기 위해서 제공된 것일 뿐, 본 발명은 상기의 실시예에 한정되는 것은 아니며, 본 발명이 속하는 분야에서 통상적인 지식을 가진 자라면 이러한 기재로부터 다양한 수정 및 변형이 가능하다.

[0096] 따라서, 본 발명의 사상은 설명된 실시예에 국한되어 정해져서는 아니되며, 후술하는 특허청구범위뿐 아니라 이 특허청구범위와 균등하거나 등가적 변형이 있는 모든 것들은 본 발명 사상의 범주에 속한다고 할 것이다.

부호의 설명

[0097] 110: 암 유발 유전자의 식별을 위한 전자 장치

111: 제1 데이터 저장부 112: 제2 데이터 저장부

113: 가중치 저장부 114: 돌연변이 정보 저장부

115: 유전자 행렬 생성부 116: 가중치 행렬 생성부

117: 분산 산출부 118: 제1 검정부

119: 유전자 선택부 120: 스코어 연산부

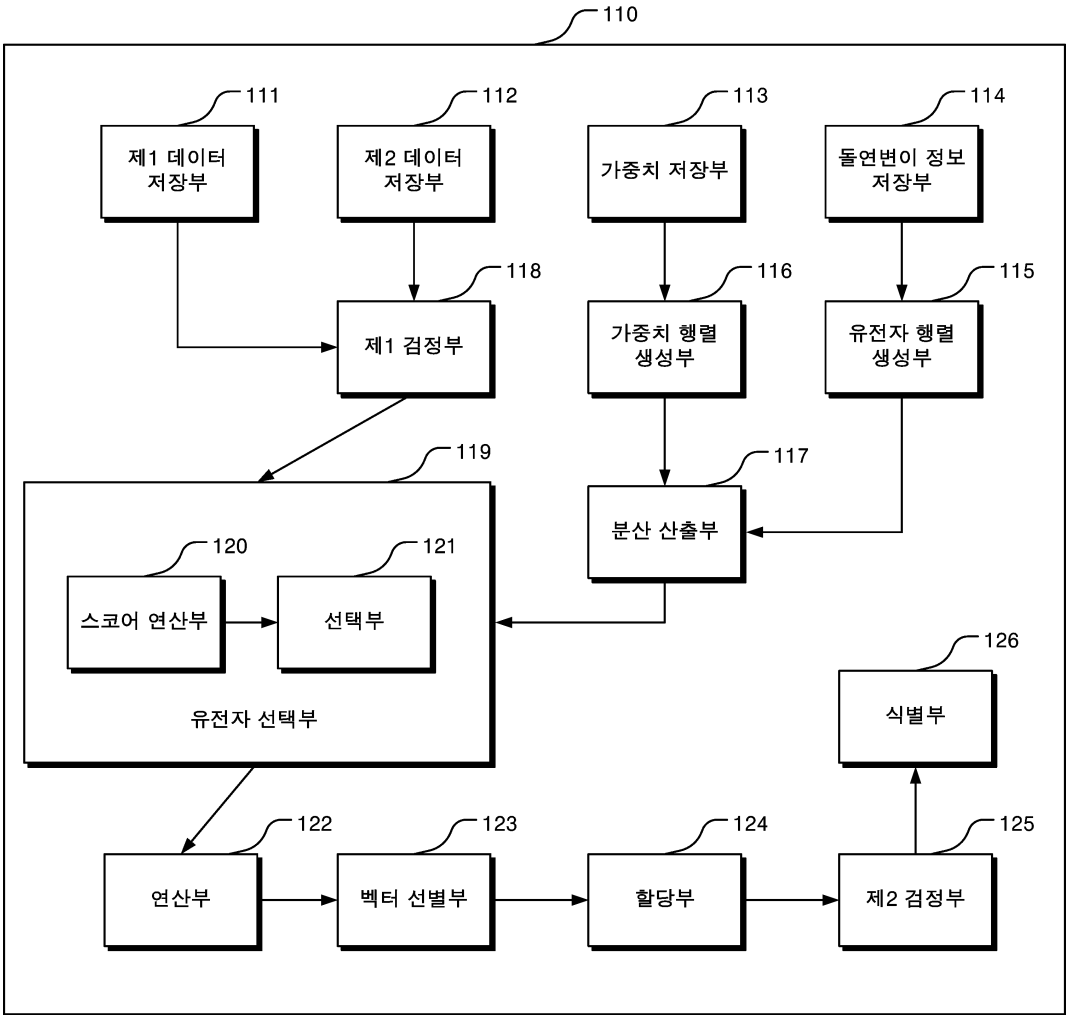
121: 선택부 122: 연산부

123: 벡터 선별부 124: 할당부

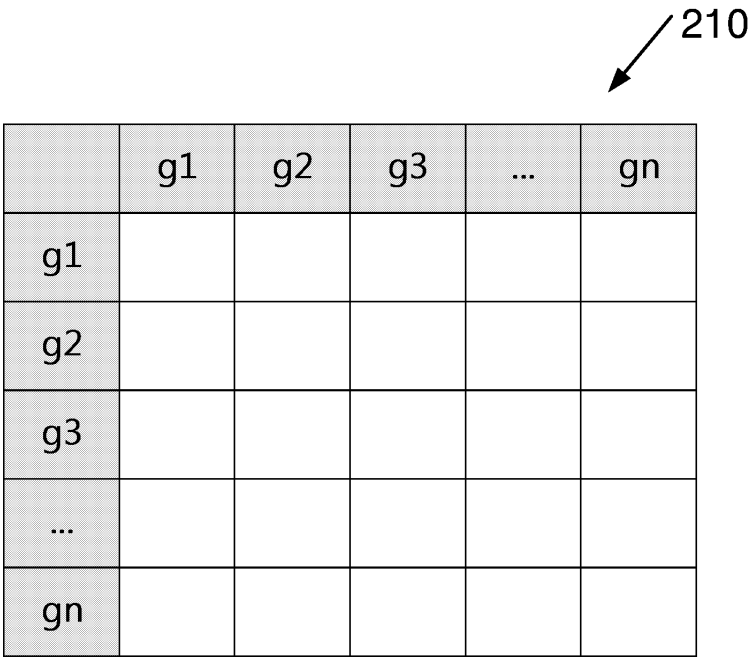
125: 제2 검정부 126: 식별부

도면

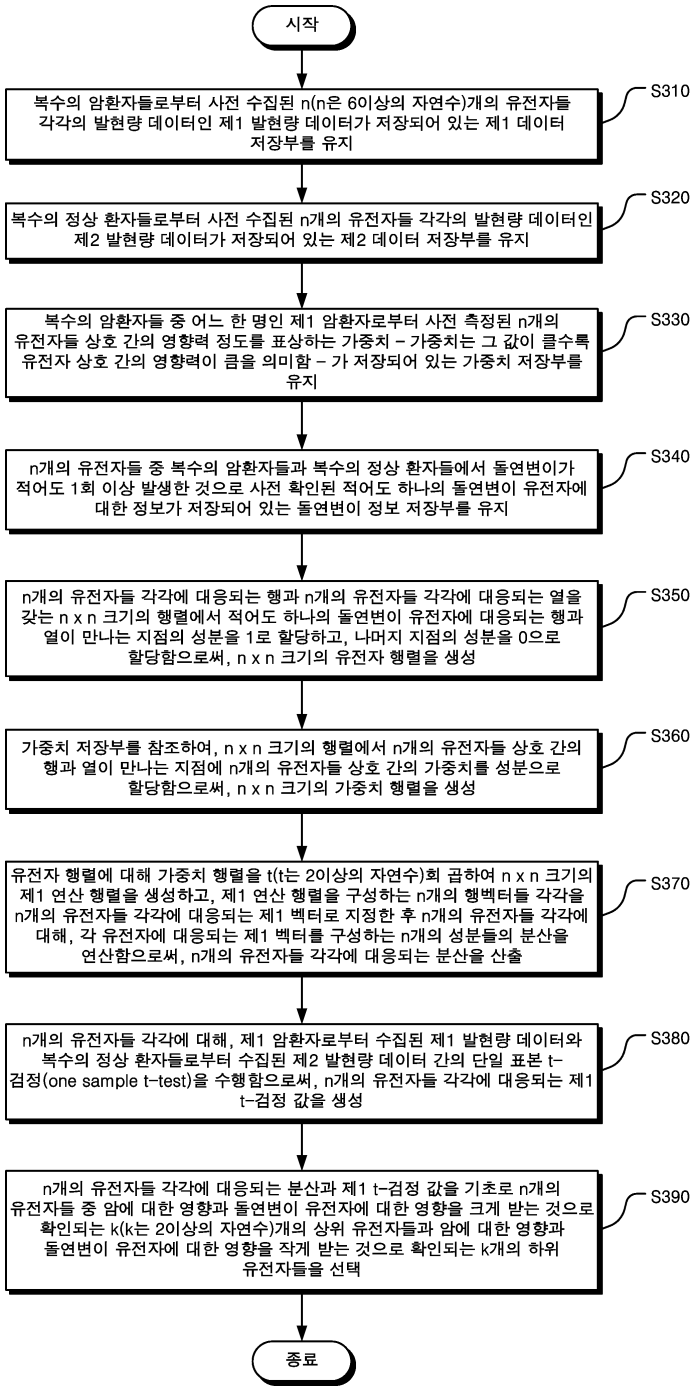
도면1



도면2



도면3


시작
복수의 암환자들로부터 사전 수집된 n(n은 6이상의 자연수)개의 유전자들 각각의 발현량 데이터인 제1 발현량 데이터가 저장되어 있는 제1 데이터 저장부를 유지
S310
복수의 정상 환자들로부터 사전 수집된 n개의 유전자들 각각의 발현량 데이터인 제2 발현량 데이터가 저장되어 있는 제2 데이터 저장부를 유지
S320
복수의 암환자들 중 어느 한 명인 제1 암환자로부터 사전 측정된 n개의 유전자들 상호 간의 영향력 정도를 표상하는 가중치 – 가중치는 그 값이 클수록 유전자 상호 간의 영향력이 큼을 의미함 – 가 저장되어 있는 가중치 저장부를 유지
S330
n개의 유전자들 중 복수의 암환자들과 복수의 정상 환자들에서 돌연변이가 적어도 1회 이상 발생한 것으로 사전 확인된 적어도 하나의 돌연변이 유전자에 대한 정보가 저장되어 있는 돌연변이 정보 저장부를 유지
S340
n개의 유전자들 각각에 대응되는 행과 n개의 유전자들 각각에 대응되는 열을 갖는 n x n 크기의 행렬에서 적어도 하나의 돌연변이 유전자에 대응되는 행과 열이 만나는 지점의 성분을 1로 할당하고, 나머지 지점의 성분을 0으로 할당함으로써, n x n 크기의 유전자 행렬을 생성
S350
가중치 저장부를 참조하여, n x n 크기의 행렬에서 n개의 유전자들 상호 간의 행과 열이 만나는 지점에 n개의 유전자들 상호 간의 가중치를 성분으로 할당함으로써, n x n 크기의 가중치 행렬을 생성
S360
유전자 행렬에 대해 가중치 행렬을 t(t는 2이상의 자연수)회 곱하여 n x n 크기의 제1 연산 행렬을 생성하고, 제1 연산 행렬을 구성하는 n개의 행벡터들 각각을 n개의 유전자들 각각에 대응되는 제1 벡터로 지정한 후 n개의 유전자들 각각에 대해, 각 유전자에 대응되는 제1 벡터를 구성하는 n개의 성분들의 분산을 연산함으로써, n개의 유전자들 각각에 대응되는 분산을 산출
S370
n개의 유전자들 각각에 대해, 제1 암환자로부터 수집된 제1 발현량 데이터와 복수의 정상 환자들로부터 수집된 제2 발현량 데이터 간의 단일 표본 t-검정(one sample t-test)을 수행함으로써, n개의 유전자들 각각에 대응되는 제1 t-검정 값을 생성
S380
n개의 유전자들 각각에 대응되는 분산과 제1 t-검정 값을 기초로 n개의 유전자들 중 암에 대한 영향과 돌연변이 유전자에 대한 영향을 크게 받는 것으로 확인되는 k(k는 2이상의 자연수)개의 상위 유전자들과 암에 대한 영향과 돌연변이 유전자에 대한 영향을 작게 받는 것으로 확인되는 k개의 하위 유전자들을 선택
S390
종료