



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2019-0032847
(43) 공개일자 2019년03월28일

(51) 국제특허분류(Int. Cl.)

G16B 5/00 (2019.01)

(52) CPC특허분류

G16B 5/00 (2019.02)

G16B 40/00 (2019.02)

(21) 출원번호 10-2017-0121201

(22) 출원일자 2017년09월20일

심사청구일자 2017년09월20일

(71) 출원인

이화여자대학교 산학협력단

서울특별시 서대문구 이화여대길 52 (대현동, 이화여자대학교)

(72) 발명자

이형준

서울특별시 성동구 매봉길 50, 120동 205호(옥수동, 옥수 파크힐스)

이민수

경기도 성남시 분당구 수내로 206, 312동 202호(수내동, 푸른마을 아파트)

(74) 대리인

특허법인(유한)아이시스

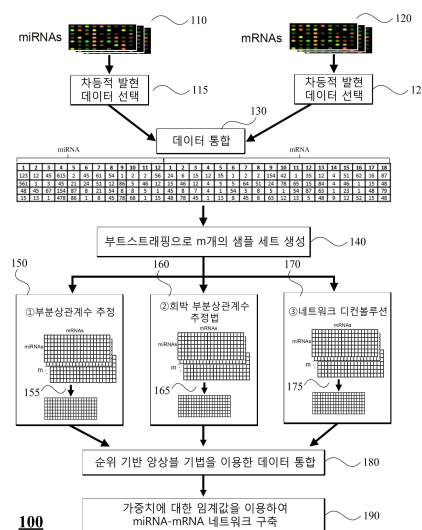
전체 청구항 수 : 총 12 항

(54) 발명의 명칭 miRNA-mRNA 연관도 분석 방법 및 miRNA-mRNA 네트워크 생성 장치

(57) 요약

miRNA-mRNA 연관도 분석 방법은 컴퓨터 장치가 복수의 miRNA의 발현 데이터 및 복수의 mRNA의 발현 데이터를 포함하는 샘플 데이터를 생성하는 단계, 상기 컴퓨터 장치가 상기 복수의 miRNA 중 적어도 하나와 상기 복수의 mRNA 중 적어도 하나에 대한 복수의 제1 연관도 모델을 생성하는 단계 및 상기 컴퓨터 장치가 상기 복수의 제1 연관도 모델을 통합하여 제2 연관도 모델을 생성하는 단계를 포함한다.

대표도 - 도3



이 발명을 지원한 국가연구개발사업

과제고유번호 2015R1D1A1A01057902

부처명 교육부

연구관리전문기관 한국연구재단

연구사업명 기초연구사업(학술진흥)-이공학개인지초연구지원사업-기본연구

연구과제명 재난 상황 시 무인비행체의 분산적 협업 탐색 및 효율적인 네트워크 센싱을 통한 네트워크 복구 시스템 연구

기 여 율 1/2

주관기관 이화여자대학교 산학협력단

연구기간 2016.11.01 ~ 2018.10.31

이 발명을 지원한 국가연구개발사업

과제고유번호 2015R1C1A1A01054305

부처명 미래창조과학부

연구관리전문기관 한국연구재단

연구사업명 이공분야 기초연구사업-신진연구자지원사업-신진연구

연구과제명 개념 변화와 불균형 데이터 문제를 고려한 적응형 빅스트림 마이닝 시스템 기술 연구

기 여 율 1/2

주관기관 이화여자대학교 산학협력단

연구기간 2015.11.01 ~ 2018.10.31

명세서

청구범위

청구항 1

컴퓨터 장치가 복수의 miRNA의 발현 데이터 및 복수의 mRNA의 발현 데이터를 포함하는 샘플 데이터를 생성하는 단계;

상기 컴퓨터 장치가 상기 복수의 miRNA 중 적어도 하나와 상기 복수의 mRNA 중 적어도 하나에 대한 복수의 제1 연관도 모델을 생성하는 단계; 및

상기 컴퓨터 장치가 상기 복수의 제1 연관도 모델을 통합하여 제2 연관도 모델을 생성하는 단계를 포함하는 miRNA-mRNA 연관도 분석 방법.

청구항 2

제1항에 있어서,

상기 컴퓨터 장치는 상기 miRNA의 발현 데이터 및 상기 mRNA의 발현 데이터에서 차별 발현한 정보를 식별하고, 식별한 miRNA의 발현 데이터 및 mRNA의 발현 데이터를 통합하여 상기 샘플 데이터를 생성하는 miRNA-mRNA 연관도 분석 방법.

청구항 3

제1항에 있어서,

상기 컴퓨터 장치는 부분상관계수 추정법, SPACE(Sparse PartialCorrelation Estimation) 및 네트워크 디컨볼루션(deconvolution) 중 적어도 하나를 사용하여 상기 제1 연관도 모델을 생성하는 miRNA-mRNA 연관도 분석 방법.

청구항 4

제1항에 있어서,

상기 컴퓨터 장치는 상기 샘플 데이터에 부분상관계수 추정법, SPACE(Sparse PartialCorrelation Estimation) 및 네트워크 디컨볼루션(deconvolution)을 각각 적용하여 상기 복수의 제1 연관도 모델을 생성하는 miRNA-mRNA 연관도 분석 방법.

청구항 5

제1항에 있어서,

상기 컴퓨터 장치는 상기 샘플 데이터에 대한 부트스트래핑(bootstrapping)을 수행하여 복수의 샘플 데이터 세트를 생성하고,

상기 복수의 샘플 데이터 세트에 각각 연관도 추정 기법인 부분상관계수 추정법, SPACE(Sparse PartialCorrelation Estimation) 및 네트워크 디컨볼루션(deconvolution) 중 적어도 하나를 적용하여 복수의 연관도 모델을 생성하고, 동일한 연관도 추정 기법으로 생성한 복수의 연관도 모델을 역 순위 기반 앙상블(ensemble) 기법을 사용하여 복수의 연관도 모델을 하나의 모델로 통합하는 단계를 더 포함하는 miRNA-mRNA 연관도 분석 방법.

청구항 6

제1항에 있어서,

상기 컴퓨터 장치는 역 순위 기반 앙상블(ensemble) 기법을 사용하여 상기 복수의 제1 연관도 모델을 상기 제2 연관도 모델로 통합하는 miRNA-mRNA 연관도 분석 방법.

청구항 7

제1항에 있어서,

상기 컴퓨터 장치는 상기 제2 연관도 모델을 이용하여 miRNA-mRNA 네트워크를 생성하는 단계를 더 포함하는 miRNA-mRNA 연관도 분석 방법.

청구항 8

컴퓨터에서 청구항 제1항 내지 제7항의 항 중 어느 하나의 항에 따른 miRNA-mRNA 연관도 분석 방법을 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체.

청구항 9

복수의 miRNA의 발현 데이터 및 복수의 mRNA의 발현 데이터를 입력받는 입력 장치;

miRNA-mRNA 네트워크 구축을 위한 프로그램을 저장하는 저장 장치; 및

상기 복수의 miRNA의 발현 데이터 및 복수의 mRNA의 발현 데이터를 입력데이터로 사용하는 상기 프로그램을 실행하여 상기 복수의 miRNA의 발현 데이터 및 복수의 mRNA의 발현 데이터를 입력데이터를 포함하는 샘플 데이터를 생성하고, 상기 샘플 데이터를 이용하여 상기 복수의 miRNA 중 적어도 하나와 상기 복수의 mRNA 중 적어도 하나에 대한 복수의 제1 연관도 모델을 생성하고, 상기 복수의 제1 연관도 모델을 통합하여 제2 연관도 모델을 생성하고, 상기 제2 연관도 모델을 기준으로 상기 복수의 miRNA 중 적어도 하나와 상기 복수의 mRNA 중 적어도 하나에 대한 연관도를 나타내는 miRNA-mRNA 네트워크를 생성하는 연산 장치를 포함하는 miRNA-mRNA 네트워크 생성 장치.

청구항 10

제9항에 있어서,

상기 연산 장치는 상기 miRNA의 발현 데이터 및 상기 mRNA의 발현 데이터에서 차별 발현한 정보를 식별하고, 식별한 miRNA의 발현 데이터 및 mRNA의 발현 데이터를 통합하여 상기 샘플 데이터를 생성하는 miRNA-mRNA 네트워크 생성 장치.

청구항 11

제9항에 있어서,

상기 연산 장치는 상기 샘플 데이터에 부분상관계수 추정법, SPACE(Sparse PARTialCorrelation Estimation) 및 네트워크 디컨볼루션(deconvolution)을 각각 적용하여 상기 복수의 제1 연관도 모델을 생성하는 miRNA-mRNA 네트워크 생성 장치.

청구항 12

제9항에 있어서,

상기 연산 장치는 상기 컴퓨터 장치는 역 순위 기반 앙상블(ensemble) 기법을 사용하여 상기 복수의 제1 연관도 모델을 상기 제2 연관도 모델로 통합하되, 상기 복수의 제1 연관도 모델에서 동일한 에지에 대한 역 순위를 곱한 값을 상기 동일한 에지의 최종 가중치로 설정하고, 상기 제2 연관도 모델로 통합하는 miRNA-mRNA 네트워크 생성 장치.

발명의 설명

기술 분야

[0001] 이하 설명하는 기술은 miRNA와 mRNA의 연관성을 분석하는 기법에 관한 것이다.

배경 기술

[0002] mRNA(messenger RNA)는 mRNA는 DNA 원형으로부터 전사되고, 유전정보에 암호(codon)를 부여하여 단백질 합성이 일어나는 장소인 리보솜으로 운반한다. mRNA는 단백질 합성을 위한 정보를 전달하여 단백질 합성에 직접 관여한

다. 마이크로 RNA(miRNA)는 식물, 동물, 바이러스 등에서 발견되며 약 22개의 뉴클레오타이드로 구성된 작은 비발현 RNA 분자이다. miRNA는 특정 mRNA에 대한 번역(translation) 과정을 억제(제어)한다. 즉 miRNA는 단백질 합성을 위한 코돈 정보를 제공하지 않지만, 단백질 합성을 일정하게 제어하는 기능을 수행한다. miRNA는 세포분화, 세포간 신호 전달 또는 특정 종양과 같은 생물학적 과정에 광범위하게 관여한다고 알려졌다. miRNA는 유전자 발현을 제어하는 과정에서 중요한 역할을 한다. 따라서 특정 miRNA와 특정 mRNA 사이의 기능적 연관성을 갖는다고 하겠다. miRNA와 mRNA 사이의 연관성에 대한 연구가 수행되고 있다.

[0003] miRNA는 mRNA의 3'-UTR(3'-untranslated regions)에 있는 상보적 서열에 결합하여 특정 mRNA에 대한 전사후 조절을 한다. 이러한 특성에 기반하여 서열의 상보성 또는 구조적 안정성을 기준으로 서열 데이터를 이용한 miRNA와 mRNA 사이의 연관성 연구가 있었다.

[0004] 특정 miRNA가 특정 mRNA를 제어하기 위하여, 해당 miRNA의 발현 수준이 변경될 수 있다. 즉, miRNA의 발현과 mRNA의 발현 사이에 일정한 상관 관계가 있다. 이러한 특성에 기반하여 miRNA와 mRNA 사이의 연관성을 확인하기 위한 네트워크 구축을 위한 연구가 있었다.

선행기술문헌

비특허문헌

[0005] (비특허문헌 0001) Joung JG, Hwang KB, Nam JW, Kim SJ, Zhang BT. Discovery of microRNA-mRNA modules via population-based probabilistic learning. Bioinformatics. 2007;23:1141-7p.

발명의 내용

해결하려는 과제

[0006] (1) 서열 데이터 기반한 분석방법은 특정 miRNA와 mRNA 사이의 연관성을 분석하는데 적합할 수 있지만, 높은 위양성(false positive) 및 위음성(false negative)을 갖는다. (2) miRNA와 mRNA의 발현의 상관성을 이용한 분석방법은 (i)유전자 발현 데이터를 이용하여 miRNA가 mRNA에 간접적으로 영향을 주는 것을 배제하기 어렵다. 즉 miRNA와 mRNA의 네트워크에 잘못된 에지(edge)가 존재할 가능성이 높다. (ii) 또한 마이크로 어레이 데이터는 기본적으로 한정된 표본으로부터 추출된다. 따라서 miRNA와 mRNA의 연관성에 대한 신뢰성에 문제(High-dimension Low-sample size problem)가 있다. (iii) 나아가 miRNA와 mRNA의 연관성 분석을 위한 특정한 발현 데이터에 대하여 어떠한 분석 기법이 더욱 유효한지 사전에 알기 어려울 수도 있다. 예컨대, 다양한 실험 환경에서 하나의 기법을 적용하면 실험 결과가 일관되지 않을 수 있다.

[0007] 이하 설명하는 기술은 종래 문제점을 해결하고 miRNA와 mRNA 사이의 연관성을 분석할 수 있는 기법을 제공하고자 한다.

과제의 해결 수단

[0008] miRNA-mRNA 연관도 분석 방법은 컴퓨터 장치가 복수의 miRNA의 발현 데이터 및 복수의 mRNA의 발현 데이터를 포함하는 샘플 데이터를 생성하는 단계, 상기 컴퓨터 장치가 상기 복수의 miRNA 중 적어도 하나와 상기 복수의 mRNA 중 적어도 하나에 대한 복수의 제1 연관도 모델을 생성하는 단계 및 상기 컴퓨터 장치가 상기 복수의 제1 연관도 모델을 통합하여 제2 연관도 모델을 생성하는 단계를 포함한다.

[0009] miRNA-mRNA 네트워크 생성 장치는 복수의 miRNA의 발현 데이터 및 복수의 mRNA의 발현 데이터를 입력받는 입력장치, miRNA-mRNA 네트워크 구축을 위한 프로그램을 저장하는 저장장치 및 상기 복수의 miRNA의 발현 데이터 및 복수의 mRNA의 발현 데이터를 입력데이터로 사용하는 상기 프로그램을 실행하여 상기 복수의 miRNA의 발현 데이터 및 복수의 mRNA의 발현 데이터를 입력데이터를 포함하는 샘플 데이터를 생성하고, 상기 샘플 데이터를 이용하여 상기 복수의 miRNA 중 적어도 하나와 상기 복수의 mRNA 중 적어도 하나에 대한 복수의 제1 연관도 모델을 생성하고, 상기 복수의 제1 연관도 모델을 통합하여 제2 연관도 모델을 생성하고, 상기 제2 연관도 모델을 기준으로 상기 복수의 miRNA 중 적어도 하나와 상기 복수의 mRNA 중 적어도 하나에 대한 연관도를 나타내는 miRNA-mRNA 네트워크를 생성하는 연산 장치를 포함한다.

발명의 효과

[0010] 이하 설명하는 기술은 적은 샘플 데이터를 이용하는 경우에도 miRNA와 mRNA의 직접적 연관성을 신뢰도 높게 나타내는 miRNA-mRNA 네트워크를 제공한다.

도면의 간단한 설명

[0011] 도 1은 miRNA-mRNA 네트워크를 구축하는 과정에 대한 예이다.

도 2는 miRNA-mRNA 네트워크에 대한 예이다.

도 3은 miRNA-mRNA 네트워크 구축을 위한 절차 흐름도의 예이다.

도 4는 miRNA-mRNA 네트워크 구축 장치에 대한 예이다.

도 5는 제안 기법의 효과를 검증한 실험 결과의 예이다.

발명을 실시하기 위한 구체적인 내용

[0012] 이하 설명하는 기술은 다양한 변경을 가할 수 있고 여러 가지 실시례를 가질 수 있는 바, 특정 실시례들을 도면에 예시하고 상세하게 설명하고자 한다. 그러나, 이는 이하 설명하는 기술을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 이하 설명하는 기술의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.

[0013] 제1, 제2, A, B 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 해당 구성요소들은 상기 용어들에 의해 한정되지는 않으며, 단지 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 이하 설명하는 기술의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. 및/또는 이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.

[0014] 본 명세서에서 사용되는 용어에서 단수의 표현은 문맥상 명백하게 다르게 해석되지 않는 한 복수의 표현을 포함하는 것으로 이해되어야 하고, "포함한다" 등의 용어는 설시된 특징, 개수, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함을 의미하는 것이지, 하나 또는 그 이상의 다른 특징들이나 개수, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 배제하지 않는 것으로 이해되어야 한다.

[0015] 도면에 대한 상세한 설명을 하기에 앞서, 본 명세서에서의 구성부들에 대한 구분은 각 구성부가 담당하는 주기능 별로 구분한 것에 불과함을 명확히 하고자 한다. 즉, 이하에서 설명할 2개 이상의 구성부가 하나의 구성부로 합쳐지거나 또는 하나의 구성부가 보다 세분화된 기능별로 2개 이상으로 분화되어 구비될 수도 있다. 그리고 이하에서 설명할 구성부 각각은 자신이 담당하는 주기능 이외에도 다른 구성부가 담당하는 기능 중 일부 또는 전부의 기능을 추가적으로 수행할 수도 있으며, 구성부 각각이 담당하는 주기능 중 일부 기능이 다른 구성부에 의해 전담되어 수행될 수도 있음은 물론이다.

[0016] 또, 방법 또는 동작 방법을 수행함에 있어서, 상기 방법을 이루는 각 과정들은 문맥상 명백하게 특정 순서를 기재하지 않은 이상 명기된 순서와 다르게 일어날 수 있다. 즉, 각 과정들은 명기된 순서와 동일하게 일어날 수도 있고 실질적으로 동시에 수행될 수도 있으며 반대의 순서대로 수행될 수도 있다.

[0018] 이하 설명하는 기술은 마이크로 RNA(micro RNA, miRNA)와 mRNA의 연관성을 분석하는 기법에 관한 것이다. 이하 설명하는 기술은 마이크로어레이(micro array) 데이터를 이용하여 분석 대상인 miRNA와 mRNA의 연관성을 분석한다.

[0019] 이하 설명하는 기술은 유전자 발현 데이터로부터 miRNA와 mRNA의 연관성을 나타내는 miRNA-mRNA 네트워크를 구축한다. miRNA-mRNA 네트워크는 miRNA와 mRNA의 연관성을 나타내기 위한 그래프 자료 구조를 의미한다.

[0021] 도 1은 miRNA-mRNA 네트워크를 구축하는 과정에 대한 예이다. 먼저 miRNA의 발현 데이터 및 mRNA의 발현 데이터를 획득한다(A). 유전자 발현 데이터는 예컨대, 마이크로어레이 실험의 결과물일 수 있다. 유전자 발현 데이터는 도 1과 같이 마이크로어레이 실험을 수행한 출력 데이터 내지 영상 데이터일 수 있다. 연구자가 유전자 발현

데이터를 이용하여 특정 유전자의 발현 상태를 도출할 수 있다. 또는 컴퓨터 장치가 자동으로 마이크로어레이 실험 결과 데이터를 읽고 분석하여 특정 유전자의 발현 상태를 도출할 수 있다. 이후 컴퓨터 장치가 miRNA-mRNA 네트워크를 구축하므로 유전자 발현 데이터는 디지털 데이터 형태로 변환된다고 전제한다. 이후 컴퓨터 장치가 miRNA 및 mRNA의 상관성을 나타내는 모델(행렬)을 도출한다(B). 마지막으로 miRNA 및 mRNA의 상관성을 나타내는 모델을 이용하여 컴퓨터 장치가 miRNA-mRNA 네트워크를 구축한다. 자세한 과정에 대해서는 후술한다.

[0023] 도 2는 miRNA-mRNA 네트워크에 대한 예이다. 도 2는 도 1(C)에서 생성되는 결과물에 해당한다. 도 2에서 원은 mRNA를 나타내고, 사각형은 miRNA를 나타낸다. 분석 결과 도출되는 그래프에서 노드(miRNA와 mRNA)가 에지로 연결되면, 해당 miRNA와 mRNA는 직접적 연관성을 갖는다는 의미이다. 도 2의 miRNA-mRNA 네트워크에서 에지로 연결된 miRNA와 mRNA는 직접적 연관성을 갖는다. 예컨대, miRNA a는 4개의 mRNA 1, 2, 3 및 4와 연관성을 갖는다. 즉 miRNA a는 mRNA 1, 2, 3 및 4의 제어 기작에 관여한다. mRNA 4는 3개의 miRNA a, b 및 d와 연관성을 갖는다. 즉, miRNA a, b 및 d는 mRNA 4의 제어 기작에 관여한다.

[0025] 이하 설명하는 기술은 직접적 연관성을 갖는 miRNA-mRNA 네트워크를 구축하기 위하여 몇 가지 기법을 사용한다. 간략하게 설명하면 (1) 직접적 연관성을 식별하기 위한 기법, (2) 부트스트래핑(Bootstrapping) 기법 및 (3) 앙상블(Ensemble) 기법을 사용한다. 첫 째, 발현 데이터로부터 예측되는 간접적 연관성을 배제한다. 직접적 연관성만을 식별하기 위하여 ①부분상관계수 추정법(Partial correlation estimation), ②희박 부분상관계수 추정법(Sparse PARTIALCorrelation Estimation, SPACE) 및 ③ 네트워크 디컨볼루션(Network deconvolution) 중 적어도 하나를 사용할 수 있다. 이 방법은 miRNA와 mRNA의 직접적인 상관성을 유추하기 위한 것이다. 둘 째, 적은 샘플로 인한 문제(High-dimension Low-sample size problem)를 해결하기 위하여 실험에서 서로 다른 양상으로 발현된 miRNA와 mRNA에 대한 데이터를 식별하여 데이터 세트의 차원을 줄인다. 또한 한정된 샘플 크기를 최대한 활용하고자 부트스트래핑 기법을 사용한다. 셋 째, 유추된 miRNA와 mRNA의 연관성의 신뢰도를 높이기 위하여, 비모수 앙상블(non-parametric Ensemble) 기법을 사용한다. 부트스트래핑 기법과 서로 다른 직접 연관성 식별 결과를 통합적으로 사용하기 위하여 순위 기반 비모수 앙상블 기법(rank-based non-parametric Ensemble)을 사용한다.

[0027] 도 3은 miRNA-mRNA 네트워크 구축을 위한 절차 흐름도의 예이다. 이하 miRNA-mRNA 네트워크 구축 과정은 컴퓨터 장치에 수행된다고 전제한다.

[0029] 먼저 miRNA와 mRNA의 연관성을 분석하기 위하여 매칭되는 miRNA와 mRNA의 발현 데이터가 필요하다. 이후 분석 대상은 매칭된 miRNA와 mRNA의 발현 데이터라고 전제한다. 컴퓨터 장치는 miRNA의 발현 데이터 및 mRNA의 발현 데이터를 입력 받는다(110, 120). 컴퓨터 장치는 발현 데이터를 이용하여 특정 유전자의 발현 상태를 확인할 수 있다고 전제한다. 한편 컴퓨터 장치는 입력된 발현 데이터를 일정하게 전처리할 수 있다. 전처리는 이후 분석을 위하여 발현 데이터를 정규화하는 과정이다.

[0030] 컴퓨터 장치는 miRNA의 발현 데이터와 mRNA의 발현 데이터에서 차별 발현 유전자(differentially expressed gene)를 식별하여 데이터의 차원을 줄인다(115, 125). 차별 발현 유전자는 대조검체에 비하여 실험검체에서 유의하게 발현이 증가하거나 감소하는 유전자를 의미한다. 통상적으로 마이크로어레이 데이터에서 평균치와 비교하면서 차별 발현 유전자를 식별한다. 예컨대, SAM(Significance Analysis of Microarray) 방법 등을 사용할 수 있다.

[0032] 이는 활성적인 miRNA와 mRNA의 연관성에 집중하기 위한 것이다. miRNA 발현 데이터와 mRNA 발현 데이터는 서로 다른 플랫폼에서 획득하기 때문에, 선택된 발현 데이터는 통합되어야 한다. 컴퓨터 장치는 miRNA와 mRNA 발현 데이터 중 선택된 데이터를 통합(integration)한다(130). 도 3에서 발현 데이터를 나타내는 테이블에서 가장 위에 위치하는 숫자는 유전자의 식별자에 해당한다. 또한 컴퓨터 장치는 miRNA 발현 데이터와 mRNA 발현 데이터의 스케일링(scaling)을 수행할 수도 있다.

[0033] 나아가 컴퓨터 장치는 샘플의 크기가 작은 문제를 해결하기 위하여 통합된 miRNA와 mRNA 발현 데이터에 대한 부

트스트래핑을 수행한다(140). 부트스트래핑은 m개의 새로운 훈련 데이터 세트를 생성한다. 부트스트래핑은 하나의 모델에 대한 다양한 버전(version)을 생성하는 기법이다. 부트스트래핑은 널리 알려진 기법이므로 자세한 설명은 생략한다. 부트스트래핑은 모집단 샘플에서 표본 데이터를 반복적으로 추출하여 새로운 데이터 세트를 생성한다. 새로운 데이터 세트에 포함된 데이터는 본래의 표본 데이터에 값으로 구성된다. 전술한 통합된 miRNA와 mRNA 발현 데이터에 대해 m번 부트스트래핑을 수행하면 m개의 새로운 데이터 세트가 생성된다.

[0034] 컴퓨터 장치는 직접적 miRNA와 mRNA의 연관성을 도출하기 위하여 통합된 miRNA 발현 데이터와 mRNA 발현 데이터에 3가지 직접적 상관도 유추 기법을 적용한다(150, 160 및 170). 직접적 연관성을 도출하기 위한 추론 기법은 표현형 데이터로부터 직접적 상관 모델을 생성한다. 상관 모델은 miRNA와 mRNA 사이에서 가능한 모든 조합을 포함하는 행렬 형태를 갖는다. 컴퓨터 장치는 3가지 직접적 상관도 유추 기법 각각에 대하여 부트스트래핑 결과인 m개의 샘플을 이용하여 직접적 상관도를 유추한다. 각 상관도 유추 기법은 m개의 샘플을 이용하여 m개의 모델을 생성한다. 컴퓨터 장치는 각 상관도 유추 기법이 생성한 m개의 모델은 순위 기반 통합 기법(rank-based aggregation)을 사용하여 하나의 모델을 생성한다(155, 165 및 175). 도 3에서는 3가지 기법을 모두 사용하는 예를 도시하였다. 다만 경우에 따라서는 도 3에서 설명한 3가지 기법(부분상관계수 추정법, 회박 부분상관계수 추정법 및 네트워크 디컨볼루션) 중 적어도 하나 이상을 이용할 수도 있다.

[0035] 서로 다른 상관도 유추 기법이 생성한 모델을 통합하기 위하여, 컴퓨터 장치는 순위 기반 앙상블 기법을 사용한다(180). 이를 통해 컴퓨터 장치는 최종적으로 하나의 직접 상관 모델을 생성한다.

[0036] 최종적인 직접 상관 모델(행렬)의 값에 대하여 일정한 임계값을 적용하여 직접 miRNA-mRNA 네트워크를 구축할 수 있다(190). 예컨대, 임계값이 50이라면, 컴퓨터 장치는 직접 상관 모델에서 50 미만은 필터링(무시)하고, 50 이상인 값을 갖는 행렬이 직접적 상관관계를 갖는다고 판단한다. 컴퓨터 장치는 행렬의 값이 50 이상인 miRNA와 mRNA 쌍에 대하여 에지를 연결하여 miRNA-mRNA 네트워크를 구축할 수 있다.

[0038] 3가지 직접적 상관도 유추 기법에 대해 구체적으로 설명한다. 직접적 상관도 유추 기법은 전술한 바와 같이 miRNA-mRNA 네트워크에 나타날 수 있는 간접적 상관을 배제하기 위한 것이다.

[0040] ①부분상관계수 추정법(Partial correlation estimation)

[0041] 부분상관계수 추정법은 무작위 변수 세트의 효과를 억제하여 두 개의 무작위 변수 사이의 연관성 가중치를 측정한다. 부분상관에 기반한 방법은 공분산 행렬의 역행렬인 집중행렬(concentration matrix)에서 0(zero)이 아닌 항목을 통해 조건부 종속성(conditional dependency)을 유추할 수 있다. 부분상관계수 추정법을 적용하여 유전자 네트워크를 분석하면, 0인 항목은 두 개의 노드 사이에 직접적인 영향이 없는 관계를 의미한다. 즉 0이 아닌 항목은 두 개의 노드 사이에 직접적 연관성이 있다는 것을 의미한다. 부분상관계수 추정을 위하여 종래 연구된 기법을 이용할 수도 있다(예컨대, Schafer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statist Appl Genet Mol Biol. 2005;4:32). 부분상관계수 추정법을 이용하여 miRNA 발현 데이터와 mRNA의 발현 데이터에서 직접적 연관성을 갖는 하나 이상의 miRNA와 mRNA를 추정할 수 있다.

[0043] ②회박 부분상관계수 추정법(SPACE)

[0044] SPACE는 적은 샘플을 갖는 상황에서 부분상관계수를 연산하는 다른 기법이다(Peng J, Wang P, Zhou N, Zhu J. Partial correlation estimation by joint sparse regression models. J Am Stat Assoc - Theory and Methods. 2009;104(486):735-46. 참조). SPACE는 부분상관을 나타내는 행렬이 희소(sparse)하고, 대부분의 변수 쌍이 조건부 독립성을 갖는다고 가정한다. SPACE는 행렬의 요소가 대부분 0인 희소 행렬을 산출한다. SPACE는 결국 0이 아닌 부분의 상관관계를 추정하는데 도움이 된다. SPACE는 희소 회귀기법을 사용하여 희소 부분상관을 추정한다.

[0046] ③ 네트워크 디컨볼루션(Network deconvolution)

[0047] 네트워크 디컨볼루션 기법은 입력 데이터로부터 종속성 네트워크의 간접적 가중치를 제거하여 직접적 종속성 네

트위크를 추론하는 기법이다(Feizi S, Marbach D, Medard M, Kellis M. Network deconvolution as a general method to distinguish direct dependencies in networks. Nat Biotechnol. 2013;31:726-33. 참조). 네트워크 디컨볼루션 기법은 입력 데이터로부터 측정된 연관성 가중치가 직접 및 간접 가중치의 합이라고 가정한다. 나아가 네트워크 디컨볼루션 기법은 간접 정보 흐름이 직접 연관성 가중치로부터 추정할 수 있다고 가정한다.

[0048] G_{obs} 를 관측된 독립성 네트워크, G_{tru} 를 진정한 독립성 네트워크, G_{ind} 를 간접적 독립성 네트워크라고 하자. G_{obs} 를 아래의 수학적 식 1과 같이 표현할 수 있다.

수학적 식 1

$$\begin{aligned} G_{obs} &= G_{tru} + G_{ind} \\ &= G_{tru} + (G_{tru}^2 + G_{tru}^3 + G_{tru}^4 + \dots) \\ &= G_{tru}(I - G_{tru})^{-1} \end{aligned}$$

[0050]

[0052] 따라서 네트워크 디컨볼루션 기법은 모든 가능한 간접적 경로를 지나가는 경우 정보의 효과를 역으로 구성하여 진정한 직접 종속성 네트워크를 도출한다. 결국 진정한 직접 종속성 네트워크는 아래의 수학적 식 2와 같다.

수학적 식 2

$$G_{tru} = G_{obs}(I + G_{obs})^{-1}$$

[0054]

[0056] 컴퓨터 장치는 상호적 정보를 사용하는 miRNA와 mRNA 표현 데이터 사이에서 관측된 상관을 추정한다. 그리고 컴퓨터 장치는 디컨볼루션 기법을 miRNA와 mRNA 사이의 상관 모델에 적용하여 간접적 상관을 배제한다.

[0058] 전술한 3가지 직접적 상관도 유추 기법은 각각 하나의 모델(행렬로 표현)을 생성한다. 이제 서로 다른 기법으로 miRNA와 mRNA의 상관성을 유추한 모델을 통합하기 위한 절차를 설명한다. 컴퓨터 장치는 전술한 순위 기반 앙상블 기법을 사용하여 서로 다른 모델을 통합한다. 이하 컴퓨터 장치가 순위 기반 비모수 앙상블 기법을 사용한다고 전제한다.

[0059] 전통적인 순위 기반 통합 기법은 순위 합계 기반(rank-sum-based) 접근, 평균 순위 기반(average-rank-based) 접근 또는 보르다 카운트 기법(Borda count election)을 사용한다. 이하 전술한 3 가지 직접적 상관도 유추 기법의 결과물을 통합하기 위하여 역 순위 기반 방법을 사용한다.

[0060] 직접 연관성 유추 기법을 통해 생성된 각 행렬은 전체 miRNA-mRNA 조합에 대한 직접적 상관관계를 표현하며 이는 miRNA-mRNA 연관 네트워크 상에서의 에지의 가중치 값을 의미한다. 직접적 상관도 유추 기법을 통해 얻은 결과물을 비모수적으로 통합하기 위해, 각 결과 행렬들의 연관 수치값을 내림차순 순위로 변환하고 값이 작은 순위에 높은 가중치를 부여하기 위해 역 순위 값을 연관성 가중치로 사용한다. 한편 통합 네트워크에서의 특정 miRNA와 mRNA 사이의 가중치는 해당 miRNA-mRNA 쌍의 모든 직접 연관성 유추 기법의 결과물에서의 역순위의 곱으로 계산한다. 즉 특정한 miRNA와 mRNA 사이의 에지는 3가지 직접 연관성 유추 결과 중 동일한 에지에 대한 역순위를 곱한 값으로 결정한다. 예컨대, 3개의 유추 결과가 모두 동일한 에지를 포함한다면, 3개의 역 순위를 모두 곱한 값이 해당 에지의 가중치가 된다. 또는 3개의 유추 결과 중 두 개의 결과(모델)가 동일한 에지를 포함한다면, 2개의 역 순위를 곱한 값이 해당 에지의 가중치가 된다.

[0061] G가 통합된 네트워크의 연관성 집합, r_{ij} 가 노드 i와 노드 j 사이의 연관도 가중치 순위라고 하자. 역 순위 기반 전략을 사용하는 통합된 그래프의 연관성 가중치 r'_{ij} 는 아래의 수학적 식 3과 같이 나타낼 수 있다. 아래 수학적 식 3은 log 연산으로 곱셈 연산을 덧셈으로 변환하여 연산 복잡도를 줄였다.

수학적 식 3

$$r'_{ij} = \frac{1}{\log\left(\prod_{m \in G} (r_{ij}^m + 1)\right)}$$

[0063]

[0065] 한편 하나의 직접 연관성 유추 기법에서 생성한 m개의 결과물을 통합할 때도 역 순위 접근 기법을 사용할 수 있다.

[0067] 전술한 설명에서 컴퓨터 장치(200)가 miRNA-mRNA 네트워크를 구축한다고 하였다. 이하 miRNA-mRNA 네트워크를 구축하는 컴퓨터 장치(200)에 대하여 설명한다. 도 4는 miRNA-mRNA 네트워크 구축 장치(200)에 대한 예이다. 도 4에서 설명하는 구축 장치가 전술한 컴퓨터 장치(200)에 해당한다.

[0068] 도 16(A)는 네트워크 단에 위치한 객체로부터 유전자 발현 데이터를 전달받아 miRNA-mRNA 네트워크를 구축하는 장치(200)에 대한 예이다. miRNA-mRNA 네트워크를 구축 장치(200)는 유전자 발현 DB(210) 및 컴퓨터 장치(220)를 포함한다.

[0069] 유전자 발현 DB(expression DB, 210)는 특정 생물체의 유전자 발현과 관련된 데이터를 저장한다. 전술한 바와 같이 유전자 발현 데이터는 마이크로어레이 등과 같은 기법을 이용하여 마련된다. 유전자 발현 데이터는 miRNA 발현 데이터 및 mRNA 발현 데이터를 포함한다.

[0070] 컴퓨터 장치(220)는 유전자 발현 DB(210)에 저장된 발현 데이터를 수신한다. 전술한 바와 같이 컴퓨터 장치(220)는 유전자 발현 데이터에 통계적 기법을 적용하여 차별 발현 유전자를 선택하고, 선택한 유전자 발현 데이터를 통합한다. 컴퓨터 장치(220)는 부트스트래핑을 통해 통합된 데이터의 샘플을 증가시킨다. 컴퓨터 장치(220)는 샘플 데이터를 기준으로 직접적 연관도 분석을 위한 3가지 기법을 적용하여 상관도 모델을 생성한다. 컴퓨터 장치(220)는 3가지 기법의 결과물을 통합하여 최종적인 상관도 모델을 생성한다. 마지막으로 컴퓨터 장치(220)는 최종적인 상관도 모델을 기반으로 miRNA-mRNA 네트워크를 구축한다. miRNA-mRNA 네트워크를 구축을 위한 각 과정은 전술한 바와 같다.

[0072] 도 4(B) miRNA-mRNA 네트워크를 구축하는 컴퓨터 장치(300)에 대한 예이다. miRNA-mRNA 네트워크를 구축하는 컴퓨터 장치(300)는 입력 장치(310), 연산장치(320), 저장장치(330) 및 출력장치(340)를 포함한다.

[0073] 입력장치(310)는 유전자 발현 데이터를 입력받는다. 입력장치(310)는 키보드, 마우스, 터치패드와 같은 물리적인 인터페이스 장치일 수 있다. 또는 입력장치(310)는 외부 저장매체(USB 등)로부터 저장된 유전자 발현 데이터를 전달받은 장치일 수도 있다. 또는 입력장치(310)는 외부 네트워크로부터 유전자 발현 데이터를 수신하는 통신 모듈일 수도 있다.

[0074] 저장장치(330)는 miRNA-mRNA 네트워크를 구축을 위한 프로그램을 저장한다. 또한 저장장치(330)는 구축된 miRNA-mRNA 네트워크를 저장할 수도 있다.

[0075] 연산장치(320)는 입력되는 유전자 발현 데이터와 저장장치(330)에 저장된 프로그램을 이용하여 miRNA-mRNA 네트워크를 구축하는 연산을 수행한다. miRNA-mRNA 네트워크를 구축 과정은 전술한 컴퓨터 장치(220)가 수행한 과정과 동일하다. 나아가 연산장치(320)는 저장장치(330)에 저장된 분석 프로그램을 이용하여 구축된 miRNA-mRNA 네트워크를 분석하는 과정을 수행할 수도 있다.

- [0076] 출력장치(340)는 miRNA-mRNA 네트워크 내지 분석 결과를 출력하는 장치이다. 출력장치(340)는 영상을 출력하는 디스플레이 장치, 텍스트를 출력하는 프린터 등일 수 있다. 나아가 출력장치(340)는 생성한 miRNA-mRNA 네트워크 내지 분석 데이터를 다른 장치에 전달하는 통신 모듈일 수도 있다.
- [0078] 이하 전술한 miRNA-mRNA 연관도를 분석하는 기법에 대한 효과를 설명한다. 이하 전술한 miRNA-mRNA 연관도를 분석하는 기법을 제안 기법이라고 명명한다. 제안 기법에 대한 효과 검증을 위하여 널리 알려진 매칭되는 miRNA와 mRAN의 발현 데이터를 사용한다. 먼저 제안 기법에서 miRNA와 mRAN의 직접적 연관도를 추정하는 기법과 함께 사용된 부트스트래핑과 앙상블 기법의 효과를 살펴보고, 제안 기법과 현재 가장 좋은 성능을 가졌다고 알려진 앙상블 기법의 효과를 비교한다.
- [0079] 종래 연구에서 사용되었던 3개의 매칭되는 miRNA와 mRAN의 발현 데이터 세트를 사용하였다(Le TD, Zhang J, Liu L, Li J. Ensemble methods for miRNA target prediction from expression data. PLoS One. 2015;10(6):e0131627 및 Le TD, Zhang J, Liu L, Liu H, Li J. miRLAB: an R based dry lab for exploring miRNA-mRNA regulatory relationships. PLoS One. 2015;10(12):e0145386. 참조). 3개의 발현 데이터 세트는 EMT(Epithelial to Mesenchymal Transition) 데이터, MCC(Multi-Class Cancer) 데이터 및 BR(Breast Cancer) 데이터이다. EMT 데이터는 11개 샘플의 상피(epithelia) 및 36개 샘플의 중간엽(mesenchymal)에 대한 매칭된 miRNA-mRNA 발현 데이터이다. MCC 데이터는 60개 샘플에서 8개 기관의 정상 조직과 종양 조직에서 획득한 매칭된 miRNA-mRNA 발현 데이터이다. BR 데이터는 50개 샘플의 유방암 조직에서 획득한 매칭된 miRNA-mRNA 발현 데이터이다.
- [0080] 상용 프로그램을 사용하여 차별 발현 데이터를 선별할 수 있다. 차별 발현을 구분하기 위한 기준으로 p 값을 사용한다. p값을 0.05 미만으로 제한하여 EMT 데이터에서 35개의 miRNA와 1,154개의 mRNA를 선별하였고, MCC 데이터에서 108개의 miRNA와 1,860개의 mRNA를 선별하였다. BR 데이터에서 p값을 0.2 미만으로 제한하여 92개의 miRNA 선별하였고, p값을 0.0001 미만으로 제한하여 1,500개의 mRNA를 선별하였다. 선별한 차별 발현 데이터는 통합되었고, 이후 과정을 위해 일정하게 전처리(표준화)하였다.
- [0081] miRNA와 mRNA의 상관도는 현재까지 실험적 결과로 유추되고 있다. 따라서 종래 연구된 결과에 따른 상관도를 진정한 상관도라고 가정하고, 제안 기술의 효과를 검증하였다. 종래 실험적 결과(진정한 상관도 판단 기준)는 4개의 데이터 베이스인 Tarbase v.6.0 (Vergoulis T, Vlachos IS, Alexiou P, Georgakilas G, Maragkakis M, Reczko M, et al. TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. Nucleic Acids Res. 2012;40(D1):D222-9.), miRecords v2013(Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA target interactions. Nucleic Acids Res. 2009;37 suppl 1:D105-10.), miRWalk v2.0(Dweep H, Sticht C, Pandey P, Gretz N. miRWalk database: prediction of possible miRNA binding sites by walking the genes of three genomes. J Biomed Inform. 2011;44(5):839-47.) 및 miRTarBase v.4.5(Hsu SD, Tseng YT, Shrestha S, Lin YL, Khaleel A, Chou CH, et al. miRTarBase update 2014: an information resource for experimentally validated miRNA target interactions. Nucleic Acids Res. 2014;42(D1):D78-85.)를 통합하여 검증 데이터로 사용하였다. 제안 기법의 정확도는 연관도를 추정하는 모든 miRNA와 mRNA 쌍에서 각 miRNA에 대해 순위가 높은 k개의 mRNA 쌍들을 모은 집합과 검증 데이터와의 교집합 원소의 수를 기준으로 판단하였다.
- [0083] 도 5는 제안 기법의 효과를 검증한 실험 결과의 예이다. 도 5(A)는 miRNA와 mRNA의 직접적 연관도를 추정을 위하여 제안 기법에서 제안한 다양한 실시예의 효과를 나타낸 예이다. 3개의 유전자 발현 데이터(EMT, MCC 및 BR)에 대하여 효과를 검증하였다. 각 miRNA에 대하여 상위 100개의 miRNA-mRNA 쌍을 기준으로 효과를 분석하였다.
- [0084] 샘플 데이터는 발현 데이터를 전체를 그대로 이용한 경우(Whole)와 부트스트래핑을 통해 샘플 데이터를 증폭한 경우(Bootstrap)로 구분된다. 단일 기법(Single method)은 직접적 연관도 유추를 위한 하나의 기법을 사용한 경우이다. 3가지 기법은 각각 부분상관계수 추정법(Corpcor(C)), SPACE(S) 및 네트워크 디컨볼루션(ND(N))이다. 앙상블 기법(Ensemble Method)은 C와 S를 통합한 경우(C&S), C와 N을 통합한 경우(C&N), S와 N을 통합한 경우(S&N), 3가지를 모두 통합한 경우(C&S&N)으로 구분한다.
- [0085] 먼저 전체 발현 데이터(Whole)에 단일 기법을 적용한 경우를 살펴보면 유전자 발현 데이터 종류에 따라 서로 다

른 기법이 효과적인 것을 알 수 있다. 예컨대, EMT는 S가, MCC는 N이, BR은 C가 가장 효과적이었다. 즉, 입력 데이터의 종류(실험의 조건 등에도 영향을 받음)에 따라 효과적인 기법이 상이한 것이다. 나아가 단일 기법을 사용한 경우 부트스트래핑을 사용하여도 부트스트래핑을 사용하지 않은 경우에 비하여 항상 효과가 뛰어난 것은 아니었다. 예컨대, MCC의 N 기법의 경우 부트스트래핑을 사용한 경우 오히려 성능이 떨어졌다.

[0086] 앙상블 기법을 적용한 경우를 살펴보면, 단일 기법을 사용한 경우에 비하여 정확도가 높아진 것을 알 수 있다. 전반적으로 3가지 기법을 모두 통합한 경우(C&S&N)에서 단일 기법에 비하여 성능이 향상되었고, 부트스트래핑을 활용한 경우가 사용하지 않은 경우에 비하여 성능이 향상되었다. 특정 샘플 데이터에 대하여 단일 기법을 사용하거나, 어느 두 개의 기법을 통합한 경우가 더 성능이 좋은 것으로 나타났으나, 입력 데이터의 종류에 따라 어떤 기법이 유효한지 사전에 알 수 없으므로, 3가지 기법을 통합한 분석 기법이 전반적으로 유의하다고 할 수 있다.

[0087] 도 5(B)는 제안 기법과 종래 분석 기법의 효과를 비교한 예이다. 종래 앙상블 기법은 Pearson, IDA 및 Lasso(Le TD, Zhang J, Liu L, Li J. Ensemble methods for miRNA target prediction from expression data. PLoS One. 2015;10(6):e0131627. 참조)를 사용하였다. Pearson(P), IDA(I) 및 Lasso(L)를 통합(ensemble)한 결과와 제안 기법을 비교하였다. 제안 기법에서 직접적 연관도 추정 기법을 어느 하나를 사용한 경우(Single)은 종래 기법보다 성능이 떨어졌다. 그러나 단일기법이라고 부트스트래핑을 적용한 경우(Bootstrap), 3가지 기법의 결과를 통합한 경우(Ensemble), 부트스트래핑과 함께 3가지 기법의 결과를 통합한 경우(Bootstrap & Ensemble)은 종래 기법에 비하여 성능이 뛰어나다. 특히 부트스트래핑과 앙상블을 같이 적용한 경우의 효과가 가장 높은 것을 알 수 있다.

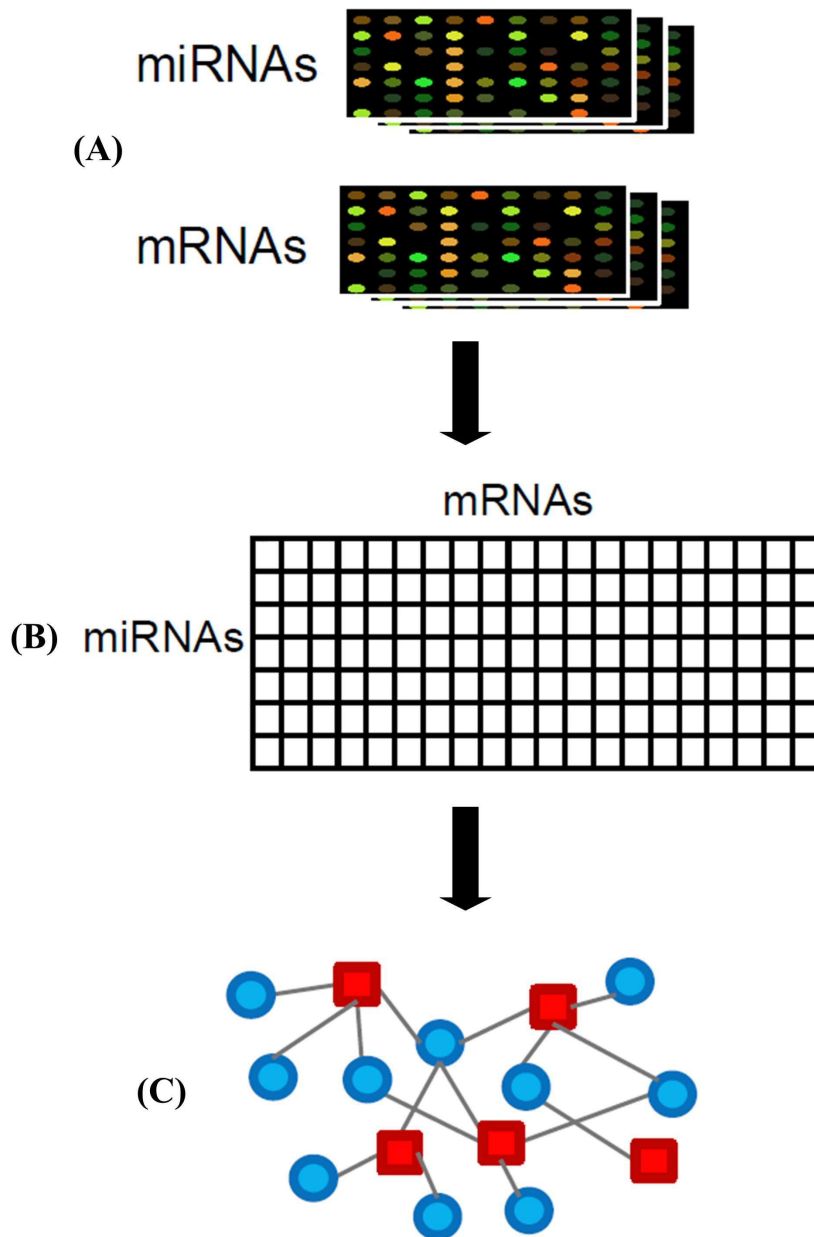
[0089] 본 실시례 및 본 명세서에 첨부된 도면은 전술한 기술에 포함되는 기술적 사상의 일부를 명확하게 나타내고 있는 것에 불과하며, 전술한 기술의 명세서 및 도면에 포함된 기술적 사상의 범위 내에서 당업자가 용이하게 유추할 수 있는 변형 예와 구체적인 실시례는 모두 전술한 기술의 권리범위에 포함되는 것이 자명하다고 할 것이다.

부호의 설명

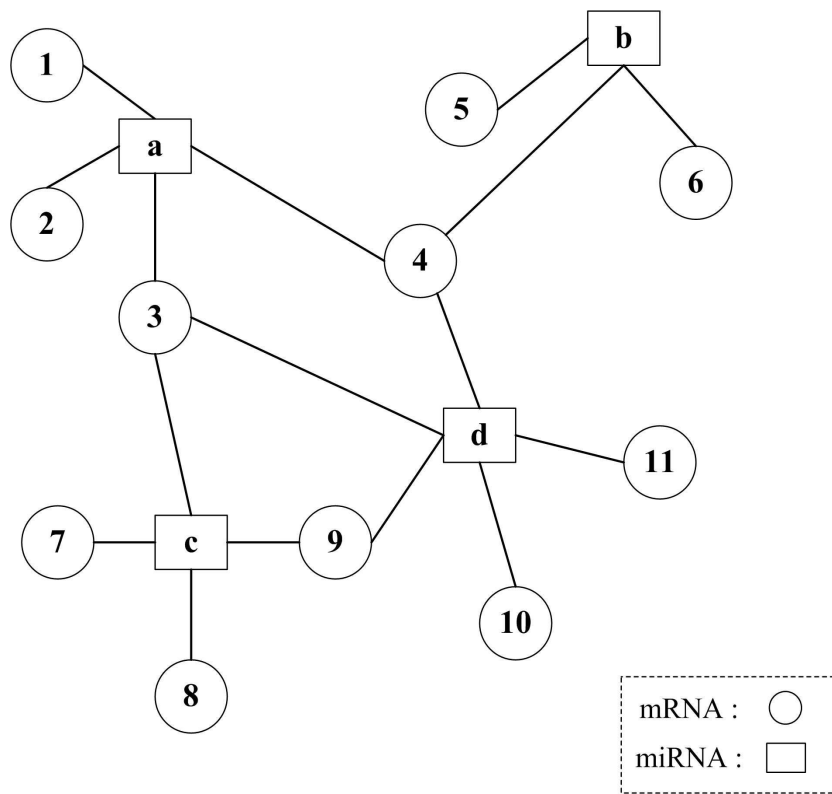
[0090] 200 : miRNA-mRNA 네트워크 구축 장치
210 : 유전자 발현 DB
220 : 컴퓨터 장치
300 : miRNA-mRNA 네트워크 구축 장치
310 : 입력장치
320 : 연산장치
330 : 저장장치
340 : 출력 장치

도면

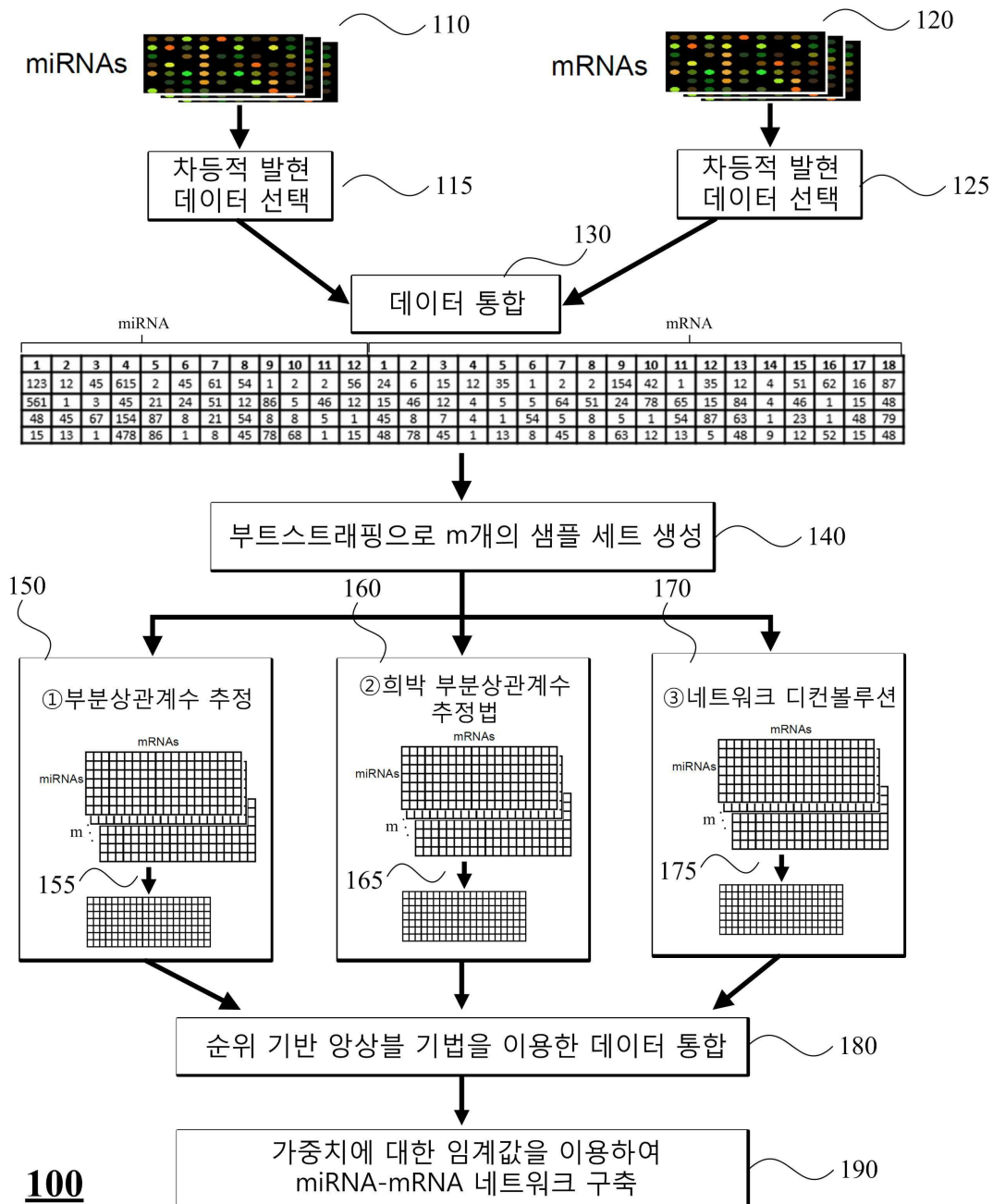
도면1



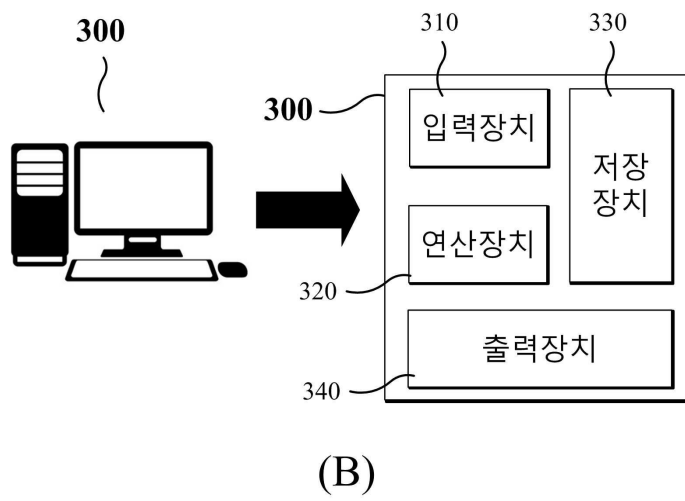
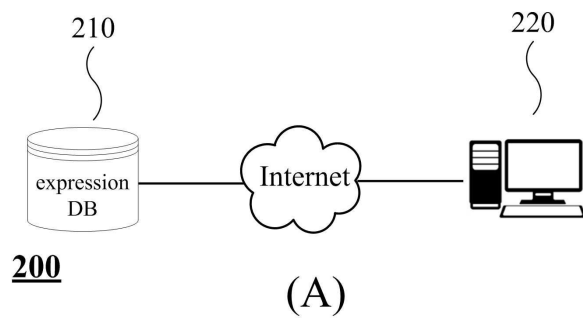
도면2



도면3



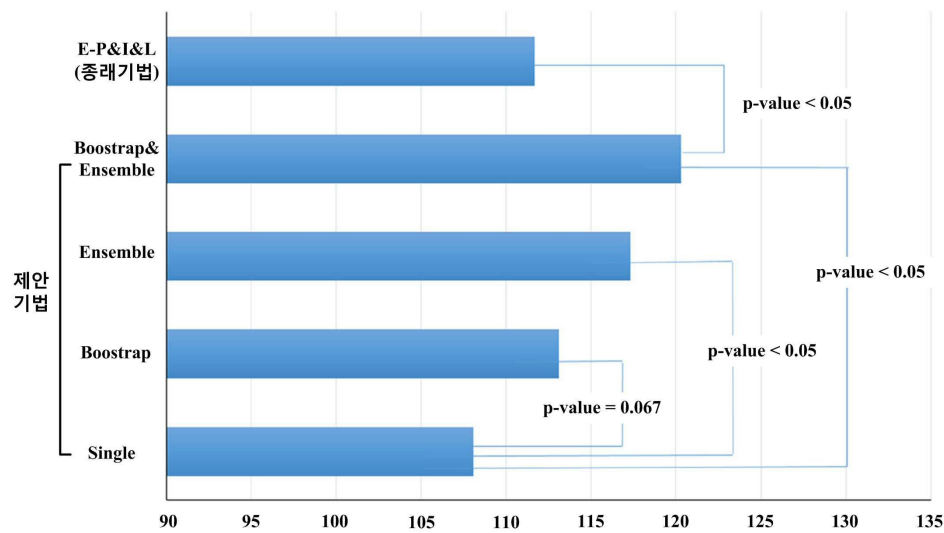
도면4



도면5

		Single Method			Ensemble Method			
		Corpcor	Space	ND	C&S	C&N	S&N	C&S&N
EMT	Whole	35	45	24	45	34	35	41
	Bootstrap	32	38	25	40	24	37	40
MCC	Whole	200	183	210	204	206	201	209
	Bootstrap	211	204	207	201	217	220	216
BR	Whole	98	83	95	90	94	97	102
	Bootstrap	107	95	99	99	102	100	105

(A)



(B)