



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0132977
(43) 공개일자 2021년11월05일

(51) 국제특허분류(Int. Cl.)
G16B 25/10 (2019.01) C12Q 1/6869 (2018.01)
G16B 30/00 (2019.01)
(52) CPC특허분류
G16B 25/10 (2019.02)
C12Q 1/6869 (2018.05)
(21) 출원번호 10-2020-0051547
(22) 출원일자 2020년04월28일
심사청구일자 2020년04월28일

(71) 출원인
인천대학교 산학협력단
인천광역시 연수구 아카데미로 119 (송도동)
(72) 발명자
정현두
인천광역시 연수구 컨벤시아대로252번길 30 송도
더샵퍼스트파크F15BL 1505동 2003호
(74) 대리인
차준용

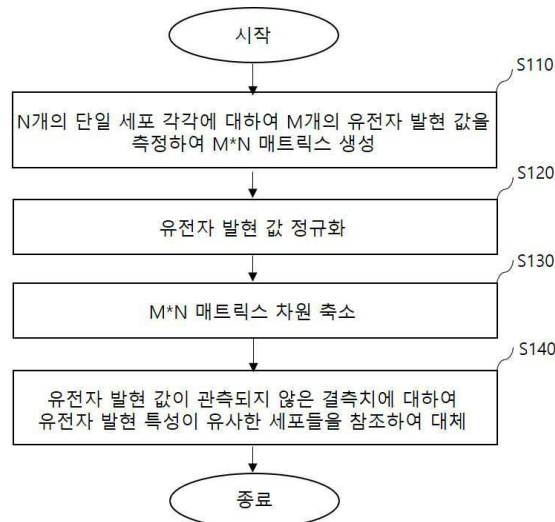
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 단일 세포 염기서열 분석 방법 및 시스템

(57) 요약

본 발명은 단일 세포 염기서열 분석 방법 및 시스템에 관한 것으로, 본 발명에 따른 단일 세포 염기서열 분석 방법은 N개(N은 2이상의 자연수)의 세포 각각에 대하여 M(M은 자연수)개의 유전체 발현 값을 측정하여 M*N 매트릭스를 생성하는 단계; 측정된 유전체 발현 값을 정규화하는 단계; 상기 매트릭스의 차원을 축소하는 단계; 상기 매트릭스 중 유전체 발현 값이 관측되지 않은 결측치를 해당 세포와 유전체 발현 특성이 유사한 세포들의 유전체 발현 값을 참조하여 획득한 값으로 대체하는 단계;를 포함한다. 본 발명에 따르면 단일 세포 염기서열 분석 과정에서 발생하는 결측치를 효과적으로 대체함으로써 분석의 신뢰성을 개선할 수 있다.

대표도 - 도1



(52) CPC특허분류

G16B 30/00 (2019.02)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711091613
과제번호	2019R1G1A1004803
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)(R&D)
연구과제명	동적 생물학적 네트워크 비교 분석을 위한 네트워크 정렬 및 군집화 알고리즘 연구
기 여 율	1/1
과제수행기관명	인천대학교
연구기간	2019.03.01 ~ 2020.02.29

명세서

청구범위

청구항 1

각 단계가 프로그램으로 구현되어 컴퓨터에서 수행되는 단일 세포 염기서열 분석 방법에 있어서,

N개(N은 2이상의 자연수)의 세포 각각에 대하여 M(M은 자연수)개의 유전체 발현 값을 측정하여 M*N 매트릭스를 생성하는 단계;

측정된 유전체 발현 값을 정규화하는 단계;

상기 매트릭스의 차원을 축소하는 단계;

상기 매트릭스 중 유전체 발현 값이 관측되지 않은 결측치를 해당 세포와 유전체 발현 특성이 유사한 세포들의 유전체 발현 값을 참조하여 획득한 값으로 대체하는 단계;를 포함하는 단일 세포 염기서열 분석 방법.

청구항 2

제 1 항에 있어서,

상기 결측치가 대체된 매트릭스에 대하여 상기 차원 축소 단계 및 상기 결측치 대체 단계를 반복 수행하는 단일 세포 염기서열 분석 방법.

청구항 3

제 2 항에 있어서, 상기 결측치 대체 단계는

상기 축소된 차원의 매트릭스 데이터를 이용하여 세포간 유사성을 확인하고, 이에 기초하여 세포간 상관성 네트워크를 구축하는 단계;

각 세포에 대하여 상관성 있는 세포가 존재하는 지역적 서브네트워크를 추정하는 단계; 및

상기 지역적 서브네트워크내 세포들의 유전체 발현 값에 기초하여 해당 세포의 유전체 발현 값을 도출하는 단계;를 포함하는 단일 세포 염기서열 분석 방법.

청구항 4

제 3 항에 있어서, 상기 상관성 네트워크 구축 단계는

피어슨(Pearson) 상관성 및 유클리드 거리(Euclidean distance) 중 적어도 하나를 이용하여 상기 세포간 유사성을 확인하는 단일 세포 염기서열 분석 방법.

청구항 5

제 4 항에 있어서, 상기 상관성 네트워크 구축 단계는

상기 유사성이 높은 세포를 적응적으로 선택하여 이웃하는 노드로 구성하는 단일 세포 염기서열 분석 방법.

청구항 6

제 5 항에 있어서, 상기 상관성 네트워크 구축 단계는

해당 세포와의 상기 피어슨(Pearson) 상관성 값이 제1 문턱 값보다 큰 세포를 이웃 노드로 선택하는 단일 세포 염기서열 분석 방법.

청구항 7

제 6 항에 있어서, 상기 제1 문턱 값은

설정된 최소값 및 다른 모든 세포들과의 상기 상관성 값의 퍼센타일(Percentile) 중 최소값을 이용하여 적응적

으로 선택되는 단일 세포 염기서열 분석 방법.

청구항 8

제 5 항에 있어서, 상기 상관성 네트워크 구축 단계는

해당 세포와의 상기 유클리드 거리(Euclidean distance) 값이 제2 문턱 값보다 큰 세포를 이웃 노드로 선택하는 단일 세포 염기서열 분석 방법.

청구항 9

제 8 항에 있어서, 상기 제2 문턱 값은

설정된 최소값 및 다른 모든 세포들과의 상기 상관성 값의 퍼센타일(Percentile) 중 최소값을 이용하여 적응적으로 선택되는 단일 세포 염기서열 분석 방법.

청구항 10

제 3 항에 있어서, 상기 지역적 서브네트워크 추정 단계는

전체 세포수의 20퍼센트 및 이웃 노드수의 1.25배 중 작은 값을 상기 지역적 서브네트워크에 포함되는 세포수 $K(K$ 는 양의 정수)로 적응적으로 선택하는 단일 세포 염기서열 분석 방법.

청구항 11

제 3 항에 있어서, 상기 유전체 발현 값 도출 단계는

해당 세포의 유전체 발현 값 및 상기 지역적 서브네트워크내에 존재하는 세포들의 유전체 발현 값의 평균값을 이용하여 추정하되, 각각 서로 다른 가중치를 적용하여 합산하는 단일 세포 염기서열 분석 방법.

청구항 12

N 개(N 은 2이상의 자연수)의 세포 각각에 대하여 M (M 은 자연수)개의 유전체 발현 값을 측정하여 $M \times N$ 매트릭스를 생성하는 측정부;

측정된 유전체 발현 값을 정규화하는 정규화부;

상기 매트릭스의 차원을 축소하는 차원 축소부; 및

상기 매트릭스 중 유전체 발현 값이 관측되지 않은 결측치를 해당 세포와 유전체 발현 특성이 유사한 세포들의 유전체 발현 값을 참조하여 획득한 값으로 대체하는 결측치 대체부;를 포함하는 단일 세포 염기서열 분석 시스템.

청구항 13

제 12 항에 있어서, 상기 결측치 대체부는

상기 축소된 차원의 매트릭스 데이터를 이용하여 세포간 유사성을 확인하고, 이에 기초하여 세포간 상관성 네트워크를 구축하는 상관성 네트워크 구축부;

각 세포에 대하여 상관성 있는 세포가 존재하는 지역적 서브네트워크를 추정하는 서브네트워크 추정부; 및

상기 지역적 서브네트워크내 세포들의 유전체 발현 값에 기초하여 해당 세포의 유전체 발현 값을 도출하는 결측치 결정부;를 포함하는 단일 세포 염기서열 분석 시스템.

청구항 14

제 13 항에 있어서, 상기 상관성 네트워크 구축부는

피어슨(Pearson) 상관성 및 유클리드 거리(Euclidean distance) 중 적어도 하나를 이용하여 상기 세포간 유사성을 확인하는 단일 세포 염기서열 분석 시스템.

청구항 15

제 14 항에 있어서, 상기 상관성 네트워크 구축부는

상기 유사성이 높은 세포를 적응적으로 선택하여 이웃하는 노드로 구성하는 단일 세포 염기서열 분석 시스템.

청구항 16

제 15 항에 있어서, 상기 상관성 네트워크 구축부는

해당 세포와의 상기 피어슨(Pearson) 상관성 값이 제1 문턱 값보다 큰 세포를 이웃 노드로 선택하는 단일 세포 염기서열 분석 시스템.

청구항 17

제 16 항에 있어서, 상기 제1 문턱 값은

설정된 최소값 및 다른 모든 세포들과의 상기 상관성 값의 퍼센타일(Percentile) 중 최소값을 이용하여 적응적으로 선택되는 단일 세포 염기서열 분석 시스템.

청구항 18

제 13 항에 있어서, 상기 상관성 네트워크 구축부는

해당 세포와의 상기 유클리드 거리(Euclidean distance) 값이 제2 문턱 값보다 큰 세포를 이웃 노드로 선택하는 단일 세포 염기서열 분석 시스템.

청구항 19

제 14 항에 있어서, 상기 제2 문턱 값은

설정된 최소값 및 다른 모든 세포들과의 상기 상관성 값의 퍼센타일(Percentile) 중 최소값을 이용하여 적응적으로 선택되는 단일 세포 염기서열 분석 시스템.

청구항 20

제 13 항에 있어서, 상기 결측치 대체부는

상기 결측치 대체 횟수가 미리 설정된 기준치 또는 미리 정해진 방식에 따라 적응적으로 선택된 값보다 작으면 상기 대체된 결측치를 포함하는 데이터를 상기 차원 축소부의 입력으로 되돌리는 판단부;를 더 포함하는 단일 세포 염기서열 분석 시스템.

발명의 설명

기술 분야

[0001] 본 발명은 단일 세포 염기서열 분석 방법 및 시스템에 관한 것으로, 보다 상세하게는 단일 세포 염기서열 분석 과정에서 실제 발현된 유전체 값이 관측되지 않아 발생하는 결측치를 효과적으로 대체 또는 보정하는 방법 및 시스템에 관한 것이다.

배경 기술

[0002] 기존 차세대 염기서열 분석 기법은 조직을 구성하는 세포들의 평균 유전체 발현 값만 측정 가능하다.

[0003] 그러나 최근 단일 세포 염기서열 기법은 세포 각각의 유전체 발현 값을 측정한다.

[0004] 적은 양의 시료를 증폭하는 과정에서 실제 발현된 유전체 값이 관측되지 못하는 결손 사건으로 인해 분석 정확도가 저감되는 문제가 발생한다.

[0005] 결측치는 실제 발현된 유전체 값이 계측기술의 한계로 인하여 일부 세포에서 관측되지 않는 것을 말하는데, 만약 이 결측치를 무시하고 관측자료만을 이용하여 단일 세포 염기서열 데이터를 분석하게 되면 많은경우에 통계적 편향이 생긴다. 따라서 통계 데이터를 전처리하는 과정에서 이러한 결측치를 적절히 보정 및 대체할 필요가 있다.

발명의 내용

해결하려는 과제

[0006] 단일 세포 RNA 염기서열 분석 과정에서 발생하는 결측치를 효과적으로 대체하는 방법을 제안한다.

과제의 해결 수단

- [0007] 본 발명의 일실시예에 따른 단일 세포 염기서열 분석 방법은 N개(N은 2이상의 자연수)의 세포 각각에 대하여 M(M은 자연수)개의 유전체 발현 값을 측정하여 $M \times N$ 매트릭스를 생성하는 단계; 측정된 유전체 발현 값을 정규화하는 단계; 상기 매트릭스의 차원을 축소하는 단계; 상기 매트릭스 중 유전체 발현 값이 관측되지 않은 결측치를 해당 세포와 유전체 발현 특성이 유사한 세포들의 유전체 발현 값을 참조하여 획득한 값으로 대체하는 단계;를 포함한다.
- [0008] 상기 단일 세포 염기서열 분석 방법은 상기 결측치가 대체된 매트릭스에 대하여 상기 축소 단계 및 상기 결측치 대체 단계를 반복 수행한다.
- [0009] 상기 결측치 대체 단계는 상기 축소된 차원의 매트릭스 데이터를 이용하여 세포간 유사성을 확인하고, 이에 기초하여 세포간 상관성 네트워크를 구축하는 단계; 각 세포에 대하여 상관성 있는 세포가 존재하는 지역적 서브네트워크를 추정하는 단계; 및 상기 지역적 서브네트워크내 세포들의 유전체 발현 값에 기초하여 해당 세포의 유전체 발현 값을 도출하는 단계;를 포함한다.
- [0010] 상기 상관성 네트워크 구축 단계는 피어슨 상관성(Pearson Correlation) 및 유클리드 거리(Euclidean distance) 중 적어도 하나를 이용하여 상기 세포간 유사성을 확인한다.
- [0011] 상기 상관성 네트워크 구축 단계는 상기 유사성이 높은 세포를 적응적으로 선택하여 이웃하는 노드로 구성한다.
- [0012] 상기 상관성 네트워크 구축 단계는 해당 세포와의 상기 피어슨 상관성이 제1 문턱 값보다 큰 세포를 이웃 노드로 선택한다.
- [0013] 상기 제1 문턱 값은 설정된 최소값 및 다른 모든 세포들과의 상기 상관성 값의 퍼센타일(Percentile) 중 최소값을 이용하여 적응적으로 선택된다.
- [0014] 상기 상관성 네트워크 구축 단계는 해당 세포와의 상기 유클리드 유사성이 제2 문턱 값보다 큰 세포를 이웃 노드로 선택한다.
- [0015] 상기 제2 문턱 값은 설정된 최소값 및 다른 모든 세포들과의 상기 상관성 값의 퍼센타일(Percentile) 중 최소값을 이용하여 적응적으로 선택된다.
- [0016] 상기 지역적 서브네트워크 추정 단계는 전체 세포수의 20퍼센트 및 이웃 노드수의 1.25배 중 작은 값을 상기 지역적 서브네트워크에 포함되는 세포수 K(K는 양의 정수)로 적응적으로 선택한다.
- [0017] 상기 유전체 발현값 도출 단계는 해당 세포의 유전체 발현 값 및 상기 지역적 서브네트워크내에 존재하는 세포들의 유전체 발현 값의 평균값을 이용하여 추정하되, 각각 서로 다른 가중치를 적용하여 합산한다.
- [0018] 본 발명의 일실시예에 따른 단일 세포 염기서열 분석 시스템은 N개(N은 2이상의 자연수)의 세포 각각에 대하여 M(M은 자연수)개의 유전체 발현 값을 측정하여 $M \times N$ 매트릭스를 생성하는 측정부; 측정된 유전체 발현 값을 정규화하는 정규화부; 상기 매트릭스의 차원을 축소하는 차원 축소부; 및 상기 매트릭스 중 유전체 발현 값이 관측되지 않은 결측치를 해당 세포와 유전체 발현 특성이 유사한 세포들의 유전체 발현 값을 참조하여 획득한 값으로 대체하는 결측치 대체부;를 포함한다.
- [0019] 상기 결측치 대체부는 상기 축소된 차원의 매트릭스 데이터를 이용하여 세포간 유사성을 확인하고, 이에 기초하여 세포간 상관성 네트워크를 구축하는 상관성 네트워크 구축부; 각 세포에 대하여 상관성 있는 세포가 존재하는 지역적 서브네트워크를 추정하는 서브네트워크 추정부; 및 상기 지역적 서브네트워크내 세포들의 유전체 발현 값에 기초하여 해당 세포의 유전체 발현 값을 도출하는 결측치 결정부;를 포함한다.
- [0020] 상기 상관성 네트워크 구축부는 피어슨상관성 및 유클리드 거리 중 적어도 하나를 이용하여 상기 세포간 유사성을 확인한다.
- [0021] 상기 상관성 네트워크 구축부는 상기 유사성이 높은 세포를 적응적으로 선택하여 이웃하는 노드로 구성한다.

- [0022] 상기 상관성 네트워크 구축부는 해당 세포와의 상기 피어슨 상관성이 제1 문턱 값보다 큰 세포를 이웃 노드로 선택한다.
- [0023] 상기 제1 문턱 값은 설정된 최소값 및 다른 모든 세포들과의 상기 상관성 값의 퍼센타일(Percentile) 중 최소값을 이용하여 적응적으로 선택된다.
- [0024] 상기 상관성 네트워크 구축부는 해당 세포와의 상기 유클리드 유사도가 제2 문턱 값보다 큰 세포를 이웃 노드로 선택한다.
- [0025] 상기 제2 문턱 값은 설정된 최소값 및 다른 모든 세포들과의 상기 상관성 값의 퍼센타일(Percentile) 중 최소값을 이용하여 적응적으로 선택된다.
- [0026] 상기 결측치 대체부는 상기 결측치 대체 횟수가 미리 설정된 기준치 또는 미리 정해진 방식에 따라 적응적으로 선택된 값보다 작으면 상기 대체된 결측치를 포함하는 데이터를 상기 차원 축소부의 입력으로 되돌리는 판단부;를 더 포함한다.

발명의 효과

- [0028] 본 발명에 따르면 단일 세포 RNA 염기서열 분석 과정에서 발생하는 유전체 발현 값의 결측치를 효과적으로 대체할 수 있다.
- [0029] 또한, 결측치가 조래하는 통계적 오류를 제거함으로써 염기서열 분석 결과에 대한 신뢰성을 향상시킬 수 있다.

도면의 간단한 설명

- [0031] 도 1은 본 발명의 일실시예에 따른 단일 세포 염기서열 분석 방법을 나타낸다.
- 도 2는 도 1에 도시된 대체하는 단계(S140)의 일실시예를 나타낸다.
- 도 3은 본 발명의 일실시예에 따른 단일 세포 염기서열 분석 시스템을 나타낸다.

발명을 실시하기 위한 구체적인 내용

- [0032] 본 명세서에서 사용된 용어는 실시예들을 설명하기 위한 것이며 본 발명을 제한하고자 하는 것은 아니다. 본 명세서에서, 단수형은 문구에서 특별히 언급되지 않는 한 복수형도 포함한다. 명세서에서 사용되는 "포함한다(comprises)" 및/또는 "포함하는(comprising)"은 언급된 구성요소 외에 하나 이상의 다른 구성요소의 존재 또는 추가를 배제하지 않는다. 명세서 전체에 걸쳐 동일한 도면 부호는 동일한 구성 요소를 지칭하며, "및/또는"은 언급된 구성요소들의 각각 및 하나 이상의 모든 조합을 포함한다. 비록 "제1", "제2" 등이 다양한 구성요소들을 서술하기 위해서 사용되나, 이들 구성요소들은 이들 용어에 의해 제한되지 않음은 물론이다. 이들 용어들은 단지 하나의 구성요소를 다른 구성요소와 구별하기 위하여 사용하는 것이다. 따라서, 이하에서 언급되는 제1 구성요소는 본 발명의 기술적 사상내에서 제2 구성요소일 수도 있음은 물론이다. 또한, 명세서에 기재된 "...부", "모듈" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어 또는 소프트웨어로 구현되거나 하드웨어와 소프트웨어의 결합으로 구현될 수 있다.
- [0033] 이하에서는 도면을 참조하여 본 발명에 대하여 보다 상세히 설명하기로 한다. 이하에 소개되는 실시예들은 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 본 발명의 사상이 충분히 전달될 수 있도록 하기 위해 예로서 제공되는 것이다.
- [0035] 도 1은 본 발명의 일실시예에 따른 단일 세포 염기서열 분석 방법을 나타낸다. 본 발명의 일실시예에 따른 단일 세포 염기서열 분석 방법은 각 단계가 프로그램으로 구현되어 컴퓨터에서 수행될 수 있다.
- [0036] 본 발명의 일실시예에 따른 단일 세포 염기서열 분석 방법은 N개(N은 2이상의 자연수)의 세포 각각에 대하여 M(M은 자연수)개의 유전체 발현 값을 측정하여 M*N 매트릭스를 생성하는 단계(S110), 측정된 유전체 발현 값을 정규화하는 단계(S120), 상기 매트릭스의 차원을 축소하는 단계(S130), 및 유전체 발현 값이 관측되지 않은 결측치는 해당 세포와 유전체 발현 특성이 유사한 세포들의 유전체 발현 값을 참조하여 대체하는 단계(S140)를 포

함한다.

[0037] 상기 생성하는 단계(S110)에서는 단일 세포 염기서열 분석 방법을 통해 각 세포의 유전체 발현 값을 측정하고, 측정 값을 원소로 하여 매트릭스 X를 생성한다. 상기 매트릭스 X는 M개의 행 및 N개의 열로 구성될 수 있다. 이 경우, M은 유전체 갯수, N은 세포의 갯수를 나타낸다.

[0038] 상기 정규화하는 단계(S120)에서는 측정된 유전체 발현 값을 정규화한다. 일실시예로 CPM(Counts per Million) 및 로그 변환을 이용하여 정규화할 수 있다. 구체적으로 M*N 매트릭스를 CPM 변환하여 얻은 매트릭스 X에 대하여 하기 수식 1 과 같이 로그 연산을 수행하여 정규화된 매트릭스 X_n을 얻을 수 있다.

수식 1

$$X_n = \log_{10}(1 + X)$$

[0039]

[0041] 상기 축소하는 단계(S130)에서는 매트릭스 X_n의 차원을 축소한다. 이를 위해 세포들간에 변동성이 높은 유전체를 선택할 수 있다. PCA(Principal Component Analysis)를 이용하여 각 세포에 대한 축소된 차원의 매트릭스를 얻을 수 있다. 일실시예로 세포간 분산 값이 높은 유전체와 PCA를 이용하여 매트릭스 X_n의 차원을 축소할 수 있다. PCA 대신에 t-SNE, UMAP 기법 등 다양한 차원 저감 기법을 적용하여 차원을 축소할 수도 있다. 매트릭스 X_n의 차원을 축소함으로써 계산 복잡도를 줄일 수 있는 장점이 있다.

[0042] 상기 대체하는 단계(S140)에서는 실제 발현된 유전체 값이 관측되지 않은 특정 세포의 결측치에 대하여 해당 세포와 유전체 발현 특성이 유사한 세포들의 유전체 발현 값을 참조하여 대체한다.

[0043] 한편, 본 발명의 일실시예에 따른 단일 세포 염기서열 분석 방법은 상기 결측치가 대체된 매트릭스를 대상으로 상기 축소하는 단계(S130) 및 상기 대체하는 단계(S140)를 여러번 반복 수행하면서 결측치를 더욱 정밀하게 예측할 수 있다. 이 경우, 반복 횟수는 미리 설정될 수 있다. 일실시예로 반복 횟수는 5로 설정될 수 있다. 다른 실시예로 반복 횟수는 적응적으로 선택될 수도 있다. 예를 들어 n+1번 반복하여 얻은 결측치 데이터와 n번 반복하여 얻은 결측치 데이터의 차이가 기준치 이하인 경우, 더 이상 반복하지 않고 종료할 수 있다. 구체적으로 수식 2을 만족하는 경우, 결측치를 대체하기 위한 반복 작업은 종료할 수 있다.

수식 2

$$\frac{|X_{n+1} - X_n|^2}{|M| \cdot |N|} \leq 0.05$$

[0044]

[0046] 도 2는 도 1에 도시된 대체하는 단계(S140)의 일실시예를 나타낸다.

[0047] 상기 대체하는 단계(S140)는 상기 차원이 축소된 매트릭스 데이터를 이용하여 세포간 유사성을 확인하고, 이에 기초하여 세포간 상관성 네트워크를 구축하는 단계(S142), 각 세포에 대하여 상관성 있는 세포가 존재하는 지역적 서브네트워크를 추정하는 단계(S144), 및 상기 지역적 서브네트워크내 세포들의 유전체 발현 값에 기초하여 해당 세포의 유전체 발현 값을 도출하는 단계(S146)를 포함한다.

[0048] 상기 구축하는 단계(S142)에서는 세포간 유사성에 관한 정보를 제공하는 세포간 상관성 네트워크를 구축한다.

세포간 상관성 네트워크는 그래프 $G = (V, E)$ 로 표현된다. 노드 V_i ($v_i \in V$)는 i번째 세포를 나타내고, 세포 상관성은 에지(edge) $e_{ij} \in E$ 으로 나타낼 수 있다. 만약 V_i 와 V_j 가 서로 유사하다면 에지 가중치 e_{ij} 는 양의 값을 가질 수 있다.

[0049] 세포간 상관성 네트워크는 세포간 유사성을 이용하여 구축할 수 있다. 상기 세포간 유사성은 피어슨(Pearson)

상관성 및 유클리드 거리(Euclidean distance) 중 적어도 하나를 이용하여 확인할 수 있다. 유사성이 높은 세포를 적응적으로 선택하여 이웃하는 노드로 구성하게 된다. 이웃하는 노드는 유전체 발현 정보가 유사한 패턴을 가지는 세포로 분류될 수 있는 세포들의 집합을 의미한다. 피어슨 상관성을 이용한 네트워크 G_C 와 유클리드 거리에 기초한 상관성 네트워크 G_E 를 구축하고, 이를 합하여 포괄적인 세포간 상관성 네트워크를 구축할 수 있다.

[0050] 피어슨(Pearson) 상관성을 이용하여 네트워크를 구축하는 경우, 이웃 노드의 선정 기준은 다음과 같다. 즉, 특정 세포와의 피어슨(Pearson) 상관성이 제1 문턱 값보다 큰 세포들을 상기 특정 세포의 이웃 노드로 선택한다. 구체적으로 수학적 식 3을 이용하여 이웃 노드를 선택할 수 있다. $C_{i,j}$ 는 i 번째 세포 V_i 와 j 번째 세포 V_j 간의 피어슨 상관성을 나타내고, $C_t(V_i)$ 는 i 번째 세포 V_i 의 이웃 노드를 선택하기 위한 문턱값을 나타낸다. 즉, i 번째 세포 V_i 와의 피어슨 상관성 $C_{i,j}$ 이 문턱값 $C_t(V_i)$ 보다 큰 j 번째 세포 V_j 를 i 번째 세포 V_i 의 이웃 노드 $N_c(V_i)$ 로 선정한다.

수학적 식 3

$$N_c(v_i) = \{v_j | C_{i,j} \geq c_t(v_i)\}$$

[0051]

[0052] 상기 문턱 값 $C_t(V_i)$ 은 수학적 식 4와 같이 사용자가 지정한 최소값 N_{th} 와 다른 모든 세포들과의 상관성 값의 퍼센타일(Percentile) 중 최소값을 이용하여 적응적으로 선택될 수 있다. 파라미터 N_{th} 및 Q_{th} 는 데이터 세트에 따라 사용자에게 의해 적절하게 선택될 수 있다. 예를 들어, N_{th} 는 0.85, Q_{th} 는 0.9로 설정할 수 있다.

수학적 식 4

$$c_t(v_i) = \min \{N_{th}, (Q_{th} \cdot \text{percentile of } C_{i,j}, \forall_j)\}$$

[0053]

[0054] 이후, i 번째 세포 V_i 및 그 이웃 노드에 속하는 j 번째 세포 V_j 를 선 또는 변으로 연결한다. 피어슨 상관성에 기초한 세포간 상관성 네트워크의 인접 매트릭스(Adjacency Matrix)는 수학적 식 5와 같이 생성할 수 있다.

수학적 식 5

$$A_{i,j}^c = \begin{cases} 1, & v_j \in N_c(v_i), \forall_{vi} \\ 0, & otherwise \end{cases}$$

[0055]

[0056] 한편, j 번째 세포 V_j 는 i 번째 세포 V_i 의 이웃 노드에 속할지라도, 그 반대의 경우 즉 i 번째 세포 V_i 가 당연히 j 번째 세포 V_j 의 이웃 노드에 속하는 것은 아니다. 따라서 상기 방법으로 구축된 세포간 상관성 네트워크는 방향성 없는 네트워크(Undirected Network)이므로 수학적 식 6을 이용하여 방향성 네트워크(Directed Network)로 변환할 수 있다. 수학적 식 6은 인접 매트릭스 A_c 와 그 전치 매트릭스의 선형 결합을 나타낸다.

수학적 식 6

$$P_c = 0.5 \times (A_c + A_c^T)$$

[0057]

[0058] 위 변환 과정에서 i 번째 세포 V_i 와 j 번째 세포 V_j 가 서로 이웃 노드로 선택된 경우에는 가중치를 더해줄 수 있다.

[0060] 한편, 유클리드 거리(Euclidean distance)를 이용하여 상관성 네트워크를 구축하는 경우, 이웃 노드의 선정 기준은 다음과 같다. 세포간의 유사도가 증가할수록 유클리드 거리는 작아지므로, 가우시안(Gaussian) 커널을 이용하여 수학적 식 7과 같이 유클리드 유사도(Euclidean Similarity)를 구할 수 있다.

수학적 식 7

$$E = \exp(-\tilde{E})$$

[0062] $\tilde{E}$ 는 수학적 식 8의 스케일링된 유클리드 거리 매트릭스 $\hat{M}$ 의 요소별 제곱(element-wise square)을 수행한 결과이다. M 은 각 세포간의 유클리드 거리에 관한

수학적 식 8

$$\hat{M} = 2 \cdot M / \max(M)$$

[0064] 유클리드 유사도를 이용하는 경우에도, 피어슨 상관도를 이용하여 인접 매트릭스 P_C 를 구하는 방법과 유사하게 상기 수학적 식 3 내지 6을 적용하여 인접 매트릭스 P_E 를 구할 수 있다. 요약하면 특정 세포와의 유클리드 유사도가 문턱 값보다 큰 세포들을 해당 세포의 이웃 노드로 선택한다. 이웃 노드와의 선 또는 변을 연결하여 인접 매트릭스를 도출하고, 방향성 있는 매트릭스가 되도록 변환을 수행한다.

[0065] 한편, 수학적 식 9에 나타난 바와 같이 인접 매트릭스 P_C 및 P_E 를 구한 후, 이들을 각각 동일 가중치를 적용하여 선형 결합할 수 있다.

수학적 식 9

$$P = 0.5 \times (P_C + P_E)$$

[0067] 이렇게 피어슨 상관성 정보와 유클리드 거리 정보를 이용하여 추정된 세포간 상관성 네트워크를 통합함으로써 서로 다른 유사성 추정 기법을 이용하더라도 일관되게 이웃 노드로 선정된 세포들에는 가중치를 더하는 단계를 더 포함할 수 있다.

[0069] 상기 추정하는 단계(S144)에서는 지역적 서브네트워크(Local subnetwork)를 추정한다. 구체적으로 랜덤 워크(Random walk) 기법을 활용하여 i 번째 세포 V_i 와 상관성 있는 세포가 존재하는 서브네트워크를 추정할 수 있다. 이 때 랜덤 워커는 세포간 상관성 네트워크에서 미리 정해진 J (J 는 자연수) 스텝동안 랜덤으로 움직일 수 있다. 상기 지역적 서브네트워크가 K (K 는 양의 정수)개의 세포를 포함하는데, K 는 수학적 식 10에 나타난 바와 같이 전체 세포수 N 에 미리 정해진 K 의 최소값 파라미터 $K_{\min}$ 을 곱한 결과 및 i 번째 세포 V_i 의 이웃 노드 개수 $N(V_i)$ 에 미리 정해진 K 의 최대값 파라미터 $K_{\max}$ 을 곱한 결과 중 최소값을 이용하여 적응적으로 선택될 수 있다. 일실시예로 J 는 3, 파라미터 $K_{\min}$ 은 0.2, 파라미터 $K_{\max}$ 은 1.25로 설정될 수 있다. 즉, 상기 지역적 서브네트워크에 포함되는

세포수 K (K 는 양의 정수)를 결정하기 위하여 전체 세포수의 20퍼센트 및 이웃 노드수의 1.25배 중의 작은 값을 적응적으로 선택할 수 있다.

수학식 10

$$K = \min\{(K_{min} \cdot N), (K_{max} \cdot |N(v_i)|)\}$$

상기 도출하는 단계(S146)에서는 결측치를 도출한다. 구체적으로 지역적 서브네트워크내 유전체 발현 값의 평균과 분산 값에 기초하여 i 번째 세포 V_i 의 M 개 유전체 발현값을 대체할 수 있다.

일실시예로 해당 세포의 유전체 발현 값 및 상기 지역적 서브네트워크내에 존재하는 세포들의 유전체 발현 값의 평균값을 이용하여 추정하되, 각각 서로 다른 가중치를 적용하여 합산한다. 만약에 i 번째 세포 V_i 의 M 번째 유전체 발현 값이 신뢰할만 하다면 즉, 지역적 서브네트워크의 평균 값에 가깝다면 비교적 큰 가중치를 부여하고, 유전체 발현 값을 조정하기 위한 이웃 세포들의 정보는 덜 이용할 수 있다. 반대로 i 번째 세포 V_i 의 M 번째 유전체 발현 값이 지역적 서브네트워크의 평균값과 차이가 크다면, 이웃 세포들의 정보에 더 큰 가중치를 둬으로써 결측치를 대체할 수 있다. 구체적으로 수학식 11에서 보는 바와 같이 i 번째 세포의 유전체 발현의 결측치를 도출하기 위하여 추정된 지역적 서브네트워크 안에 존재하는 세포의 유전체 발현값의 평균값 벡터 μ_i 를 이용한다.

수학식 11

$$X_i^{n+1} = p \cdot X_i^n + (1 - p) \cdot \mu_i$$

가중치 P 는 수학식 12의 시그모이드(sigmoid) 함수를 이용하여 계산될 수 있다.

수학식 12

$$P = \frac{1}{1 + \exp(-\alpha \cdot (x - \mu_i))}$$

α 가 매우 작은 값으로 설정되면 시그모이드 함수가 스텝 함수에 가까워진다. 따라서 극단적인 경우를 피하기 위해 마진을 주어서 스케일링 계수 α 를 1로 설정할 수 있다. 스케일링 계수 α 는 시그모이드 함수의 스케일링 계수로서 수학식 13을 이용하여 결정된다. 스케일 계수 α 는 추정된 지역 서브네트워크 안에 존재하는 세포의 유전체 발현 값의 분산값 벡터 σ_i 을 이용하여 계산한다.

수학식 13

$$\alpha = \min\{2 \cdot \exp(0.05 \cdot \sigma_i), 1\}$$

도 3은 본 발명의 일실시예에 따른 단일 세포 염기서열 분석 시스템을 나타낸다.

도 3을 참조하면, 본 발명의 일실시예에 따른 단일 세포 염기서열 분석 시스템은 측정부(310), 변환부(320) 및 결측치 대체부(330)를 포함한다.

- [0082] 상기 측정부(310)는 단일 세포 염기서열 분석 방법을 통해 각 세포의 유전체 발현 값을 측정하고, 측정 값을 원소로 하여 매트릭스 X 를 생성한다. 상기 매트릭스 X 는 M 개의 행 및 N 개의 열로 구성될 수 있다. 이 경우, M 은 유전체 갯수, N 은 세포의 갯수를 나타낸다.
- [0083] 상기 변환부(320)는 매트릭스 X 를 정규화하고 차원을 축소한다. 이를 위해, 변환부(320)는 정규화부(322) 및 차원축소부(324)를 포함할 수 있다.
- [0084] 상기 정규화부(322)는 측정된 유전체 발현 값을 정규화한다. 일실시예로 CPM(Counts per Million) 및 로그 변환을 이용하여 정규화할 수 있다. 구체적으로 $M \times N$ 매트릭스를 CPM 변환하여 얻은 매트릭스 X 에 대하여 상기 수학식 1 과 같이 로그 연산을 수행하여 정규화된 매트릭스 X_n 을 얻을 수 있다.
- [0085] 상기 차원축소부(324)는 매트릭스 X_n 의 차원을 축소한다. 이를 위해 세포들간에 변동성이 높은 유전체를 선택할 수 있다. PCA(Principal Component Analysis)를 이용하여 각 세포에 대한 축소된 차원의 매트릭스를 얻을 수 있다. 일실시예로 세포간 분산 값이 높은 유전체와 PCA를 이용하여 매트릭스 X_n 의 차원을 축소할 수 있다. PCA 대신에 t-SNE, UMAP 기법을 적용하여 차원을 축소할 수도 있다. 매트릭스 X_n 의 차원을 축소함으로써 계산 복잡도를 줄일 수 있는 장점이 있다.
- [0086] 상기 결측치 대체부(330)는 유전체 발현 값이 관측되지 않은 특정 세포의 결측치에 대하여 해당 세포와 유전체 발현 특성이 유사한 세포들의 유전체 발현 값을 참조하여 대체한다. 이를 위해, 결측치 대체부(330)는 상관성 네트워크 구축부(332), 서브네트워크 추정부(334) 및 결측치 결정부(336)를 포함할 수 있다.
- [0087] 상기 상관성 네트워크 구축부(332)는 상기 차원이 축소된 매트릭스 데이터를 이용하여 세포간 유사성을 확인하고, 이에 기초하여 세포간 상관성 네트워크를 구축한다. 구체적으로 상기 상관성 네트워크 구축부(332)는 세포간 유사성에 관한 정보를 제공하는 세포간 상관성 네트워크를 구축한다. 세포간 상관성 네트워크는 그래프 $G = (V, E)$ 로 표현된다. 노드 $V_i (v_i \in V)$ 는 i 번째 세포를 나타내고, 세포 상관성은 에지(edge) $e_{i,j} \in E$ 로 나타낼 수 있다. 만약 V_i 와 V_j 가 서로 유사하다면 에지 가중치 $e_{i,j}$ 는 양의 값을 가질 수 있다. 상기 상관성 네트워크 구축부(332)는 세포간 유사성을 이용하여 세포간 상관성 네트워크를 구축할 수 있다. 상기 세포간 유사성은 피어슨(Pearson) 상관성 및 유클리드 거리(Euclidean distance) 중 적어도 하나를 이용하여 확인할 수 있다. 유사성이 높은 세포를 적응적으로 선택하여 이웃하는 노드로 구성하게 된다. 이웃하는 노드는 유사 형질 패턴을 갖는 같은 타입의 세포로 분류될 수 있는 세포들의 집합을 의미한다. 피어슨 상관성을 이용한 네트워크 G_C 와 유클리드 거리에 기초한 상관성 네트워크 G_E 를 구축하고, 이를 합하여 포괄적인 세포간 상관성 네트워크를 구축할 수 있다.
- [0088] 상기 상관성 네트워크 구축부(332)는 피어슨(Pearson) 상관성을 이용하여 네트워크를 구축할 수 있다. 이 경우, 이웃 노드의 선정 기준은 다음과 같다. 즉, 특정 세포와의 피어슨(Pearson) 상관성이 제1 문턱 값보다 큰 세포들을 상기 특정 세포의 이웃 노드로 선택한다. 구체적으로 수학식 3을 이용하여 이웃 노드를 선택할 수 있다. $C_{i,j}$ 는 i 번째 세포 V_i 와 j 번째 세포 V_j 간의 피어슨 상관성을 나타내고, $C_t(V_i)$ 는 이웃 노드를 선택하기 위한 문턱 값을 나타낸다. 즉, i 번째 세포 V_i 와의 피어슨 상관성 $C_{i,j}$ 이 문턱값 $C_t(V_i)$ 보다 큰 j 번째 세포 V_j 를 i 번째 세포 V_i 의 이웃 노드 $N_c(V_i)$ 로 선정한다.
- [0089] 상기 문턱 값 $C_t(V_i)$ 은 수학식 4와 같이 사용자가 지정한 최소값 N_{th} 와 다른 모든 세포들과의 상관성 값의 퍼센타일(Percentile) 중 최소값을 이용하여 적응적으로 선택될 수 있다. 파라미터 N_{th} 및 Q_{th} 는 데이터 세트에 따라 사용자에게 의해 적절하게 선택될 수 있다. 예를 들어, N_{th} 는 0.85, Q_{th} 는 0.9로 설정할 수 있다.
- [0090] 한편, j 번째 세포 V_j 는 i 번째 세포 V_i 의 이웃 노드에 속할지라도, 그 반대의 경우 즉 i 번째 세포 V_i 가 당연히 j 번째 세포 V_j 의 이웃 노드에 속하는 것은 아니다. 따라서 상기 방법으로 구축된 세포간 상관성 네트워크는 방향성 없는 네트워크(Undirected Network)이므로 상기 상관성 네트워크 구축부(332)는 수학식 6을 이용하여 방향성 네트워크(Directed Network)로 변환할 수 있다. 상기 변환 과정에서 i 번째 세포 V_i 와 j 번째 세포 V_j 가 서로 이웃 노드로 선택된 경우에는 가중치를 더해줄 수 있다.

- [0091] 상기 상관성 네트워크 구축부(332)는 유클리드 거리(Euclidean distance)를 이용하여 상관성 네트워크를 구축할 수 있다. 이 경우, 이웃 노드의 선정 기준은 다음과 같다. 세포간의 유사도가 증가할수록 유클리드 거리는 작아지므로, 가우시안(Gaussian) 커널을 이용하여 수학적 7과 같이 유클리드 유사도(Euclidean Similarity)를 구할 수 있다.
- [0092] 상기 상관성 네트워크 구축부(332)는 유클리드 유사도를 이용하는 경우에도, 피어슨 상관도를 이용하여 인접 매트릭스 P_c 를 구하는 방법과 유사하게 상기 수학적 3 내지 6을 적용하여 인접 매트릭스 P_E 를 구할 수 있다. 요약하면 특정 세포와의 유클리드 유사도가 문턱 값보다 큰 세포들을 해당 세포의 이웃 노드로 선택한다. 이웃 노드와의 선 또는 변을 연결하여 인접 매트릭스를 도출하고, 방향성 있는 매트릭스가 되도록 변환을 수행한다.
- [0093] 상기 상관성 네트워크 구축부(332)는 수학적 9에 나타난 바와 같이 인접 매트릭스 P_c 및 P_E 를 구한 후, 이들을 각각 동일 가중치를 적용하여 선형 결합할 수 있다. 이렇게 피어슨 상관성 정보와 유클리드 거리 정보를 이용하여 추정된 세포간 상관성 네트워크를 통합함으로써 서로 다른 유사성 추정 기법을 이용하더라도 일관되게 이웃 노드로 선정된 세포들에는 가중치를 더 줄 수 있다.
- [0095] 상기 서브네트워크 추정부(334)는 각 세포에 대하여 상관성 있는 세포가 존재하는 지역적 서브네트워크(Local subnetwork)를 추정한다. 구체적으로 상기 서브네트워크 추정부(334)는 랜덤 워크(Random walk) 기법을 활용하여 i 번째 세포 V_i 와 상관성 있는 세포가 존재하는 서브네트워크를 추정할 수 있다. 이 때 랜덤 워커는 세포간 상관성 네트워크에서 미리 정해진 J (J 는 자연수) 스텝동안 랜덤으로 움직일 수 있다. 상기 지역적 서브네트워크가 K (K 는 양의 정수)개의 세포를 포함하는데, K 는 수학적 10에 나타난 바와 같이 전체 세포수 N 에 미리 정해진 K 의 최소값 파라미터 K_{min} 을 곱한 결과 및 i 번째 세포 V_i 의 이웃 노드 개수 $N(v_i)$ 에 미리 정해진 K 의 최대값 파라미터 K_{max} 을 곱한 결과 중 최소값을 이용하여 적응적으로 선택될 수 있다. 일실시예로 J 는 3, 파라미터 K_{min} 은 0.2, 파라미터 K_{max} 은 1.25로 설정될 수 있다.
- [0097] 상기 결측치 결정부(336)는 상기 지역적 서브네트워크내 세포들의 유전체 발현 값에 기초하여 해당 세포의 유전체 발현 값을 도출한다. 구체적으로 상기 결측치 결정부(336)는 지역적 서브네트워크내 유전체 발현 값의 평균과 분산 값에 기초하여 i 번째 세포 V_i 의 M 개 유전체 발현값을 대체할 수 있다.
- [0098] 일실시예로 상기 결측치 결정부(336)는 해당 세포의 유전체 발현 값 및 상기 지역적 서브네트워크내에 존재하는 세포들의 유전체 발현 값의 평균값을 이용하여 추정하되, 각각 서로 다른 가중치를 적용하여 합산한다. 만약에 i 번째 세포 V_i 의 M 번째 유전체 발현 값이 신뢰할만 하다면 즉, 지역적 서브네트워크의 평균 값에 가깝다면 비교적 큰 가중치를 부여하고, 유전체 발현 값을 조정하기 위한 이웃 세포들의 정보는 덜 이용할 수 있다. 반대로 i 번째 세포 V_i 의 M 번째 유전체 발현 값이 지역적 서브네트워크의 평균값과 차이가 크다면, 이웃 세포들의 정보에 더 큰 가중치를 줌으로써 결측치를 대체할 수 있다. 구체적으로 상기 결측치 결정부(336)는 수학적 11에서 보는 바와 같이 i 번째 세포의 유전체 발현의 결측치를 도출하기 위하여 추정된 지역적 서브네트워크 안에 존재하는 세포의 유전체 발현값의 평균값 벡터 μ_i 을 이용한다.
- [0099] 상기 결측치 결정부(336)는 수학적 12의 시그모이드(sigmoid) 함수를 이용하여 가중치 P 를 계산될 수 있다. α 가 매우 작은 값으로 설정되면 시그모이드 함수가 스텝 함수에 가까워진다. 따라서 극단적인 경우를 피하기 위해 마진을 주어서 스케일링 계수 α 를 1로 설정할 수 있다. 상기 결측치 결정부(336)는 수학적 13을 이용하여 시그모이드 함수의 스케일링 계수 α 를 결정한다. 스케일 계수 α 는 추정된 지역 서브네트워크 안에 존재하는 세포의 유전체 발현 값의 분산값 벡터 σ_i 을 이용하여 계산된다.
- [0101] 한편, 본 발명의 일실시예에 따른 단일 세포 염기서열 분석 시스템은 상기 결측치 대체 프로세스를 여러번 반복 수행하면서 결측치를 더욱 정밀하게 예측할 수 있다. 이를 위해 결측치 대체부(330)는 판단부(338)를 더 포함할

수 있다. 판단부(338)는 반복 횟수를 카운트하고, 설정된 횟수 이내인 경우 획득한 데이터를 가지고 결측치 대체 프로세스를 반복하도록 한다. 반복 횟수는 미리 설정될 수 있다. 일실시예로 반복 횟수는 5로 설정될 수 있다. 다른 실시예로 반복 횟수는 적응적으로 선택될 수도 있다. 예를 들어 판단부(338)는 $n+1$ 번 반복하여 얻은 결측치 데이터와 n 번 반복하여 얻은 결측치 데이터의 차이가 기준치 이하인 경우, 더 이상 결측치 대체 프로세스를 반복하지 않고 종료할 수 있다. 구체적으로 수학적 2을 만족하는 경우, 판단부(338)는 결측치를 대체하기 위한 반복 작업을 종료할 수 있다.

[0103] 본 발명의 실시예와 관련하여 설명된 방법 또는 알고리즘의 단계들은 하드웨어로 직접 구현되거나, 하드웨어에 의해 실행되는 소프트웨어 모듈로 구현되거나, 또는 이들의 결합에 의해 구현될 수 있다. 소프트웨어 모듈은 RAM(Random Access Memory), ROM(Read Only Memory), EPROM(Erasable Programmable ROM), EEPROM(Electrically Erasable Programmable ROM), 플래시 메모리(Flash Memory), 하드 디스크, 착탈형 디스크, CD-ROM, 또는 본 발명이 속하는 기술 분야에서 잘 알려진 임의의 형태의 컴퓨터 판독가능한 기록매체에 상주할 수도 있다.

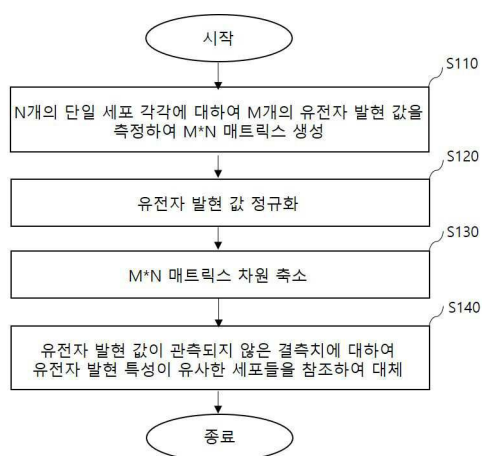
[0105] 이상, 첨부된 도면을 참조하여 본 발명의 실시예를 설명하였지만, 본 발명이 속하는 기술분야의 통상의 기술자는 본 발명이 그 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 실시될 수 있다는 것을 이해할 수 있을 것이다. 그러므로, 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며, 제한적이지 않은 것으로 이해해야만 한다.

부호의 설명

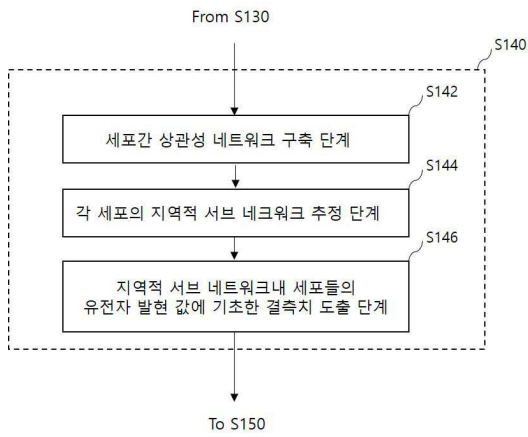
[0107] 310: 측정부 320: 변환부
322: 정규화부 324: 차원축소부
330: 결측치 대체부 332: 상관성 네트워크 구축부
334: 서브네트워크 추정부 336: 결측치 추정부
338: 판단부

도면

도면1



도면2



도면3

