(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2017-0108517
(43) 공개일자 2017년09월27일

(51) 국제특허분류(Int. Cl.)
G06F 19/18 (2011.01) *G06F 19/28* (2011.01)
(52) CPC특허분류
G06F 19/18 (2013.01)
G06F 19/28 (2013.01)
(21) 출원번호 10-2016-0032570
(22) 출원일자 2016년03월18일
심사청구일자 2016년03월18일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
석준희
서울특별시 강북구 삼양로 256, 106동 203호
태동현
전라북도 전주시 완산구 삼천천변1길 45, 202동 104호
장성용
경기도 의정부시 경의로146번길 33-12
(74) 대리인
심경식, 홍성욱

전체 청구항 수 : 총 15 항

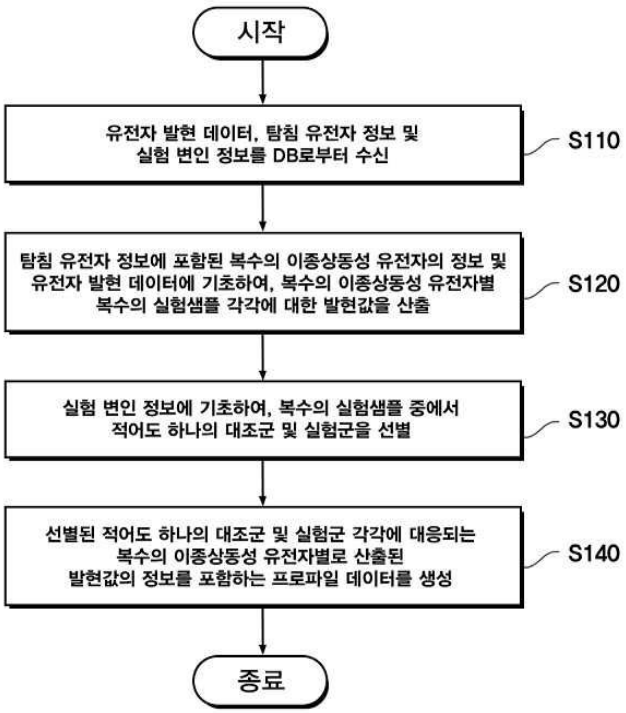
(54) 발명의 명칭 **이종상동유전자 발현 반응을 분석하기 위한 데이터의 재구성 방법 및 장치**

(57) 요 약

이종상동유전자 발현 반응을 분석하기 위한 데이터의 재구성 방법을 제공한다. 본 발명의 일 실시예에 따른 이종상동유전자 발현 반응을 분석하기 위한 데이터의 재구성 방법은 적어도 한 종류의 유전자를 탐침하는 프로브셋(probe set)과 유전자 발현 실험의 대상인 생물의 조직을 포함하는 실험샘플을 이용한 유전자 발현 반응을 분석

(뒷면에 계속)

대 표 도 - 도1



하기 위한 데이터의 재구성 방법에 있어서, 복수의 상기 프로브셋을 복수의 상기 실험샘플 각각과 반응시킨 발현 실험 결과인 발현값을 포함하는 유전자 발현 데이터, 상기 복수의 프로브셋별 탐침 유전자 정보 및 상기 복수의 실험샘플별 실험 변인 정보를 DB로부터 수신하는 단계; 상기 탐침 유전자 정보에 포함된 복수의 이종상동성 유전자(orthologous gene)의 정보 및 상기 유전자 발현 데이터에 기초하여, 상기 복수의 이종상동성 유전자별 상기 복수의 실험샘플 각각에 대한 상기 발현값을 산출하는 단계; 상기 실험 변인 정보에 기초하여, 상기 복수의 실험샘플 중에서 적어도 하나의 대조군 및 실험군을 선별하는 단계; 및 상기 선별된 적어도 하나의 대조군 및 실험군 각각에 대응되는 상기 복수의 이종상동성 유전자별로 산출된 상기 발현값의 정보를 포함하는 프로파일 데이터를 생성하는 단계를 포함한다.

이 발명을 지원한 국가연구개발사업

과제고유번호	1345233414
부처명	교육부
연구관리전문기관	한국연구재단
연구사업명	이공학개인기초연구지원
연구과제명	인간 면역 질환에 대한 모델 실험의 계산적 해석
기 여 율	1/1
주관기관	고려대학교
연구기간	2014.09.01 ~ 2017.04.30

명 세 서

청구범위

청구항 1

적어도 한 종류의 유전자를 탐침하는 프로브셋(probe set)과 유전자 발현 실험의 대상인 생물의 조직을 포함하는 실험샘플을 이용한 유전자 발현 반응을 분석하기 위한 데이터의 재구성 방법에 있어서,

복수의 상기 프로브셋을 복수의 상기 실험샘플 각각과 반응시킨 발현 실험 결과인 발현값을 포함하는 유전자 발현 데이터, 상기 복수의 프로브셋별 탐침 유전자 정보 및 상기 복수의 실험샘플별 실험 변인 정보를 DB로부터 수신하는 단계;

상기 탐침 유전자 정보에 포함된 복수의 이종상동성 유전자(orthologous gene)의 정보 및 상기 유전자 발현 데이터에 기초하여, 상기 복수의 이종상동성 유전자별 상기 복수의 실험샘플 각각에 대한 상기 발현값을 산출하는 단계;

상기 실험 변인 정보에 기초하여, 상기 복수의 실험샘플 중에서 적어도 하나의 대조군 및 실험군을 선별하는 단계; 및

상기 선별된 적어도 하나의 대조군 및 실험군 각각에 대응되는 상기 복수의 이종상동성 유전자별로 산출된 상기 발현값의 정보를 포함하는 프로파일 데이터를 생성하는 단계

를 포함하는 것을 특징으로 하는 이종상동유전자 발현 반응을 분석하기 위한 데이터의 재구성 방법.

청구항 2

제1항에 있어서,

상기 유전자 발현 데이터가 상기 복수의 프로브셋 및 상기 복수의 실험샘플 간의 상기 발현값을 나타내는 행렬의 형태일 때,

상기 행렬에 포함된 발현값의 크기에 기초하여, 선택적으로 상기 발현값을 상기 발현값의 로그값으로 대체하는 단계; 및

상기 행렬의 행 또는 열에 포함된 결손 원소의 개수에 기초하여, 선택적으로 상기 행 또는 열을 삭제하거나 상기 결손 원소를 보정값으로 대체하는 단계

를 더 포함하는 것을 특징으로 하는 이종상동유전자 발현 반응을 분석하기 위한 데이터의 재구성 방법.

청구항 3

제2항에 있어서,

상기 결손 원소를 보정값으로 대체하는 단계는

상기 행렬의 행 또는 열에 포함된 결손 원소의 개수 및 전체 원소의 개수 간의 비율이 소정의 임계비율 이상이면, 상기 행 또는 열을 삭제하고,

상기 비율이 상기 임계비율 미만이면, 상기 결손 원소를 인접한 다른 원소의 상기 발현값에 기초한 보정값으로 대체하는 것을 특징으로 하는 이종상동유전자 발현 반응을 분석하기 위한 데이터의 재구성 방법.

청구항 4

제1항에 있어서,

상기 복수의 실험샘플 각각에 대한 상기 발현값을 산출하는 단계는

상기 탐침 유전자 정보를 이용하여 상기 복수의 프로브셋과 상기 복수의 이종상동성 유전자 간의 다대다 매칭 관계를 결정하는 단계;

상기 다대다 매칭 관계를 다대일 매칭 관계로 변환하는 단계; 및

상기 다대일 매칭 관계를 일대일 매칭 관계로 변환하는 단계를 포함하고,

상기 다대일 매칭 관계 및 일대일 매칭 관계로 변환하는 단계는 상기 변환된 매칭 관계에 기초하여, 상기 복수의 이종상동성 유전자별 상기 복수의 실험샘플 각각에 대한 상기 발현값을 산출하는 것을 특징으로 하는 이종상동유전자 발현 반응을 분석하기 위한 데이터의 재구성 방법.

청구항 5

제4항에 있어서,

상기 다대다 매칭 관계를 다대일 매칭 관계로 변환하는 단계는

복수의 상기 이종상동성 유전자에 매칭된 하나의 상기 프로브셋에 대하여,

상기 복수의 이종상동성 유전자 중 선택된 하나를 매칭하거나, 상기 복수의 이종상동성 유전자 각각에 대응되는 복수의 가상프로브셋으로 분할하여 매칭하는 것을 특징으로 하는 이종상동유전자 발현 반응을 분석하기 위한 데이터의 재구성 방법.

청구항 6

제4항에 있어서,

상기 다대일 매칭 관계를 일대일 매칭 관계로 변환하는 단계는

복수의 상기 프로브셋에 매칭된 하나의 상기 이종상동성 유전자에 대하여,

상기 복수의 프로브셋별로 상기 복수의 실험샘플에 대하여 산출된 상기 발현값의 분산 또는 평균에 기초하여 선택된 하나의 상기 프로브셋을 매칭하거나, 상기 복수의 프로브셋을 하나의 가상프로브셋으로 통합하여 매칭하는 것을 특징으로 하는 이종상동유전자 발현 반응을 분석하기 위한 데이터의 재구성 방법.

청구항 7

제1항에 있어서,

상기 적어도 하나의 대조군 및 실험군을 선별하는 단계는

상기 실험 변인 정보를 이용하여 상기 복수의 실험샘플에 포함된 대조군당 실험군의 개수 및 시계열적인 실험의 수행 여부를 판단하는 단계;

상기 판단결과에 따라, 상기 복수의 실험샘플을 이용하는 실험구성을 단조건 실험구성, 다조건 실험구성, 단조건 시계열 실험구성 및 다조건 시계열 실험구성 중의 하나로 결정하는 단계; 및

상기 결정된 실험구성에 대응되는 상기 적어도 하나의 대조군 및 실험군을 선별하는 단계

를 포함하는 것을 특징으로 하는 이종상동유전자 발현 반응을 분석하기 위한 데이터의 재구성 방법.

청구항 8

제1항에 있어서,

상기 프로파일 데이터를 생성하는 단계는

상기 선별된 적어도 하나의 대조군 및 실험군 각각에 대응되는 상기 복수의 이종상동성 유전자별로 산출된 상기 발현값의 평균, 분산 및 평균의 차이값 중 적어도 하나를 더 포함하는 프로파일 데이터를 생성하는 것을 특징으로 하는 이종상동유전자 발현 반응을 분석하기 위한 데이터의 재구성 방법.

청구항 9

적어도 한 종류의 유전자를 탐침하는 프로브셋과 유전자 발현 실험의 대상인 생물의 조직을 포함하는 실험샘플을 이용한 유전자 발현 반응을 분석하기 위한 데이터 재구성 장치에 있어서,

복수의 상기 프로브셋을 복수의 상기 실험샘플 각각과 반응시킨 발현 실험 결과인 발현값을 포함하는 유전자 발

현 데이터, 상기 복수의 프로브셋별 탐침 유전자 정보 및 상기 복수의 실험샘플별 실험 변인 정보를 DB로부터 수신하는 수신부;

상기 탐침 유전자 정보에 포함된 복수의 이종상동성 유전자의 정보 및 상기 유전자 발현 데이터에 기초하여, 상기 복수의 이종상동성 유전자별 상기 복수의 실험샘플 각각에 대한 상기 발현값을 산출하는 매칭부;

상기 실험 변인 정보에 기초하여, 상기 복수의 실험샘플 중에서 적어도 하나의 대조군 및 실험군을 선별하는 선별부; 및

상기 선별된 적어도 하나의 대조군 및 실험군 각각에 대응되는 상기 복수의 이종상동성 유전자별로 산출된 상기 발현값의 정보를 포함하는 프로파일 데이터를 생성하는 생성부

를 포함하는 것을 특징으로 하는 이종상동유전자 발현 반응을 분석하기 위한 데이터 재구성 장치.

청구항 10

제9항에 있어서,

상기 유전자 발현 데이터가 상기 복수의 프로브셋 및 상기 복수의 실험샘플 간의 상기 발현값을 나타내는 행렬의 형태일 때,

상기 행렬에 포함된 발현값의 크기에 기초하여, 선택적으로 상기 발현값을 상기 발현값의 로그값으로 대체하고, 상기 행렬의 행 또는 열에 포함된 결손 원소의 개수에 기초하여, 선택적으로 상기 행 또는 열을 삭제하거나 상기 결손 원소를 보정값으로 대체하는 데이터변환부

를 더 포함하는 것을 특징으로 하는 이종상동유전자 발현 반응을 분석하기 위한 데이터 재구성 장치.

청구항 11

제10항에 있어서,

상기 데이터변환부는

상기 행렬의 행 또는 열에 포함된 결손 원소의 개수 및 전체 원소의 개수 간의 비율이 소정의 임계비율 이상이면, 상기 행 또는 열을 삭제하고,

상기 비율이 상기 임계비율 미만이면, 상기 결손 원소를 인접한 다른 원소의 상기 발현값에 기초한 보정값으로 대체하는 것을 특징으로 하는 이종상동유전자 발현 반응을 분석하기 위한 데이터 재구성 장치.

청구항 12

제9항에 있어서,

상기 매칭부는

상기 탐침 유전자 정보를 이용하여 상기 복수의 프로브셋과 상기 복수의 이종상동성 유전자 간의 다대다 매칭 관계를 결정하고,

상기 다대다 매칭 관계를 다대일 매칭 관계로 변환하고,

상기 다대일 매칭 관계를 일대일 매칭 관계로 변환하며,

상기 다대일 매칭 관계 및 일대일 매칭 관계로 변환할 때, 상기 변환된 매칭 관계에 기초하여, 상기 복수의 이종상동성 유전자별 상기 복수의 실험샘플 각각에 대한 상기 발현값을 산출하는 것을 특징으로 하는 이종상동유전자 발현 반응을 분석하기 위한 데이터 재구성 장치.

청구항 13

제12항에 있어서,

상기 다대다 매칭 관계를 다대일 매칭 관계로 변환할 때,

복수의 상기 이종상동성 유전자에 매칭된 하나의 상기 프로브셋에 대하여,

상기 복수의 이종상동성 유전자 중 선택된 하나를 매칭하거나, 상기 복수의 이종상동성 유전자 각각에 대응되는

복수의 가상프로브셋으로 분할하여 매칭하는 것을 특징으로 하는 이종상동유전자 발현 반응을 분석하기 위한 데이터 재구성 장치.

청구항 14

제12항에 있어서,

상기 다대일 매칭 관계를 일대일 매칭 관계로 변환할 때,

복수의 상기 프로브셋에 매칭된 하나의 상기 이종상동성 유전자에 대하여,

상기 복수의 프로브셋별로 상기 복수의 실험샘플에 대하여 산출된 상기 발현값의 분산 또는 평균에 기초하여 선택된 하나의 상기 프로브셋을 매칭하거나, 상기 복수의 프로브셋을 하나의 가상프로브셋으로 통합하여 매칭하는 것을 특징으로 하는 이종상동유전자 발현 반응을 분석하기 위한 데이터 재구성 장치.

청구항 15

제9항에 있어서,

상기 선별부는

상기 실험 변인 정보를 이용하여 상기 복수의 실험샘플에 포함된 대조군당 실험군의 개수 및 시계열적인 실험의 수행 여부를 판단하고,

상기 판단결과에 따라, 상기 복수의 실험샘플을 이용하는 실험구성을 단조건 실험구성, 다조건 실험구성, 단조건 시계열 실험구성 및 다조건 시계열 실험구성 중의 하나로 결정하고,

상기 결정된 실험구성에 대응되는 상기 적어도 하나의 대조군 및 실험군을 선별하는 것을 특징으로 하는 이종상동유전자 발현 반응을 분석하기 위한 데이터 재구성 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 이종상동유전자의 발현 반응을 분석하기 위하여 데이터를 재구성하는 방법 및 장치에 관한 것이며, 보다 상세하게는 서로 다른 종들 간의 유전자 발현(gene expression) 반응 데이터 분석을 위하여 기존의 복잡한 형태의 유전자 발현 반응 데이터를 단순한 형태의 데이터로 재구성하는 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 인간의 질병에 대한 원리 및 치료법을 찾기 위해 쥐와 같은 동물 모델을 이용한 실험이 많이 이용되고 있다. 이들 실험 중 질병이나 약물에 대한 유전자 발현 반응에 대한 실험은 질병의 발병이나 약물의 효과에 대한 근본적인 원리를 밝히는 주요한 연구 수단이다.

[0003] 지난 수 십 년간 인간과 동물 모델에 대한 다양한 유전자 발현 반응 실험 데이터가 축적되어 왔고, 이를 자유롭게 이용할 수 있게 되었다. 이러한 대규모 데이터의 사용 가능성은 인간과 동물 모델 사이의 차이점을 체계적으로 분석할 수 있는 기회를 제공한다.

[0004] 이러한 이종(異種)간의 데이터 분석을 위해, 과거에 축적된 대규모 유전자 발현 반응 데이터를 바탕으로, 이종간의 차이를 극복하여 분석할 수 있는 방법들이 제시되어 왔다. 예를 들어, 2015년도에 PLoS One지에 발표된 Evidence-based Translation은 면역관련 질환에 대하여 실험용 생쥐의 반응을 인간에 맞춰 분석할 수 있는 방법을 제시하였다.

[0005] 이와 같은 새로운 분석법은 공통적으로 대규모로 축적된 실험 데이터를 바탕으로 한다. 하지만, 현재 제공되는 데이터는 이러한 이종 간의 분석을 위해 최적화되어 있지 않아, 더 효율적인 방식의 데이터 재구성 방법이 필요하다.

[0006] 따라서, 이종상동유전자 발현 반응을 분석하기 위한 데이터의 재구성 방법 및 장치에 대한 필요성이 대두되고 있다.

[0007] 관련 선행기술로는 공개특허공보 제10-2008-0063156호(발명의 명칭: 표준 발현 유전자를 발굴하기 위한 유전자 발현 데이터 처리, 분석 방법, 공개일자: 2008년 7월 3일)가 있다.

발명의 내용

해결하려는 과제

[0008] 본 발명은 이종간의 유전자 발현 데이터 분석을 쉽게 하기 위하여 공공저장소에 축적된 복잡한 데이터를 사전에 단순한 프로파일 단위로 재구성하는 방법 및 장치를 제공하고자 한다.

과제의 해결 수단

[0009] 상기 목적을 달성하기 위해, 본 발명에서 제공하는 이종상동유전자 발현 반응을 분석하기 위한 데이터의 재구성 방법은 적어도 한 종류의 유전자를 탐침하는 프로브셋(probe set)과 유전자 발현 실험의 대상인 생물의 조직을 포함하는 실험샘플을 이용한 유전자 발현 반응을 분석하기 위한 데이터의 재구성 방법에 있어서, 복수의 상기 프로브셋을 복수의 상기 실험샘플 각각과 반응시킨 발현 실험 결과인 발현값을 포함하는 유전자 발현 데이터, 상기 복수의 프로브셋별 탐침 유전자 정보 및 상기 복수의 실험샘플별 실험 변인 정보를 DB로부터 수신하는 단계; 상기 탐침 유전자 정보에 포함된 복수의 이종상동성 유전자(orthologous gene)의 정보 및 상기 유전자 발현 데이터에 기초하여, 상기 복수의 이종상동성 유전자별 상기 복수의 실험샘플 각각에 대한 상기 발현값을 산출하는 단계; 상기 실험 변인 정보에 기초하여, 상기 복수의 실험샘플 중에서 적어도 하나의 대조군 및 실험군을 선별하는 단계; 및 상기 선별된 적어도 하나의 대조군 및 실험군 각각에 대응되는 상기 복수의 이종상동성 유전자별로 산출된 상기 발현값의 정보를 포함하는 프로파일 데이터를 생성하는 단계를 포함한다.

[0010] 바람직하게는, 상기 유전자 발현 데이터가 상기 복수의 프로브셋 및 상기 복수의 실험샘플 간의 상기 발현값을 나타내는 행렬의 형태일 때, 상기 행렬에 포함된 발현값의 크기에 기초하여, 선택적으로 상기 발현값을 상기 발현값의 로그값으로 대체하는 단계; 및 상기 행렬의 행 또는 열에 포함된 결손 원소의 개수에 기초하여, 선택적으로 상기 행 또는 열을 삭제하거나 상기 결손 원소를 보정값으로 대체하는 단계를 더 포함할 수 있다.

[0011] 바람직하게는, 상기 결손 원소를 보정값으로 대체하는 단계는 상기 행렬의 행 또는 열에 포함된 결손 원소의 개수 및 전체 원소의 개수 간의 비율이 소정의 임계비율 이상이면, 상기 행 또는 열을 삭제하고, 상기 비율이 상기 임계비율 미만이면, 상기 결손 원소를 인접한 다른 원소의 상기 발현값에 기초한 보정값으로 대체할 수 있다.

[0012] 바람직하게는, 상기 복수의 실험샘플 각각에 대한 상기 발현값을 산출하는 단계는 상기 탐침 유전자 정보를 이용하여 상기 복수의 프로브셋과 상기 복수의 이종상동성 유전자 간의 다대다 매칭 관계를 결정하는 단계; 상기 다대다 매칭 관계를 다대일 매칭 관계로 변환하는 단계; 및 상기 다대일 매칭 관계를 일대일 매칭 관계로 변환하는 단계를 포함하고, 상기 다대일 매칭 관계 및 일대일 매칭 관계로 변환하는 단계는 상기 변환된 매칭 관계에 기초하여, 상기 복수의 이종상동성 유전자별 상기 복수의 실험샘플 각각에 대한 상기 발현값을 산출할 수 있다.

[0013] 바람직하게는, 상기 다대다 매칭 관계를 다대일 매칭 관계로 변환하는 단계는 복수의 상기 이종상동성 유전자에 매칭된 하나의 상기 프로브셋에 대하여, 상기 복수의 이종상동성 유전자 중 선택된 하나를 매칭하거나, 상기 복수의 이종상동성 유전자 각각에 대응되는 복수의 가상프로브셋으로 분할하여 매칭할 수 있다.

[0014] 바람직하게는, 상기 다대일 매칭 관계를 일대일 매칭 관계로 변환하는 단계는 복수의 상기 프로브셋에 매칭된 하나의 상기 이종상동성 유전자에 대하여, 상기 복수의 프로브셋별로 상기 복수의 실험샘플에 대하여 산출된 상기 발현값의 분산 또는 평균에 기초하여 선택된 하나의 상기 프로브셋을 매칭하거나, 상기 복수의 프로브셋을 하나의 가상프로브셋으로 통합하여 매칭할 수 있다.

[0015] 바람직하게는, 상기 적어도 하나의 대조군 및 실험군을 선별하는 단계는 상기 실험 변인 정보를 이용하여 상기 복수의 실험샘플에 포함된 대조군당 실험군의 개수 및 시계열적인 실험의 수행 여부를 판단하는 단계; 상기 판단결과에 따라, 상기 복수의 실험샘플을 이용하는 실험구성을 단조건 실험구성, 다조건 실험구성, 단조건 시계열 실험구성 및 다조건 시계열 실험구성 중의 하나로 결정하는 단계; 및 상기 결정된 실험구성에 대응되는 상기 적어도 하나의 대조군 및 실험군을 선별하는 단계를 포함할 수 있다.

[0016] 바람직하게는, 상기 프로파일 데이터를 생성하는 단계는 상기 선별된 적어도 하나의 대조군 및 실험군 각각에 대응되는 상기 복수의 이종상동성 유전자별로 산출된 상기 발현값의 평균, 분산 및 평균의 차이값 중 적어도 하

나를 더 포함하는 프로파일 데이터를 생성할 수 있다.

[0017] 또한, 상기 목적을 달성하기 위해, 본 발명에서 제공하는 이종상동유전자 발현 반응을 분석하기 위한 데이터 재구성 장치는 적어도 한 종류의 유전자를 탐침하는 프로브셋(probe set)과 유전자 발현 실험의 대상인 생물의 조직을 포함하는 실험샘플을 이용한 유전자 발현 반응을 분석하기 위한 데이터 재구성 장치에 있어서, 복수의 상기 프로브셋을 복수의 상기 실험샘플 각각과 반응시킨 발현 실험 결과인 발현값을 포함하는 유전자 발현 데이터, 상기 복수의 프로브셋별 탐침 유전자 정보 및 상기 복수의 실험샘플별 실험 변인 정보를 DB로부터 수신하는 수신부; 상기 탐침 유전자 정보에 포함된 복수의 이종상동성 유전자(orthologous gene)의 정보 및 상기 유전자 발현 데이터에 기초하여, 상기 복수의 이종상동성 유전자별 상기 복수의 실험샘플 각각에 대한 상기 발현값을 산출하는 매칭부; 상기 실험 변인 정보에 기초하여, 상기 복수의 실험샘플 중에서 적어도 하나의 대조군 및 실험군을 선별하는 선별부; 및 상기 선별된 적어도 하나의 대조군 및 실험군 각각에 대응되는 상기 복수의 이종상동성 유전자별로 산출된 상기 발현값의 정보를 포함하는 프로파일 데이터를 생성하는 생성부를 포함한다.

[0018] 바람직하게는, 상기 유전자 발현 데이터가 상기 복수의 프로브셋 및 상기 복수의 실험샘플 간의 상기 발현값을 나타내는 행렬의 형태일 때, 상기 행렬에 포함된 발현값의 크기에 기초하여, 선택적으로 상기 발현값을 상기 발현값의 로그값으로 대체하고, 상기 행렬의 행 또는 열에 포함된 결손 원소의 개수에 기초하여, 선택적으로 상기 행 또는 열을 삭제하거나 상기 결손 원소를 보정값으로 대체하는 데이터변환부를 더 포함할 수 있다.

[0019] 바람직하게는, 상기 데이터변환부는 상기 행렬의 행 또는 열에 포함된 결손 원소의 개수 및 전체 원소의 개수 간의 비율이 소정의 임계비율 이상이면, 상기 행 또는 열을 삭제하고, 상기 비율이 상기 임계비율 미만이면, 상기 결손 원소를 인접한 다른 원소의 상기 발현값에 기초한 보정값으로 대체할 수 있다.

[0020] 바람직하게는, 상기 매칭부는 상기 탐침 유전자 정보를 이용하여 상기 복수의 프로브셋과 상기 복수의 이종상동성 유전자 간의 다대다 매칭 관계를 결정하고, 상기 다대다 매칭 관계를 다대일 매칭 관계로 변환하고, 상기 다대일 매칭 관계를 일대일 매칭 관계로 변환하며, 상기 다대일 매칭 관계 및 일대일 매칭 관계로 변환할 때, 상기 변환된 매칭 관계에 기초하여, 상기 복수의 이종상동성 유전자별 상기 복수의 실험샘플 각각에 대한 상기 발현값을 산출할 수 있다.

[0021] 바람직하게는, 상기 다대다 매칭 관계를 다대일 매칭 관계로 변환할 때, 복수의 상기 이종상동성 유전자에 매칭된 하나의 상기 프로브셋에 대하여, 상기 복수의 이종상동성 유전자 중 선택된 하나를 매칭하거나, 상기 복수의 이종상동성 유전자 각각에 대응되는 복수의 가상프로브셋으로 분할하여 매칭할 수 있다.

[0022] 바람직하게는, 상기 다대일 매칭 관계를 일대일 매칭 관계로 변환할 때, 복수의 상기 프로브셋에 매칭된 하나의 상기 이종상동성 유전자에 대하여, 상기 복수의 프로브셋별로 상기 복수의 실험샘플에 대하여 산출된 상기 발현값의 분산 또는 평균에 기초하여 선택된 하나의 상기 프로브셋을 매칭하거나, 상기 복수의 프로브셋을 하나의 가상프로브셋으로 통합하여 매칭할 수 있다.

[0023] 바람직하게는, 상기 선별부는 상기 실험 변인 정보를 이용하여 상기 복수의 실험샘플에 포함된 대조군당 실험군의 개수 및 시계열적인 실험의 수행 여부를 판단하고, 상기 판단결과에 따라, 상기 복수의 실험샘플을 이용하는 실험구성을 단조건 실험구성, 다조건 실험구성, 단조건 시계열 실험구성 및 다조건 시계열 실험구성 중의 하나로 결정하고, 상기 결정된 실험구성에 대응되는 상기 적어도 하나의 대조군 및 실험군을 선별할 수 있다.

발명의 효과

[0024] 본 발명은 복잡한 구성을 가진 유전자 발현 반응 실험의 구성을 단순한 대조군-실험군의 구성으로 단순화하는 효과가 있다.

[0025] 또한, 본 발명은 자연어 형태로 표현되어 있는 실험 구성을 구조화된 형태로 재표현하여, 사용자가 실험의 구성을 빠르게 이해할 수 있고 이에 대한 대규모 처리가 가능하게 되는 효과가 있다.

도면의 간단한 설명

[0026] 도 1은 본 발명의 일 실시예에 따른 이종상동유전자 발현 반응을 분석하기 위한 데이터의 재구성 방법을 설명하기 위하여 도시한 흐름도이다.

도 2는 본 발명의 일 실시예에 따른 복수의 이종상동성 유전자별 복수의 실험샘플 각각에 대한 발현값 산출 방법을 설명하기 위하여 도시한 흐름도이다.

도 3은 본 발명의 일 실시예에 따른 적어도 하나의 대조군 및 실험군을 선별하는 방법을 설명하기 위하여 도시한 흐름도이다.

도 4는 본 발명의 일 실시예에 따른 이종상동유전자 발현 반응을 분석하기 위한 데이터 재구성 장치를 설명하기 위하여 도시한 도면이다.

도 5는 본 발명의 일 실시예에 따른 유전자 발현 데이터를 설명하기 위하여 도시한 도면이다.

도 6은 본 발명의 일 실시예에 따른 복수의 프로브셋과 복수의 이종상동성 유전자 간의 매칭 관계를 설명하기 위하여 도시한 도면이다.

도 7은 본 발명의 일 실시예에 따른 다조건 실험구성에 대응되는 대조군 및 실험군의 선별을 설명하기 위하여 도시한 도면이다.

도 8은 본 발명의 일 실시예에 따른 단조건 시계열 실험구성에 대응되는 대조군 및 실험군의 선별을 설명하기 위하여 도시한 도면이다.

도 9는 본 발명의 일 실시예에 따른 다조건 시계열 실험구성에 대응되는 대조군 및 실험군의 선별을 설명하기 위하여 도시한 도면이다.

도 10은 본 발명의 일 실시예에 따른 프로파일 데이터를 설명하기 위하여 도시한 도면이다.

도 11은 본 발명의 다른 실시예에 따른 이종상동유전자 발현 반응을 분석하기 위한 데이터의 재구성 방법을 설명하기 위하여 도시한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0027] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세한 설명에 상세하게 설명하고자 한다. 그러나, 이는 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 사용하였다.

[0028] 제1, 제2, A, B 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 발명의 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.

[0029] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다.

[0030] 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0031] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0032] 이하, 본 발명에 따른 바람직한 실시예를 첨부된 도면을 참조하여 상세하게 설명한다.

[0033] 도 1은 본 발명의 일 실시예에 따른 이종상동유전자 발현 반응을 분석하기 위한 데이터의 재구성 방법을 설명하기 위하여 도시한 흐름도이다.

[0034] 유전자 발현 반응(gene expression)은 적어도 한 종류의 유전자를 탐침하는 프로브셋(probe set)과 유전자 발현 실험의 대상인 생물의 조직을 포함하는 실험샘플을 반응시키는 실험의 결과일 수 있다. 한편, 프로브는 특정한 한 종류의 유전자를 탐침하고, 프로브셋은 여러 종류의 유전자를 탐침하기 위하여 복수의 프로브를 이용하여 구성될 수 있다.

[0035] 이종상동유전자 발현 반응을 분석하는 이유는 인간의 질병에 대한 원인 및 치료법을 찾기 위해 진행되어 온 동물을 이용한 유전자 발현 반응 실험의 결과로부터 이종상동유전자의 발현 반응 결과를 획득하는 경우, 이를 인간의 질병 치료 등에 이용할 수 있는 가능성이 있기 때문이다.

[0036] 단계 S110에서는, 데이터 재구성 장치가 복수의 프로브셋을 복수의 실험샘플 각각과 반응시킨 발현 실험 결과인 발현값을 포함하는 유전자 발현 데이터, 그 복수의 프로브셋별 탐침 유전자 정보 및 그 복수의 실험샘플별 실험 변인 정보를 DB로부터 수신한다.

[0037] 발현값은 복수의 프로브셋 중 하나와 복수의 실험샘플 중 하나를 반응시켰을 때의 반응 결과에 따른 값으로, 그 하나의 실험샘플에 포함된 유전자의 발현 정도를 나타내는 값으로 볼 수 있다. 또한, 유전자 발현 데이터는 복수의 프로브셋과 복수의 실험샘플 간의 모든 개별 조합에 대한 발현값을 포함할 수 있으며, 행렬(gene expression data matrix)의 형태로도 표시될 수 있다. 도 5를 참조하면, 유전자 발현 데이가 m개의 프로브셋과 n개의 실험샘플을 이용한 유전자 발현 실험의 결과인 발현값을 포함하고 있는 것을 확인할 수 있다.

[0038] 탐침 유전자 정보는 프로브셋이 탐침하는 적어도 하나의 종류의 유전자에 대한 정보이다. 이때, 프로브셋에 따라서 탐침할 수 있는 유전자의 종류 및 개수가 달라질 수 있다.

[0039] 실험 변인 정보는 각 실험샘플에 대한 정보로써, 각 실험샘플이 추출된 생물의 종(organism), 조직(tissue) 및 실험 환경(experiment conditions)에 대한 정보를 포함할 수 있다. 예를 들어, 어떤 실험샘플의 실험 변인 정보는 인간(종), 말초혈액단핵세포(조직) 및 고령자는 한달, 비고령자는 2주간 백신 접종 미실시(실험 환경)에 대한 정보를 포함할 수 있다.

[0040] 한편, 데이터 재구성 장치가 이러한 유전자 발현 데이터, 탐침 유전자 정보 및 실험 변인 정보를 DB로부터 수신할 수 있다.

[0041] DB는 유전자 발현 데이터, 탐침 유전자 정보 및 실험 변인 정보 등의 유전자 발현 반응 데이터를 대규모로 저장하고 있으면서, 이를 연구자들에게 제공하는 데이터 저장소일 수 있다. 데이터 재구성 장치는 DB로부터 대규모로 저장된 유전자 발현 반응 데이터를 수신하여, 그 수신된 실험의 결과를 인간의 측면에서, 즉, 이종상동유전자의 발현 반응 측면에서, 편리하게 해석할 수 있는 방법을 제공할 수 있다.

[0042] 또한, 도 11을 참조하면, DB는 포함된 각 유전자 발현 데이터에 대하여 플랫폼 설명 데이터 (platform description), 샘플 설명 데이터 (sample description) 및 실험 설명 데이터(experiment description)을 포함할 수 있다.

[0043] 유전자 발현 데이터는 m개의 프로브셋을 행으로 하고 n개의 실험샘플을 열로 하는 m x n 행렬일 수 있다. 이때, 행렬의 [i, j] 위치의 값은 j번째 실험샘플에서 i번째 프로브셋을 통해 검출된 유전자 발현값을 나타낼 수 있다.

[0044] 플랫폼 설명 데이터는 각 프로브셋에 대한 정보를 포함하고 있고, 특히 프로브셋이 측정하고자 하는 유전자 정보를 필수적으로 포함할 수 있다. 즉, 플랫폼 설명 데이터는 그 m개의 프로브셋에 대한 설명을 제공할 수 있다. 또한, 탐침 유전자 정보는 플랫폼 설명 데이터에 포함될 수 있다.

[0045] 샘플 설명 데이터는 각 실험샘플에 대한 정보를 포함하고, 특히, 실험샘플이 추출된 생물의 종 (organism), 실험샘플이 얻어진 조직 (tissue), 실험 환경 (experiment conditions)을 필수적으로 포함할 수 있다. 또한, 실험샘플에 대한 유전자 발현 데이터가 얻어진 방식 또는 조건에 대한 설명을 포함할 수 있다. 샘플 설명 데이터는 그 n개의 샘플에 대한 설명을 제공할 수 있다. 또한, 실험 변인 정보는 샘플 설명 데이터에 포함될 수 있다.

[0046] 마지막으로 실험 설명 데이터는 전체 실험의 구성 및 실험을 통해 확인하고자 하는 과학적 가설 등을 포함할 수 있다.

[0047] 다른 실시예에서는, 데이터 재구성 장치가, 유전자 발현 데이터가 복수의 프로브셋 및 복수의 실험샘플 간의 발현값을 나타내는 행렬의 형태일 때, 그 행렬을 재구성할 수 있다.

[0048] 이때, 데이터 재구성 장치가 그 행렬을 재구성하는 이유는, 데이터 재구성 장치를 통한 자동적이고 일괄적인 재

구성 과정을 거치려면, 그 행렬이 일괄적인 처리에 보다 적합한 형태로 존재하는 것이 효율적이기 때문일 수 있다.

[0049] 예컨대, 유전자 발현 데이터 행렬의 원소가 모두 발현값에 로그(log)함수를 취한 로그값으로 구성되거나, 결손된 원소가 없는 경우에, 데이터 재구성 장치가 그 유전자 발현 데이터 행렬을 이용하여 자동적이고 일괄적으로 데이터 재구성을 처리할 수 있다.

[0050] 보다 구체적으로는, 데이터 재구성 장치가 그 행렬에 포함된 발현값의 크기에 따라서, 선택적으로 그 발현값을 그 발현값의 로그값으로 대체할 수 있다.

[0051] 데이터 재구성 장치가 발현값을 발현값의 로그값으로 대체하는 것은 발현값이 너무 큰 값을 가져 발현값 간의 비교가 어려워지는 것을 방지하는 효과가 있기 때문이다.

[0052] 또한, 데이터 재구성 장치는 다음과 같은 발현값에 관한 조건에 따라, 그 발현값을 그 발현값의 로그값으로 대체할 수 있다.

[0053] 첫째로, 실험 변인 정보에 발현값의 단위가 로그값이라고 명시된 경우, 데이터 재구성 장치는 그 행렬에 포함된 발현값을 로그값으로 대체하지 않을 수 있다.

[0054] 둘째로, 데이터 재구성 장치는 그 행렬에 포함된 모든 발현값이 소정의 임계치(예, 20) 이상인 경우, 발현값이 로그값이 아닌 것으로 판단하여, 발현값을 로그값으로 대체할 수 있다.

[0055] 셋째로, 데이터 재구성 장치는 그 행렬에 포함된 모든 발현값이 그 임계치(예, 20)보다 작더라도, i) 모든 발현값이 0보다 크고, ii) 1보다 작은 발현값의 비율이 40~60%이고, iii) 1이상인 발현값이 60~40%이면, 발현값이 로그값이 아니라고 판단하여, 발현값을 로그값으로 대체할 수 있다.

[0056] 이때, 발현값의 로그값은 아래의 수학식 1에 의하여 산출될 수 있다.

수학식 1

[0057] $$y = \log_2(\max(x,0) + \alpha) - \log_2\alpha$$

[0058] 여기서, y는 발현값의 로그값이고, x는 발현값이고, max(a,b)는 a와 b 중에서 작지 않은 값을 선택하는 함수이고, α는 로그값이 가지는 최소값을 결정하는 파라미터이다.

[0059] 이때, 최소값을 결정하는 α는 아래와 같은 방법으로 결정될 수 있다.

[0060] 첫째로, 행렬에 포함된 발현값에 대하여, 1) 모든 발현값이 0보다 크고, 2) 1보다 작은 발현값이 전체 발현값의 40~60% 사이이고 3) 1 이상인 발현값이 전체 발현값의 40~60% 사이이면, α는 전체 발현값의 하위 1%에 해당하는 값으로 정해진다.

[0061] 둘째로, 첫째의 경우에 해당하지 않고, 상위 10%에 해당하는 값이 20보다 큰 경우, α는 1로 정해진다.

[0062] 셋째로, 첫째 및 둘째의 경우에 해당하지 않는 모든 경우, α는 모든 발현값의 최소값의 1/3로 정해진다.

[0063] 이와 같이, α는 발현값의 로그값이 너무 작은 음수가 되지 않도록 결정될 수 있다.

[0064] 또한, 데이터 재구성 장치가 그 행렬의 행 또는 열에 포함된 결손 원소의 개수에 따라서, 선택적으로 그 행 또는 열을 삭제하거나 그 결손 원소를 보정값으로 대체할 수 있다.

[0065] 예컨대, 데이터 재구성 장치는 그 행렬의 행 또는 열에 포함된 원소가 (1) 숫자가 아닌 형태로 주어지는 값 (2) 양의 무한 혹은 음의 무한으로 주어지는 값 (3) 0으로 나눈 숫자와 같이 정의되지 않은 값 (4) 기타 측정되지 않은 값 인 경우에 그 원소를 결손 원소로 판단할 수 있다.

[0066] 그리고, 데이터 재구성 장치가 그 행렬의 행 또는 열에 포함된 결손 원소의 개수 및 전체 원소의 개수 간의 비율과 소정의 임계비율의 크기 비교 결과에 따라서, 선택적으로 그 행 또는 열을 삭제하거나 그 결손 원소를 보정값으로 대체할 수 있다.

[0067] 보다 구체적으로는, 데이터 재구성 장치가 동일한 행 또는 열에서의 결손 원소와 전체 원소 간의 비율이 임계비

율 이상이면 그 행 또는 열을 삭제하고, 그 비율이 임계비율 미만이면 결손 원소를 인접한 다른 원소의 발현값에 기초한 보정값으로 대체할 수 있다.

[0068] 예컨대, 데이터 재구성 장치가 행렬에 포함된 각 행 또는 열에 포함된 결손원소의 비율이 임계비율 0.3(30%) 이상인 경우, 그 행 또는 열을 삭제할 수 있다. 그리고, 결손원소의 비율이 그 임계비율 0.3미만인 경우에는 그 결손 원소 주변의 다른 원소값(발현값)을 이용하여 그 결손 원소를 보정할 수 있다.

[0069] 이때, 데이터 재구성 장치는 k-nearest neighbor 기반의 결손값 보정 방식을 사용할 수 있고, 그 외의 다른 방식을 사용할 수도 있다.

[0070] 단계 S120에서는, 데이터 재구성 장치가 탐침 유전자 정보에 포함된 복수의 이종상동성 유전자(orthologous gene)의 정보 및 유전자 발현 데이터에 기초하여, 복수의 이종상동성 유전자별 복수의 실험샘플 각각에 대한 발현값을 산출한다.

[0071] 이종상동성 유전자는 서로 다른 유전자가 공통의 조상으로부터 종분화에 의하여 유래하였을 때, 그 유전자를 이종상동성 유전자라고 한다. 예를 들어, 사람과 쥐의 알파(α)헤모글로빈유전자는 이종상동성 유전자이다.

[0072] 데이터 재구성 장치는 탐침 유전자 정보에 포함된 탐침 유전자 중에서 특히 이종상동성 유전자에 대한 정보와 유전자 발현 데이터에 기초하여, 복수의 이종상동성 유전자별로 복수의 실험샘플 각각에 대한 발현값을 산출한다.

[0073] 즉, 데이터 재구성 장치가 복수의 프로브셋별로 복수의 실험샘플 각각에 대한 발현값을 포함한 유전자 발현 데이터 및 이종상동성 유전자에 대한 정보를 포함하는 탐침 유전자 정보를 이용하여, 복수의 이종상동성 유전자별로 복수의 실험샘플 각각에 대한 발현값을 산출할 수 있다.

[0074] 이때, 복수의 이종상동성 유전자별로 복수의 실험샘플 각각에 대한 발현값을 산출하는 방법은 도 2에 대한 설명에서 구체적으로 후술한다.

[0075] 단계 S130에서는, 데이터 재구성 장치가 실험 변인 정보에 기초하여, 그 복수의 실험샘플 중에서 적어도 하나의 대조군 및 실험군을 선별한다.

[0076] 대조군은 실험 결과가 제대로 도출되었는지의 여부를 판단하기 위해 어떤 조작이나 조건도 가하지 않은 실험샘플이고, 실험군은 실험 결과를 도출하기 위해 인위적 또는 어떤 조작을 통해 환경 설정을 한 실험샘플이다.

[0077] 실험 변인 정보는 복수의 실험샘플 각각에 대한 변인 정보(예, 종, 조직, 실험환경)를 포함하고 있으므로, 데이터 재구성 장치가 그 변인 정보에 따라서 복수의 실험샘플로부터 적어도 하나의 대조군 및 실험군을 선별할 수 있다.

[0078] 예컨대, 실험샘플의 실험대상이 젊은이, 나이든 남성, 나이든 여성의 3개로 나누어 지는 경우에, 데이터 재구성 장치가 {대조군, 실험군}을 {젊은이, 나이든 남성} 및 {젊은이, 나이든 여성}으로 선별할 수 있다.

[0079] 다른 실시예에서는, 데이터 재구성 장치가 사용자로부터 적어도 하나의 대조군 및 실험군의 선별 결과에 대한 정보를 수신할 수 있다.

[0080] 이때, 데이터 재구성 장치는 사용자로부터 실험 변인 정보를 분석한 결과에 따라 선별된 적어도 하나의 대조군 및 실험군에 대한 정보를 수신할 수 있다.

[0081] 마지막으로 단계 S140에서는, 데이터 재구성 장치가 그 선별된 적어도 하나의 대조군 및 실험군 각각에 대응되는 복수의 이종상동성 유전자별로 산출된 발현값의 정보를 포함하는 프로파일 데이터를 생성한다.

[0082] 프로파일 데이터는 그 적어도 하나의 대조군 및 실험군 각각에 대한 정보 및 그 적어도 하나의 대조군 및 실험군 각각에 대응되는 복수의 이종상동성 유전자별로 산출된 발현값의 정보를 포함할 수 있다.

[0083] 예컨대, 도 10의 좌측(유전자 발현 데이터 행렬)을 참조하면, 프로파일 데이터에 대조군 샘플 및 실험군 샘플에 각각 대응되는 유전자 1 내지 유전자 m”의 발현값 정보가 포함된 것을 확인할 수 있다.

[0084] 다른 실시예에서는, 데이터 재구성 장치가 그 선별된 적어도 하나의 대조군 및 실험군 각각에 대응되는 복수의 이종상동성 유전자별로 산출된 발현값의 평균, 분산 및 평균의 차이값 중 적어도 하나를 더 포함하는 프로파일 데이터를 생성할 수 있다.

[0085] 즉, 데이터 재구성 장치가 대조군 및 실험군에 대응되는 복수의 이종상동성 유전자별로 산출된 발현값의 평균,

분산 및 평균의 차이값과 같은 기초적인 통계 데이터를 추가적으로 포함하는 프로파일 데이터를 생성할 수 있다.

[0086] 예컨대, 도 10의 우측(기초 통계값 행렬)을 참조하면, 프로파일 데이터에 대조군 평균, 대조군 분산, 실험군 평균, 실험군 분산 및 평균의 차이에 대한 기초적인 통계데이터가 더 포함된 것을 확인할 수 있다.

[0087] 또 다른 실시예에서는, 데이터 재구성 장치가 단계 S120과 단계 S130의 순서를 서로 바꾸어 수행할 수 있다.

[0088] 단계 S120은 탐침 유전자 정보 및 유전자 발현 데이터를 이용하고 있으며, 단계 S130은 실험 변인 정보를 이용하고 있으므로, 데이터 재구성 장치가 서로 순서를 바꾸어 수행하여도 프로파일 데이터의 생성 결과에는 영향을 미치지 않을 수 있다.

[0089] 이와 같이, 본 발명의 일 실시예에 따른 이종상동성유전자 발현 반응을 분석하기 위한 데이터의 재구성 방법은 복잡한 실험의 구성을 대조군-실험군의 구성으로 단순화하고, 구조화하여, 실험의 구성을 빠르게 이해할 수 있고 데이터에 대한 자동화되고 일괄적인 대규모 처리가 가능하게 되는 효과가 있다.

[0090] 도 2는 본 발명의 일 실시예에 따른 복수의 이종상동성 유전자별 복수의 실험샘플 각각에 대한 발현값 산출 방법을 설명하기 위하여 도시한 흐름도이다.

[0091] 단계 S210에서는, 데이터 재구성 장치가 탐침 유전자 정보를 이용하여 복수의 프로브셋과 복수의 이종상동성 유전자 간의 다대다 매칭 관계를 결정한다.

[0092] 데이터 재구성 장치는 탐침 유전자 정보로부터 각 프로브셋이 탐침하는 이종상동성 유전자에 대한 정보를 획득할 수 있다. 그리고, 결과적으로 이종상동성 유전자를 탐침하는 적어도 하나의 프로브셋에 대한 정보 또한 획득할 수 있다.

[0093] 따라서, 도 6의 (a)를 참조하면, 데이터 재구성 장치는 복수의 프로브셋과 그 복수의 프로브셋이 탐침할 수 있는 복수의 이종상동성 유전자 간의 다대다 매칭 관계를 결정할 수 있다.

[0094] 단계 S220에서는, 데이터 재구성 장치가 그 다대다 매칭 관계를 다대일 매칭 관계로 변환한다.

[0095] 예컨대, 도 6의 (b)를 참조하면, 데이터 재구성 장치가 복수의 이종상동성 유전자와 매칭된 하나의 프로브셋에 대하여, 그 복수의 이종상동성 유전자 중에서 선택된 하나와만 매칭되도록 하여 다대다 매칭 관계를 다대일 매칭 관계로 변환할 수 있다.

[0096] 다른 실시예에서는, 데이터 재구성 장치가 복수의 이종상동성 유전자에 매칭된 하나의 프로브셋에 대하여, 그 복수의 이종상동성 유전자 중 선택된 하나를 매칭하거나, 그 복수의 이종상동성 유전자 각각에 대응되는 복수의 가상프로브셋으로 분할하여 매칭할 수 있다.

[0097] 즉, 데이터 재구성 장치가 균일 분할 매칭 방법을 적용하여 다대다 매칭 관계를 다대일 매칭 관계로 변환할 수 있다.

[0098] 예컨대, 프로브셋 p가 유전자 g_1, g_2, …, g_k의 k개의 유전자와 매칭되어 있을 수 있다. 이때, k개의 유전자에 대하여 각각 대응되는 가상프로브셋 p_1, p_2, …, p_k에 대하여, 각 가상프로브셋의 유전자 발현값은 아래의 수학식 2를 이용하여 균일한 값으로 산출될 수 있다.

수학식 2

[0099]
$$E[p_i] = \frac{E[p]}{k}, \qquad i=1,2,\ldots,k$$

[0100] 여기서, p_i는 가상프로브셋이고, p는 프로브셋이고, E[p]는 프로브셋 p의 발현값이고, k는 프로브셋 p와 매칭되는 유전자의 개수이다.

[0101] 이 경우, 프로브셋 p가 k개의 가상프로브셋에 균일하게 나누어져 발현값이 할당됨으로써 데이터 재구성 장치가 다대일 매칭을 구현할 수 있다.

[0102] 또는, 데이터 재구성 장치가 임의 매칭 방법을 적용하여 다대다 매칭 관계를 다대일 매칭 관계로 변환할 수 있다.

[0103] 예컨대, 데이터 재구성 장치가 복수의 이종상동성 유전자와 매칭된 하나의 프로브셋에 대하여, 그 복수의 이종상동성 유전자 중에서 임의로 선택된 하나와만 매칭되도록 하여 다대다 매칭 관계를 다대일 매칭 관계로 변환할 수 있다.

[0104] 또는, 데이터 재구성 장치가 상관관계 기반 매칭 방법을 적용하여 다대다 매칭 관계를 다대일 매칭 관계로 변환할 수 있다.

[0105] 예컨대, 프로브셋 p가 유전자 g_1, g_2, …, g_k의 k개의 유전자와 매칭되고, 유전자 g_i에는 p_{i1}, p_{i2}, …, p_{il}의 l개의 프로브셋이 매칭되어 있을 수 있다. (단, 0 < i =< k) 이때, 데이터 재구성 장치는, 복수의 실험샘플에 대하여, 수학식 2에 의해 산출된 프로브셋 p의 유전자 g_1, g_2, …, g_k에 대한 발현값의 분포 및 복수의 프로브셋(p_{i1}, p_{i2}, …, p_{il})의 유전자 g_i에 대한 평균 발현값의 분포 간의 상관관계를 산출하여, 가장 상관관계가 높은 유전자 하나를 선택하여 매칭할 수 있다. 이때, 상관관계는 두 변수간의 관련성을 구하기 위하여 보편적으로 이용되는 피어슨 상관관계(Pearson's correlation)을 이용하여 산출될 수 있다. 즉, 피어슨 상관관계가 1에 가까워질수록 더 높은 상관관계를 갖는다고 할 수 있다.

[0106] 이때, 복수의 프로브셋(p_{i1}, p_{i2}, …, p_{il})의 유전자 g_i에 대한 평균 발현값은 아래의 수학식 3을 이용하여 산출될 수 있다.

수학식 3

[0107]
$$E[g_i] = \sum_{j=1}^{l} \frac{E[p_{ij}]}{l}$$

[0108] 여기서, g_i는 프로브셋 p와 매칭되는 i번째 유전자이고, p_{ij}는 유전자 g_i와 매칭되는 j번째 프로브셋이고, $E[p_{ij}]$는 프로브셋 p_{ij}의 발현값이고, l은 유전자 g_i에 매칭되는 프로브셋의 개수이다.

[0109] 단계 S230에서는, 데이터 재구성 장치가 그 다대일 매칭 관계를 일대일 매칭 관계로 변환한다.

[0110] 예컨대, 도 6의 (c)를 참조하면, 데이터 재구성 장치가 복수의 프로브셋과 매칭된 하나의 이종상동성 유전자에 대하여, 그 복수의 프로브셋 중에서 선택된 하나와만 매칭되도록 하여 다대일 매칭 관계를 일대일 매칭 관계로 변환할 수 있다.

[0111] 다른 실시예에서는, 데이터 재구성 장치가 복수의 프로브셋에 매칭된 하나의 이종상동성 유전자에 대하여, 그 복수의 프로브셋별로 복수의 실험샘플에 대하여 산출된 발현값의 분산 또는 평균에 기초하여 선택된 하나의 프로브셋을 매칭하거나, 그 복수의 프로브셋을 하나의 가상프로브셋으로 통합하여 매칭할 수 있다.

[0112] 즉, 데이터 재구성 장치가 최대 분산 매칭 방법을 적용하여 다대일 매칭 관계를 일대일 매칭 관계로 변환할 수 있다.

[0113] 예컨대, 유전자 g에는 p_1, p_2, …, p_k의 k개의 프로브셋이 매칭되어 있을 수 있다. 이때, s번째 실험샘플에 대한 프로브셋 p_i(단, 0 < i =< k)의 발현값을 $E[p_i,s]$라고 하면, 프로브셋 p_i의 유전자 발현값의 분산은 아래의 수학식 4를 이용하여 산출될 수 있다.

수학식 4

[0114] $$S_{p_i} = \frac{1}{n-1} \sum_{s=1}^{n} \left(E[p_i, s] - \frac{1}{n} \sum_{t=1}^{n} E[p_i, t] \right)^2$$

[0115] 여기서, S_{pi}는 프로브셋 p_i의 발현값의 분산이고, p_i는 유전자 g에 매칭되는 i번째 프로브셋이고, n는 복수의 실험샘플의 개수이고, $E[p_i,s]$는 s번째 실험샘플에 대한 프로브셋 p_i의 발현값이다.

[0116] 그리고, 데이터 재구성 장치가 그 산출된 분산이 가장 큰 프로브셋을 유전자 g에 매칭시킬 수 있다.

[0117] 또는, 데이터 재구성 장치가 최대 평균 매칭 방법을 적용하여 다대일 매칭 관계를 일대일 매칭 관계로 변환할 수 있다.

[0118] 예컨대, 유전자 g에는 p_1, p_2, …, p_k의 k개의 프로브셋이 매칭되어 있을 수 있다. 이때, 복수의 실험샘플 s에 대한 프로브셋 p_i(단, $0 < i =< k$)의 발현값을 $E[p_i,s]$라고 하면, 프로브셋 p_i의 유전자 발현값의 평균은 아래의 수학식 5를 이용하여 산출될 수 있다.

수학식 5

[0119] $$A_{p_i} = \frac{1}{n} \sum_{s=1}^{n} E[p_i, s]$$

[0120] 여기서, A_{pi}는 프로브셋 p_i의 발현값의 평균이고, p_i는 유전자 g에 매칭되는 i번째 프로브셋이고, n는 복수의 실험샘플의 개수이고, $E[p_i,s]$는 s번째 실험샘플에 대한 프로브셋 p_i의 발현값이다.

[0121] 그리고, 데이터 재구성 장치가 그 산출된 평균이 가장 큰 프로브셋을 유전자 g에 매칭시킬 수 있다.

[0122] 또는, 데이터 재구성 장치가 평균 매칭 방법을 적용하여 다대일 매칭 관계를 일대일 매칭 관계로 변환할 수 있다.

[0123] 예컨대, 유전자 g에는 p_1, p_2, …, p_k의 k개의 프로브셋이 매칭되어 있을 수 있다. 이때, k개의 프로브셋이 통합된 가상프로브셋 p_g를 생성하고, 그 가상프로브셋 p_g의 유전자 발현값은 아래의 수학식 6을 이용하여 산출될 수 있다.

수학식 6

[0124] $$E[p_g] = \sum_{i=1}^{k} E[p_i]/k$$

[0125] 여기서, $E[p_g]$는 가상프로브셋 p_g의 유전자 발현값이고, p_g는 통합된 가상프로브셋이고, p_i는 유전자 g에 매칭되는 i번째 프로브셋이고, k는 유전자 g에 매칭되는 프로브셋의 개수이다.

[0126] 그리고, 데이터 재구성 장치가 그 가상프로브셋을 유전자 g에 매칭시킬 수 있다.

[0127] 한편, 데이터 재구성 장치는 다대일 매칭 관계 및 일대일 매칭 관계로 변환할 때, 그 변환된 매칭 관계에 기초하여, 복수의 이종상동성 유전자별 복수의 실험샘플 각각에 대한 발현값을 산출할 수 있다.

[0128] 이와 같이, 본 발명의 일 실시예에 따른 복수의 이종상동성 유전자별 복수의 실험샘플 각각에 대한 발현값 산출

방법은, 복수의 프로브셋과 복수의 실험샘플간의 유전자 발현 데이터를 복수의 이종상동성 유전자와 복수의 실험샘플간의 유전자 발현 데이터로 변환할 수 있는 효과가 있다.

[0129] 도 3은 본 발명의 일 실시예에 따른 적어도 하나의 대조군 및 실험군을 선별하는 방법을 설명하기 위하여 도시한 흐름도이다.

[0130] 단계 S310에서는, 데이터 재구성 장치가 실험 변인 정보를 이용하여 복수의 실험샘플에 포함된 대조군당 실험군의 개수 및 시계열적인 실험의 수행 여부를 판단한다.

[0131] 이는, 대조군당 실험군의 개수 및 시계열적인 실험의 수행 여부에 따라 다음 단계에서 실험구성에 대한 결정 결과가 달라질 수 있기 때문이다.

[0132] 예컨대, 대조군당 실험군의 개수는 단수인지 복수인지의 판단이 필요하고, 시계열적인 실험의 수행 여부는 시간의 흐름에 따라서 실험이 수행(예, T시간 시점, T+1시간 시점, T+2시간 시점으로 나뉘어 실험)되었는지의 판단이 필요할 수 있다.

[0133] 단계 S320에서는, 데이터 재구성 장치가, 그 판단결과에 따라, 그 복수의 실험샘플을 이용하는 실험구성을 단조건 실험구성, 다조건 실험구성, 단조건 시계열 실험구성 및 다조건 시계열 실험구성 중의 하나로 결정한다.

[0134] 단조건 실험구성은 원 실험 구성이 하나의 대조군과 하나의 실험군으로 단순 구성된 경우이다. 다조건 실험구성은, 도 7을 참조하면, 원 실험이 하나의 대조군과 N개의 실험군으로 구성된 경우이다. 단조건 시계열 실험 구성은, 도 8을 참조하면, 원 실험이 하나의 조건에서 T_0, T_1, …, T_N의 시간에서 측정한 시간에 따른 실험샘플의 변화에 대한 측정인 경우이다. 다조건 시계열 실험 구성은, 도 9를 참조하면, 원 실험이 두 개의 조건에서 T_0, T_1, …, T_N의 시간에서 측정한 시간에 따른 실험샘플의 변화에 대한 측정인 경우이다.

[0135] 단계 S330에서는, 데이터 재구성 장치가 그 결정된 실험구성에 대응되는 적어도 하나의 대조군 및 실험군을 선별한다.

[0136] 데이터 재구성 장치는 그 결정된 실험구성이 단조건 실험구성인 경우, 하나의 대조군 및 실험군을 선별할 수 있다.

[0137] 또한, 도 7을 참조하면, 데이터 재구성 장치는 그 결정된 실험구성이 다조건 실험구성(N개의 실험군)인 경우, N개의 대조군 및 실험군을 선별할 수 있다.

[0138] 또한, 도 8을 참조하면, 데이터 재구성 장치는 그 결정된 실험구성이 단조건 시계열 실험구성(T_0, T_1, …, T_N의 시간에서 측정)인 경우, 기준 시간(T_0 또는 T_N)의 실험을 대조군으로 하고, 나머지 시간에 대한 실험을 다수의 실험군으로 설정하여, 모두 N개의 대조군 및 실험군을 선별할 수 있다.

[0139] 또한, 도 9를 참조하면, 데이터 재구성 장치는 그 결정된 실험구성이 2개의 다조건 시계열 실험구성(T_0, T_1, …, T_N의 시간에서 측정)인 경우, 먼저 각 조건에 대하여 단조건 시계열 실험 구성과 같이 N개의 대조군 및 실험군을 선별할 수 있다. 그리고, 추가적으로 동시간 대의 두 조건의 실험을 비교하여 N+1개의 대조군 및 실험군을 선별할 수 있다. 이때, 최종적으로 3N+1개의 대조군 및 실험군을 선별할 수 있다.

[0140] 이상까지 이종상동유전자 발현 반응을 분석하기 위한 데이터의 재구성 방법을 설명하였으며, 다음으로는 실제의 데이터 재구성 과정에 기초한 실시예를 설명한다.

[0141] 다음은, 실시예1의 데이터 재구성 과정에 대한 설명이다.

[0142] 1) 유전자 발현 데이터를 얻기 위해 Gene Expression Omnibus로부터 면역 관련 질병을 연구하기 위한 모델 실험 데이터를 수집하였다. GSE65219는 연령이나 성별에 따른 면역체계의 변화를 관찰하는 실험 데이터이다.

[0143] 2) 유전자 발현 데이터 행렬은 47,322개의 프로브셋을 행으로 하고 176개의 실험샘플을 열로 하는 47,322 x 176 행렬로 주어진다.

[0144] 3) 플랫폼 설명 데이터는 47,322개의 프로브셋에 ID, GenBank Accession number, Gene symbol, Description 4가지 정보를 제공하고 있으며, 이 중에서 Gene symbol 항목을 통해 측정하고자 하는 이종상동상 유전자를 매칭

시킬 수 있다.

[0145] 4) 샘플 설명 데이터의 실험샘플이 추출된 생물의 종은 인간, 실험샘플이 얻어진 조직은 말초혈액단핵세포, 실험 환경은 고령자의 경우 한달, 비고령자의 경우 2주간 감염이나 백신 접종을 하지 않도록 했다. 유전자 발현 데이터가 얻어진 방식은 오전 8시에서 12시 사이 혈액응고방지제가 포함된 튜브로 혈액을 채취해 백혈구를 따로 제외하고 수집했다.

[0146] 5) 로그값 변화 여부 확인을 위해 확인 결과 명시적으로 표기되어 있지 않지만 행렬의 최대값이 20보다 작고, 1보다 작은 값이 전체 값의 40~60% 사이라는 조건에 해당하지 않으므로, 로그화되어있다고 판단한다.

[0147] 6) 로그화 되어있으므로 로그값 변환 단계는 건너뛴다.

[0148] 7) 결손값 확인 단계에서 4가지 종류의 결손값 모두 발견되지 않았으므로 결손 행/열 제거 단계, 결손값 보정 단계 또한 건너 뛴다.

[0149] 8) 유전자 매칭부에서 다대다 프로브셋-유전자 매칭을 다대일 매칭으로 변환하는 방법으로는 임의 매칭 방법을 사용했다.

[0150] 9) 다대일 프로브셋-유전자 매칭을 일대일 매칭으로 변환하는 법은 최대 분산 매칭 방식을 이용했다. 매칭 결과 22,452개의 프로브셋과 유전자가 매칭되었다.

[0151] 10) GSE65219는 다조건 실험 구성이다. 실험샘플은 young control, old male, old female로 나뉘어져있다. 이를 two-class 실험 디자인으로 변환하면, young control vs. old male, young control vs. old female 두 개의 프로필을 얻을 수 있다.

[0152] 다음은, 실시예2의 데이터 재구성 과정에 대한 설명이다.

[0153] 1) 유전자 발현 데이터를 얻기 위해 Gene Expression Omnibus로부터 면역 관련 질병을 연구하기 위한 모델 실험 데이터를 수집하였다. GSE46600는 형질과 시간에 따른 CD4T세포의 반응을 관찰하는 실험 데이터이다.

[0154] 2) 유전자 발현 데이터 행렬은 45,101개의 프로브셋을 행으로 하고 44개의 실험샘플을 열로 하는 45101 x 44 행렬로 주어진다.

[0155] 3) 플랫폼 설명 데이터는 45,101개의 프로브셋에 ID, GenBank Accession number, Gene symbol, Description 등 등 16가지 정보를 제공하고 있으며, 이 중에서 Gene symbol 항목을 통해 측정하고자 하는 이종상동상 유전자를 매칭시킬 수 있다.

[0156] 4) 샘플 설명 데이터의 실험샘플이 추출된 생물의 종은 쥐, 실험샘플이 얻어진 조직은 비장의 백혈구, 실험 환경과 유전자 발현 데이터가 얻어진 방식은 명시되어 있지 않았다.

[0157] 5) 로그값 변화 여부 확인을 위해 확인 결과, 명시적으로 표기되어 있지 않지만 전체 데이터 행렬의 최대값이 20보다 크므로 로그화되지 않았다고 판단한다.

[0158] 6) 로그값 변환 단계에서 유전자 발현 데이터의 값을 x라고 하였을 때, 로그화된 값 y는 다음과 같은 수식을 통해 계산된다.

[0159] 7) 이때, 전체 행렬 데이터에서 1보다 작은 값이 전체 값의 40~60%를 초과 하고, 상위 10%에 해당하는 값이 20보다 크므로 α는 1로 정해진다.

[0160] 8) 결손값 확인 단계에서 4가지 종류의 결손값 모두 발견되지 않았으므로 결손 행/열 제거 단계, 결손값 보정 단계 또한 건너 뛴다.

[0161] 9) 유전자 매칭부에서 다대다 프로브셋-유전자 매칭을 다대일 매칭으로 변환하는 방법으로는 임의 매칭 방법을 사용했다.

[0162] 10) 다대일 프로브셋-유전자 매칭을 일대일 매칭으로 변환하는 법은 최대 분산 매칭 방식을 이용했다. 매칭 결과 21,678개의 프로브셋과 유전자가 매칭되었다.

[0163] 11) GSE46600는 다조건 시계열 실험 구성이다. 실험샘플은 NOD 형질 2/3/4주차, NOR 형질 2/3/4주차, 그리고 C57BL/6 형질 2/3/4주차로 나뉘어져있다. 이를 two-class 실험 디자인으로 변환하면, NOD 형질 2주차 vs. NOD

형질 3/4주차, NOR 형질 2주차 vs. NOR 형질 3/4주차, C57BL/6 형질 2주차 vs. C57BL/6 형질 3/4주차, 그리고 NOD 형질 2주차 vs. NOR 형질 2주차, NOD 형질 2주차 vs. C57BL/6 형질 2주차, NOD 형질 3주차 vs. NOR 형질 3주차, NOD 형질 3주차 vs. C57BL/6 형질 3주차, NOD 형질 4주차 vs. NOR 형질 4주차, NOD 형질 4주차 vs. C57BL/6 형질 4주차, 총 열두 개의 프로필을 얻을 수 있다.

[0164] 도 4는 본 발명의 일 실시예에 따른 이종상동유전자 발현 반응을 분석하기 위한 데이터 재구성 장치를 설명하기 위하여 도시한 도면이다.

[0165] 도 4를 참조하면, 이종상동유전자 발현 반응을 분석하기 위한 데이터 재구성 장치(400)는 수신부(410), 매칭부(420), 선별부(430) 및 생성부(440)를 포함한다. 또한, 선택적으로 데이터변환부(미도시)를 더 포함할 수 있다.

[0166] 한편, 데이터 재구성 장치(400)는 서버, 데스크탑 PC, 노트북PC, 태블릿 및 스마트폰등에 탑재될 수 있다.

[0167] 수신부(410)는 복수의 프로브셋을 복수의 실험샘플 각각과 반응시킨 발현 실험 결과인 발현값을 포함하는 유전자 발현 데이터, 복수의 프로브셋별 탐침 유전자 정보 및 복수의 실험샘플별 실험 변인 정보를 DB로부터 수신한다.

[0168] 매칭부(420)는 그 탐침 유전자 정보에 포함된 복수의 이종상동성 유전자의 정보 및 그 유전자 발현 데이터에 기초하여, 복수의 이종상동성 유전자별 복수의 실험샘플 각각에 대한 발현값을 산출한다.

[0169] 다른 실시예에서는, 매칭부(420)는 탐침 유전자 정보를 이용하여 복수의 프로브셋과 복수의 이종상동성 유전자 간의 다대다 매칭 관계를 결정하고, 그 다대다 매칭 관계를 다대일 매칭 관계로 변환하고, 그 다대일 매칭 관계를 일대일 매칭 관계로 변환할 수 있다. 그리고, 다대일 매칭 관계 및 일대일 매칭 관계로 변환할 때, 그 변환된 매칭 관계에 기초하여, 복수의 이종상동성 유전자별 복수의 실험샘플 각각에 대한 발현값을 산출할 수 있다.

[0170] 또 다른 실시예에서는, 매칭부(420)가 다대다 매칭 관계를 다대일 매칭 관계로 변환할 때, 복수의 이종상동성 유전자에 매칭된 하나의 프로브셋에 대하여, 그 복수의 이종상동성 유전자 중 선택된 하나를 매칭하거나, 그 복수의 이종상동성 유전자 각각에 대응되는 복수의 가상프로브셋으로 분할하여 매칭할 수 있다.

[0171] 또 다른 실시예에서는, 매칭부(420)가 다대일 매칭 관계를 일대일 매칭 관계로 변환할 때, 복수의 프로브셋에 매칭된 하나의 이종상동성 유전자에 대하여, 그 복수의 프로브셋별로 복수의 실험샘플에 대하여 산출된 발현값의 분산 또는 평균에 기초하여 선택된 하나의 프로브셋을 매칭하거나, 그 복수의 프로브셋을 하나의 가상프로브셋으로 통합하여 매칭할 수 있다.

[0172] 선별부(430)는 그 실험 변인 정보에 기초하여, 복수의 실험샘플 중에서 적어도 하나의 대조군 및 실험군을 선별한다.

[0173] 다른 실시예에서는, 선별부(430)는 실험 변인 정보를 이용하여 복수의 실험샘플에 포함된 대조군당 실험군의 개수 및 시계열적인 실험의 수행 여부를 판단하고, 그 판단결과에 따라, 복수의 실험샘플을 이용하는 실험구성을 단조건 실험구성, 다조건 실험구성, 단조건 시계열 실험구성 및 다조건 시계열 실험구성 중의 하나로 결정하고, 그 결정된 실험구성에 대응되는 적어도 하나의 대조군 및 실험군을 선별할 수 있다.

[0174] 생성부(440)는 그 선별된 적어도 하나의 대조군 및 실험군 각각에 대응되는 복수의 이종상동성 유전자별로 산출된 발현값의 정보를 포함하는 프로파일 데이터를 생성한다.

[0175] 마지막으로 데이터변환부(미도시)는 유전자 발현 데이터가 복수의 프로브셋 및 복수의 실험샘플 간의 발현값을 나타내는 행렬의 형태일 때, 그 행렬에 포함된 발현값의 크기에 기초하여, 선택적으로 발현값을 발현값의 로그값으로 대체하고, 그 행렬의 행 또는 열에 포함된 결손 원소의 개수에 기초하여, 선택적으로 그 행 또는 열을 삭제하거나 그 결손 원소를 보정값으로 대체한다.

[0176] 다른 실시예에서는, 데이터변환부는 그 행렬의 행 또는 열에 포함된 결손 원소의 개수 및 전체 원소의 개수 간의 비율이 소정의 임계비율 이상이면, 그 행 또는 열을 삭제하고, 그 비율이 임계비율 미만이면, 그 결손 원소를 인접한 다른 원소의 발현값에 기초한 보정값으로 대체할 수 있다.

[0177] 한편, 상술한 본 발명의 실시예들은 컴퓨터에서 실행될 수 있는 프로그램으로 작성가능하고, 컴퓨터로 읽을 수 있는 기록매체를 이용하여 상기 프로그램을 동작시키는 범용 디지털 컴퓨터에서 구현될 수 있다.

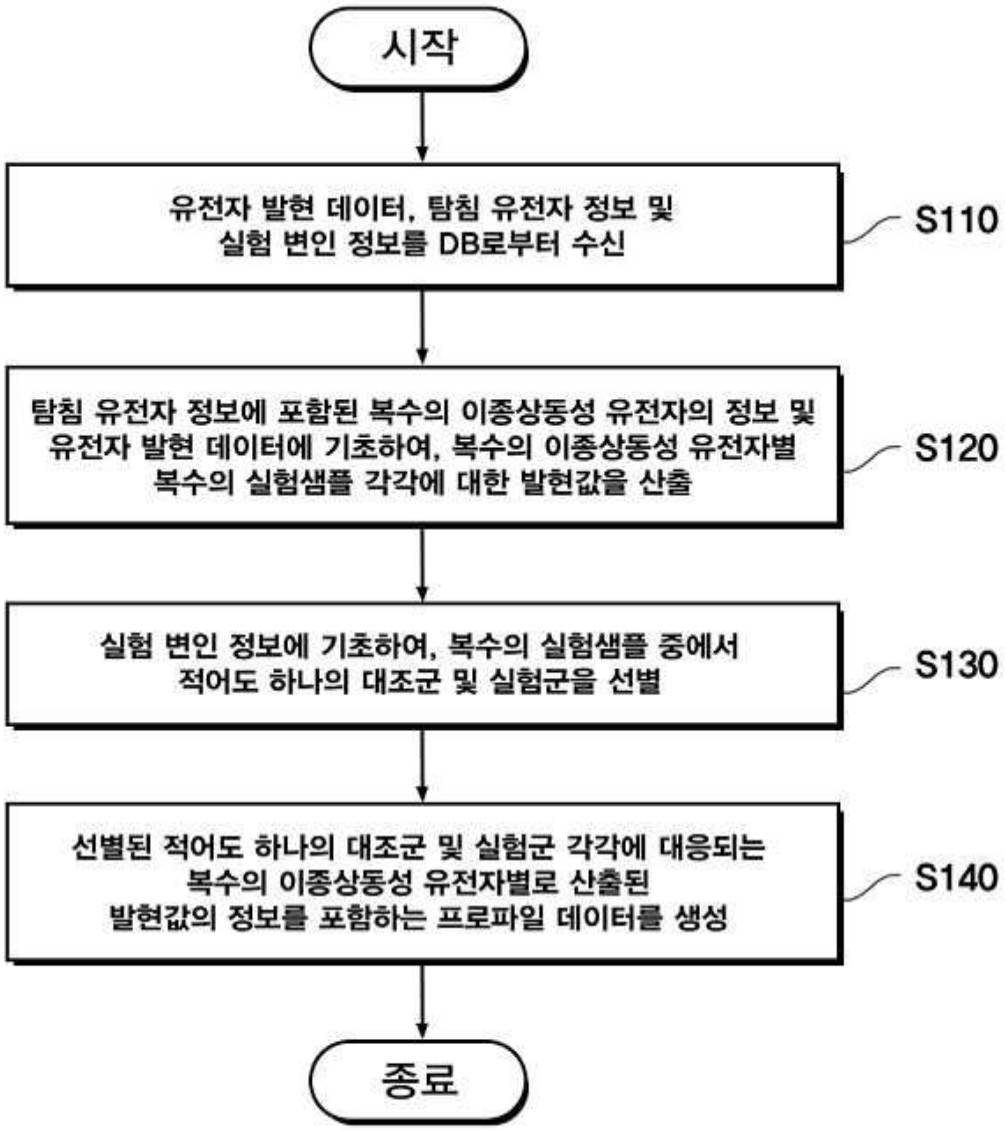
[0178] 상기 컴퓨터로 읽을 수 있는 기록매체는 마그네틱 저장매체(예를 들면, 롬, 플로피 디스크, 하드디스크 등), 광

학적 판독 매체(예를 들면, 시디롬, 디브이디 등) 를 포함한다.

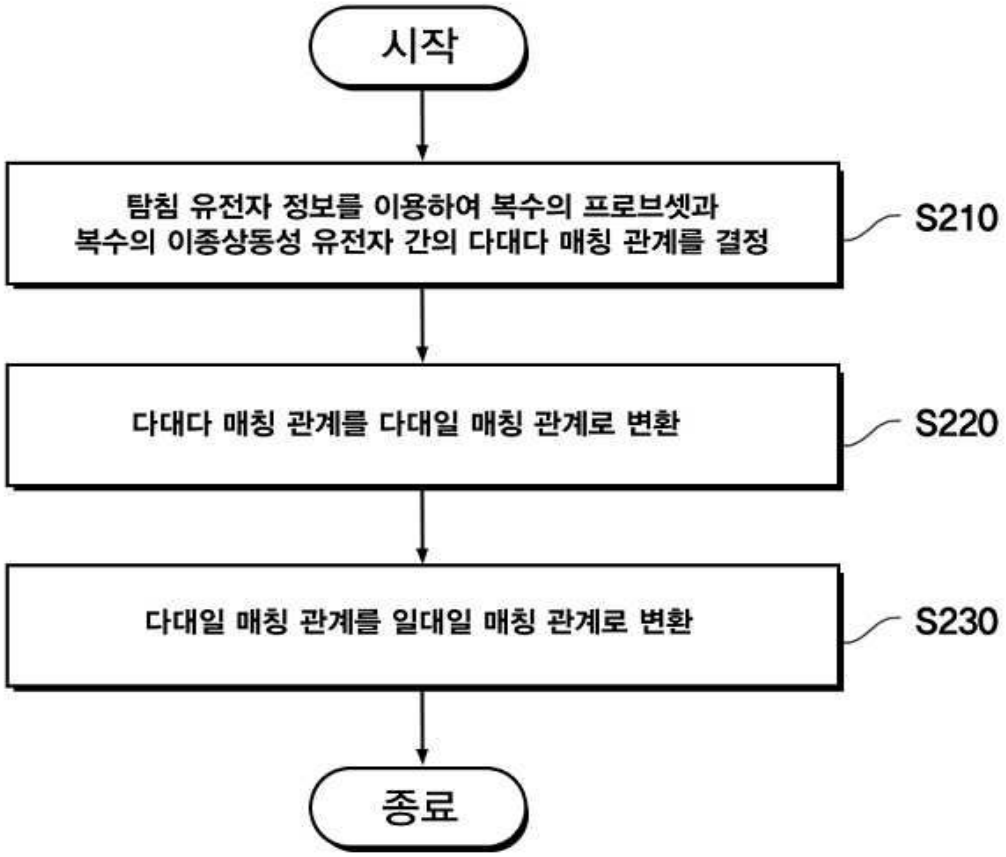
[0179] 이제까지 본 발명에 대하여 그 바람직한 실시예들을 중심으로 살펴보았다. 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현될 수 있음을 이해할 수 있을 것이다. 그러므로 개시된 실시예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다. 본 발명의 범위는 전술한 설명이 아니라 특허청구범위에 나타나 있으며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명에 포함된 것으로 해석되어야 할 것이다.

도면

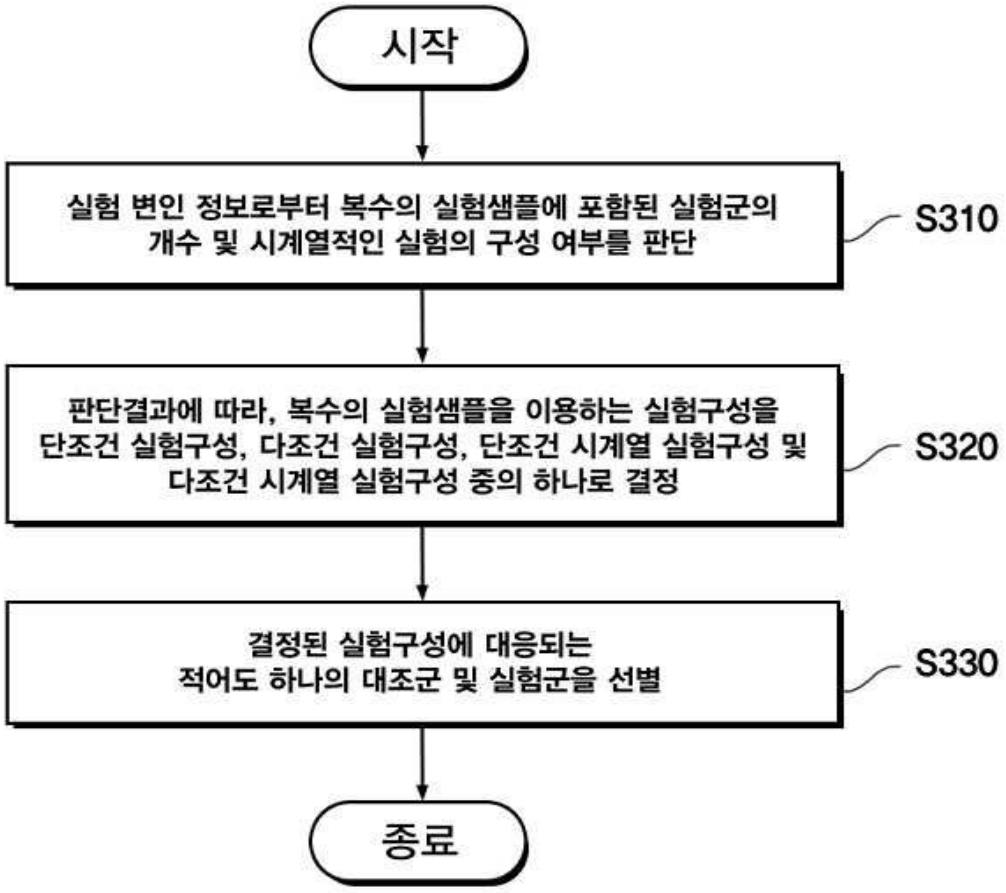
도면1



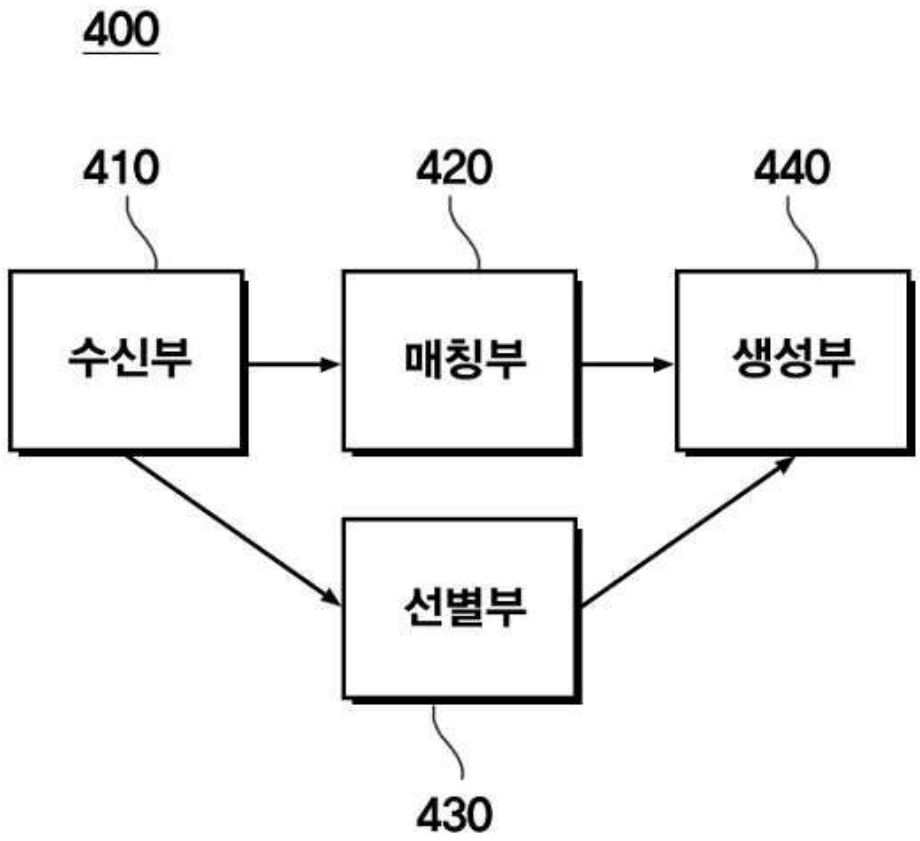
도면2


시작
탐침 유전자 정보를 이용하여 복수의 프로브셋과
복수의 이종상동성 유전자 간의 다대다 매칭 관계를 결정
S210
다대다 매칭 관계를 다대일 매칭 관계로 변환
S220
다대일 매칭 관계를 일대일 매칭 관계로 변환
S230
종료

도면3


시작
실험 변인 정보로부터 복수의 실험샘플에 포함된 실험군의 개수 및 시계열적인 실험의 구성 여부를 판단
S310
판단결과에 따라, 복수의 실험샘플을 이용하는 실험구성을 단조건 실험구성, 다조건 실험구성, 단조건 시계열 실험구성 및 다조건 시계열 실험구성 중의 하나로 결정
S320
결정된 실험구성에 대응되는 적어도 하나의 대조군 및 실험군을 선별
S330
종료

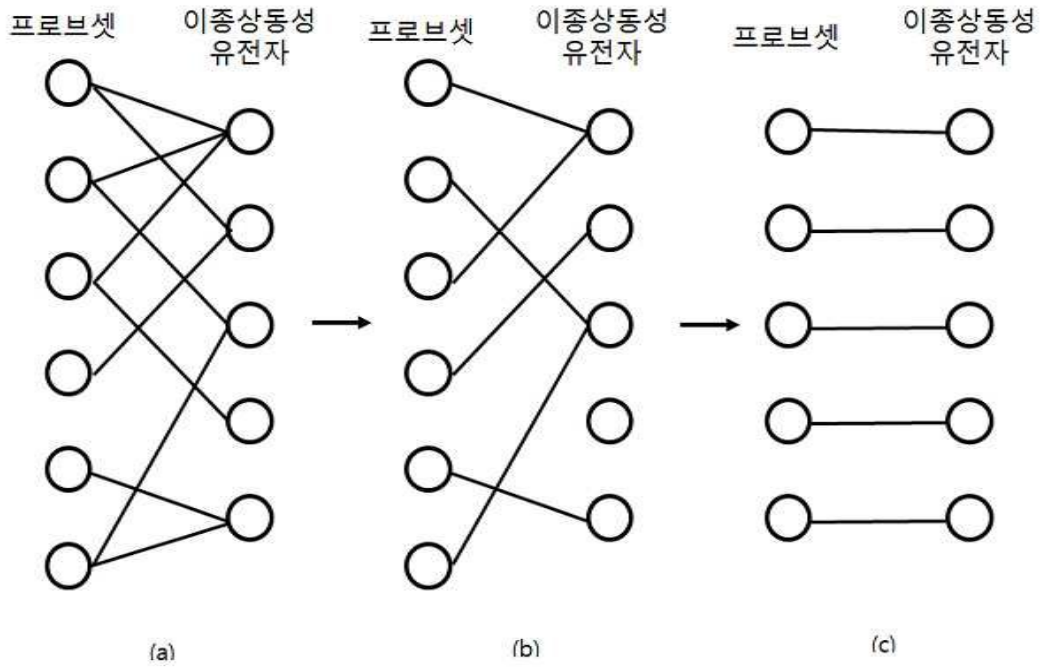
도면4



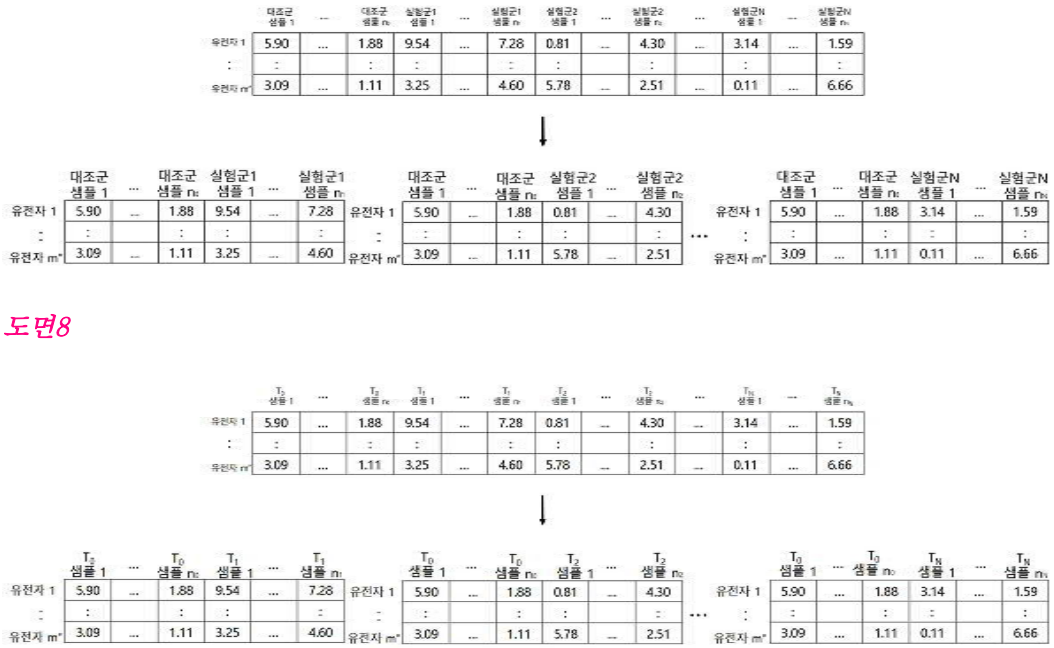
도면5

	샘플 1	...			...			...	샘플 n
프로브셋 1	59.71	...	3.68	744.43	...	155.41	1.75	...	19.69
프로브셋 2	1234.7	...	245.57	missing	...	31.12	1214.6	...	1.04
:	:		:	:		:	:		:
프로브셋 m	8.51	...	2.15	9.51	...	24.25	54.94	...	5.96

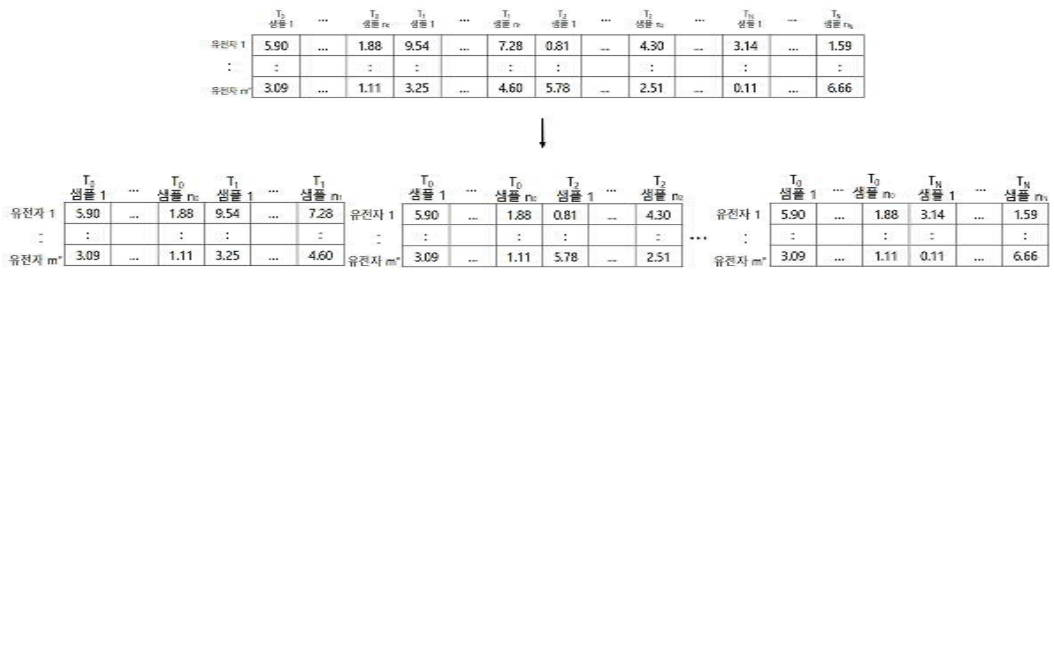
도면6


프로브셋
이종상동성 유전자
프로브셋
이종상동성 유전자
프로브셋
이종상동성 유전자
(a)
(b)
(c)

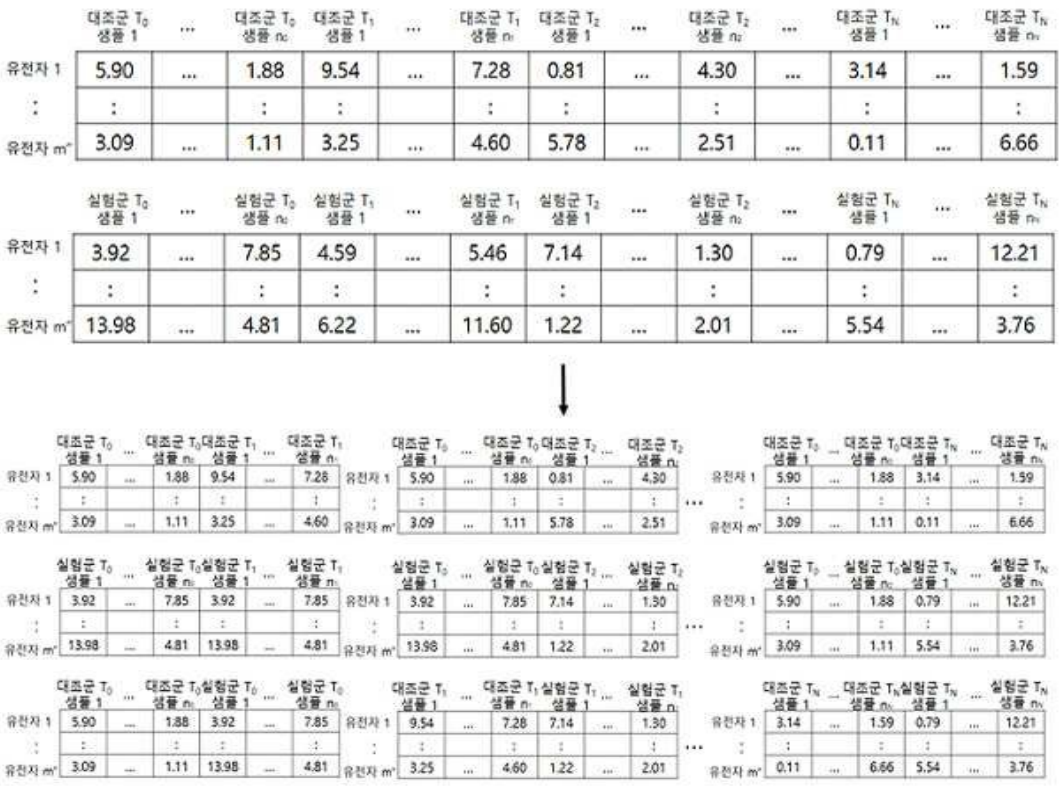
도면7


대조군 샘플 1
대조군 샘플 n0
실험군1 샘플 1
실험군1 샘플 n1
실험군2 샘플 1
실험군2 샘플 n2
실험군N 샘플 1
실험군N 샘플 nN
유전자 1
유전자 m'
5.90 ... 1.88 9.54 ... 7.28 0.81 ... 4.30 ... 3.14 ... 1.59
3.09 ... 1.11 3.25 ... 4.60 5.78 ... 2.51 ... 0.11 ... 6.66

도면8


T0 샘플 1
T0 샘플 n0
T1 샘플 1
T1 샘플 n1
T2 샘플 1
T2 샘플 n2
TN 샘플 1
TN 샘플 nN
유전자 1
유전자 m'
5.90 ... 1.88 9.54 ... 7.28 0.81 ... 4.30 ... 3.14 ... 1.59
3.09 ... 1.11 3.25 ... 4.60 5.78 ... 2.51 ... 0.11 ... 6.66

도면9



도면10

	대조군 샘플 1	...	대조군 샘플 n_1	실험군 샘플 1	...	실험군 샘플 n_2	대조군 평균	대조군 분산	실험군 평균	실험군 분산	평균의 차이
유전자 1	5.90	...	1.88	9.54	...	27.28	3.21	1.20	14.30	2.49	11.09
:	:		:	:		:	:	:	:	:	:
유전자 m'	3.09	...	1.11	3.25	...	74.6	5.47	0.89	4.88	1.39	-0.59

유전자 발현 데이터 행렬 기초 통계값 행렬

도면11

