



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0117598
(43) 공개일자 2022년08월24일

(51) 국제특허분류(Int. Cl.)
G16B 20/20 (2019.01) G06N 20/00 (2019.01)
G06N 3/08 (2006.01) G16B 30/00 (2019.01)
G16B 40/20 (2019.01) G16B 50/00 (2019.01)
(52) CPC특허분류
G16B 20/20 (2019.02)
C12Q 1/6869 (2018.05)
(21) 출원번호 10-2021-0021143
(22) 출원일자 2021년02월17일
심사청구일자 2021년02월17일

(71) 출원인
지니너스 주식회사
서울특별시 송파구 정의로 70, 5층(문정동, 케이디유타워)
충남대학교산학협력단
대전광역시 유성구 대학로 99 (궁동, 충남대학교)
(72) 발명자
안강현
대전광역시 유성구 어은로 57, 116동 205호(어은동, 한빛아파트)
김현재
대전광역시 유성구 진잠로66번길 24 (원내동), 2층
보이첸코 벨리
대전광역시 서구 만년남로 3번길 114, 203 호 (만년동)
(74) 대리인
리엔목특허법인

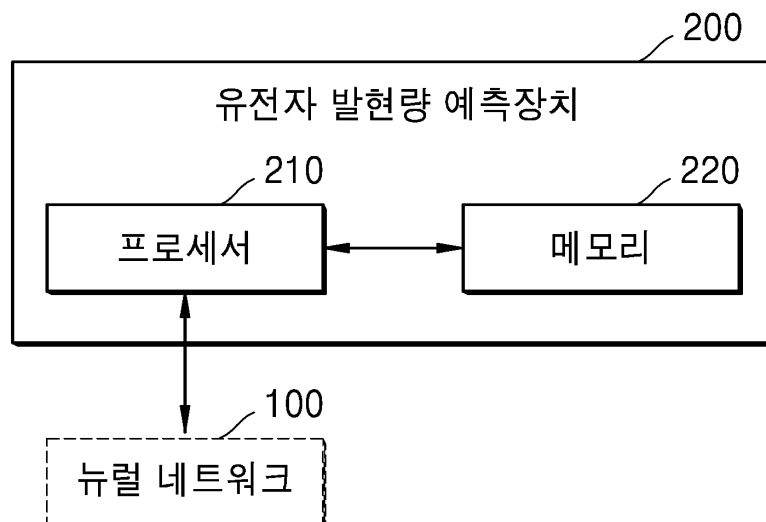
전체 청구항 수 : 총 12 항

(54) 발명의 명칭 유전자 발현량 예측 방법 및 장치

(57) 요약

유전자 발현량 예측 방법은 뉴럴 네트워크를 학습시키기 위한 학습 샘플로부터 제1 SNP 데이터를 추출하는 단계, 제1 SNP 데이터를 입력 데이터로 하고 제1 SNP 데이터에 대응되는 제1 유전자의 발현량을 목표 데이터로 하여 뉴럴 네트워크를 학습시키는 단계, 유전자 발현량을 예측하기 위한 테스트 샘플로부터 제2 SNP 데이터를 추출하는 단계 및 제2 SNP 데이터를 입력 데이터로 하여 학습된 뉴럴 네트워크로부터 제2 SNP 데이터에 대응되는 제2 유전자의 발현량을 획득하는 단계를 포함할 수 있다.

대표도 - 도4



(52) CPC특허분류

G06N 20/00 (2021.08)
G06N 3/08 (2013.01)
G16B 30/00 (2019.02)
G16B 40/20 (2019.02)
G16B 50/00 (2019.02)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711131842
과제번호	2018M3C9A6017315
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	포스트게놈산업육성을위한다부처유전체사업(R&D)(과기정통부)
연구과제명	진행성 대장암 예후 예측 바이오마커 발굴 및 임상유용성 검증
기 여 율	1/2
과제수행기관명	지니너스 주식회사
연구기간	2021.01.01 ~ 2021.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711137915
과제번호	202011B12-01
부처명	다부처
과제관리(전문)기관명	(재단)범부처전주기의료기기연구개발사업단
연구사업명	범부처전주기의료기기연구개발사업(R&D)
연구과제명	전립선암의 맞춤형 치료를 위한 전주기 진단용 액체생검 통합 플랫폼 개발
기 여 율	1/2
과제수행기관명	지니너스 주식회사
연구기간	2021.03.01 ~ 2021.12.31

명세서

청구범위

청구항 1

유전자 발현량 예측 방법에 있어서,

뉴럴 네트워크를 학습시키기 위한 학습 샘플로부터 제1 SNP 데이터를 추출하는 단계;

상기 제1 SNP 데이터를 입력 데이터로 하고 상기 제1 SNP 데이터에 대응되는 제1 유전자의 발현량을 목표 데이터로 하여 상기 뉴럴 네트워크를 학습시키는 단계;

유전자 발현량을 예측하기 위한 테스트 샘플로부터 제2 SNP 데이터를 추출하는 단계; 및

상기 제2 SNP 데이터를 입력 데이터로 하여 상기 학습된 뉴럴 네트워크로부터 상기 제2 SNP 데이터에 대응되는 제2 유전자의 발현량을 획득하는 단계를 포함하는, 방법.

청구항 2

제 1항에 있어서,

상기 제1 SNP 데이터를 추출하는 단계는,

상기 학습 샘플로부터 SNP 염기 및 상기 SNP 염기 주변의 염기들을 포함하는 상기 제1 SNP 데이터를 추출하는 단계를 포함하고,

상기 뉴럴 네트워크를 학습시키는 단계는,

상기 SNP 염기 및 상기 주변 염기들 각각에 웨이트를 적용하는 단계를 포함하는, 방법.

청구항 3

제 2항에 있어서,

상기 웨이트를 적용하는 단계는,

상기 SNP 염기가 상기 제1 유전자의 발현량에 미치는 영향에 기초하여 상기 SNP 염기에 웨이트를 적용하는 단계;

상기 주변 염기들 각각이 상기 제1 유전자의 발현량에 미치는 영향에 기초하여 상기 주변 염기들 각각에 웨이트를 적용하는 단계; 및

상기 주변 염기들 각각이 상기 SNP 염기에 미치는 영향에 기초하여 상기 주변 염기들 각각에 웨이트를 적용하는 단계를 포함하는, 방법.

청구항 4

제 1항에 있어서,

상기 제1 SNP 데이터를 추출하는 단계는,

상기 제1 SNP 데이터를 연속되는 복수의 청크들로 구분하는 단계를 포함하고,

상기 뉴럴 네트워크를 학습시키는 단계는,

상기 복수의 청크들 간의 상호 작용을 고려한 연산을 수행하도록 상기 뉴럴 네트워크를 학습시키는 단계를 포함하는, 방법.

청구항 5

제 4항에 있어서,

상기 제1 SNP 데이터를 연속되는 복수의 청크들로 구분하는 단계는,

상기 제1 유전자에 포함되는 염기쌍의 개수에 비례하도록 상기 청크들의 개수를 결정하는 단계를 포함하는, 방법.

청구항 6

제 1항에 있어서,

상기 제1 SNP 데이터를 추출하는 단계는,

상기 제1 SNP 데이터에 포함된 SNP 영역에 표현된 SNP 염기가 상기 SNP 영역에 표현될 수 있는 염기들 중 몇 번째로 흔하게 표현되는 염기인지에 따라 0, 1 또는 2로 레이블링하는 단계를 포함하고,

상기 뉴럴 네트워크를 학습시키는 단계는,

상기 SNP 염기에 레이블링된 번호에 기초하여 연산을 수행하도록 상기 뉴럴 네트워크를 학습시키는 단계를 포함하는, 방법.

청구항 7

제 1항에 있어서,

상기 제1 SNP 데이터를 추출하는 단계는,

상기 제1 SNP 데이터에 대해 One-Hot 인코딩을 수행하는 단계를 포함하고,

상기 뉴럴 네트워크를 학습시키는 단계는,

상기 One-Hot 인코딩된 제1 SNP 데이터를 입력 데이터로 하여 상기 뉴럴 네트워크를 학습시키는 단계를 포함하는, 방법.

청구항 8

제 1항에 있어서,

상기 뉴럴 네트워크를 학습시키는 단계는,

상기 제1 SNP 데이터의 서로 구분되는 영역들 각각에 개별적인 커널을 적용하고, 스트라이드 없이 연산을 수행하도록 상기 뉴럴 네트워크를 학습시키는 단계를 포함하는, 방법.

청구항 9

제 1항에 있어서,

상기 뉴럴 네트워크는 FCN이며, 상기 제2 유전자는 CCNA2 유전자인, 방법.

청구항 10

제 1항에 있어서,

상기 뉴럴 네트워크는 CNN이며, 상기 제2 유전자는 SERINC5 유전자인, 방법.

청구항 11

유전자 발현량 예측 장치에 있어서,

적어도 하나의 프로그램이 저장된 메모리; 및

상기 적어도 하나의 프로그램을 실행함으로써 유전자 발현량을 예측하는 프로세서를 포함하고,

상기 프로세서는,

뉴럴 네트워크를 학습시키기 위한 학습 샘플로부터 제1 SNP 데이터를 추출하고, 상기 제1 SNP 데이터를 입력 데이터로 하고 상기 제1 SNP 데이터에 대응되는 제1 유전자의 발현량을 목표 데이터로 하여 상기 뉴럴 네트워크를 학습시키고, 유전자 발현량을 예측하기 위한 테스트 샘플로부터 제2 SNP 데이터를 추출하고, 상기 제2 SNP 데이

터를 입력 데이터로 하여 상기 학습된 뉴럴 네트워크로부터 상기 제2 SNP 데이터에 대응되는 제2 유전자의 발현량을 획득하는, 장치

청구항 12

제 1항의 방법을 컴퓨터에서 실행하기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체.

발명의 설명

기술 분야

[0001] 본 개시는 유전자 발현량 예측 방법 및 장치를 제공한다.

배경 기술

[0002] 인간의 DNA 속 기능의 단위인 유전자(gene)는 단백질을 암호화하고 있고, 유전자가 암호화하고 있는 단백질은 세포의 기능을 결정할 수 있다. 세포 안에서 생성되는 유전자는 세포의 기능을 결정하고, 각각의 세포는 다른 유전자를 생성함으로써 다른 기능을 수행할 수 있다. 유전자 발현(gene expression)이란 DNA가 최종 생산물인 단백질을 생성하는 것을 의미하며 유전자 발현으로 인해 생성되는 단백질의 양에 따라 세포의 기능이 결정될 수 있다. 세포의 정상 기능을 위해 생물의 발생과 신진대사 과정에서 유전자 발현은 필요한 부분에서 필요한 만큼만 일어나야 하며, 특정 유전자의 발현량(gene expression level)에 따라 질병의 발병 가능성이 결정될 수 있다. 따라서, 질병의 발병 가능성을 판단하기 위해 유전자가 발현되기 전에 유전자의 발현량을 예측하는 연구가 지속되고 있다.

발명의 내용

해결하려는 과제

[0003] 본 발명이 해결하고자 하는 기술적 과제는, 유전자 발현량 예측 방법 및 장치를 제공하는데 있다. 본 실시예가 이루고자 하는 기술적 과제는 상기된 바와 같은 기술적 과제들로 한정되지 않으며, 이하의 실시예들로부터 또 다른 기술적 과제들이 유추될 수 있다.

과제의 해결 수단

[0004] 상술한 기술적 과제를 달성하기 위한 기술적 수단으로서, 일 측면에 따른 유전자 발현량 예측 방법은 뉴럴 네트워크를 학습시키기 위한 학습 샘플로부터 제1 SNP 데이터를 추출하는 단계; 상기 제1 SNP 데이터를 입력 데이터로 하고 상기 제1 SNP 데이터에 대응되는 제1 유전자의 발현량을 목표 데이터로 하여 상기 뉴럴 네트워크를 학습시키는 단계; 유전자 발현량을 예측하기 위한 테스트 샘플로부터 제2 SNP 데이터를 추출하는 단계; 및 상기 제2 SNP 데이터를 입력 데이터로 하여 상기 학습된 뉴럴 네트워크로부터 상기 제2 SNP 데이터에 대응되는 제2 유전자의 발현량을 획득하는 단계를 포함할 수 있다.

[0005] 다른 측면에 따른 유전자 발현량 예측 장치는 적어도 하나의 프로그램이 저장된 메모리; 및 상기 적어도 하나의 프로그램을 실행함으로써 유전자 발현량을 예측하는 프로세서를 포함하고, 상기 프로세서는, 뉴럴 네트워크를 학습시키기 위한 학습 샘플로부터 제1 SNP 데이터를 추출하고, 상기 제1 SNP 데이터를 입력 데이터로 하고 상기 제1 SNP 데이터에 대응되는 제1 유전자의 발현량을 목표 데이터로 하여 상기 뉴럴 네트워크를 학습시키고, 유전자 발현량을 예측하기 위한 테스트 샘플로부터 제2 SNP 데이터를 추출하고, 상기 제2 SNP 데이터를 입력 데이터로 하여 상기 학습된 뉴럴 네트워크로부터 상기 제2 SNP 데이터에 대응되는 제2 유전자의 발현량을 획득할 수 있다.

[0006] 또 다른 측면에 따른 컴퓨터로 읽을 수 있는 기록매체는 일 측면에 따른 방법을 컴퓨터에서 실행하기 위한 프로그램을 기록한 기록매체를 포함할 수 있다.

발명의 효과

[0007] 본 발명에 따르면, 유전자 발현량 예측 장치는 뉴럴 네트워크를 이용하여 SNP 데이터로부터 유전자의 발현량을 예측할 수 있다. 유전자의 발현량으로부터 질병의 발병 가능성이 예측될 수 있으므로 질병이 예방될 수 있다.

[0008] 본 발명의 효과가 상술한 효과들로 제한되는 것은 아니며, 언급되지 아니한 효과들은 본 명세서 및 첨부된 도면

으로부터 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 명확히 이해될 수 있을 것이다.

도면의 간단한 설명

[0009]

- 도 1은 뉴럴 네트워크 아키텍처의 일 예를 설명하기 위한 도면이다.
- 도 2는 뉴럴 네트워크에서 수행되는 연산을 설명하기 위한 도면이다.
- 도 3은 뉴럴 네트워크에서 입력 피쳐맵 및 출력 피쳐맵의 관계의 일 예를 설명하기 위한 도면이다.
- 도 4는 일 실시예에 따른 유전자 발현량 예측 장치를 나타낸 블록도이다.
- 도 5는 SNP를 설명하기 위한 염기 서열의 일 예를 나타낸 도면이다.
- 도 6은 샘플로부터 SNP 데이터를 출력하는 방법을 설명하기 위한 도면이다.
- 도 7은 일 실시예에 따른 유전자 발현량 예측 장치의 동작 방법을 설명하기 위한 도면이다.
- 도 8은 SNP 데이터에 커널을 적용한 예시들을 도시한 도면이다.
- 도 9는 일 실시예에 따른 뉴럴 네트워크의 연산 과정을 도시한 도면이다.
- 도 10은 다른 실시예에 따른 뉴럴 네트워크의 연산 과정을 도시한 도면이다.
- 도 11은 선형 모델에 의한 유전자 발현량 예측 결과를 도시한 그래프이다.
- 도 12는 일 실시예에 따른 유전자 발현량 예측 결과를 도시한 그래프들이다.
- 도 13은 다른 실시예에 따른 유전자 발현량 예측 결과를 도시한 그래프들이다.
- 도 14는 상위 번호 유전자들에 대한 예측 결과를 도시한 그래프들이다.
- 도 15는 일 실시예에 따른 전자 시스템을 나타내는 블록도이다.
- 도 16은 일 실시예에 따른 유전자 발현량 예측 방법을 나타낸 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0010]

아래에서는 첨부한 도면을 참고하여 본 발명의 실시예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다.

[0011]

실시예들에서 사용되는 용어는 본 발명에서의 기능을 고려하면서 가능한 현재 널리 사용되는 일반적인 용어들을 선택하였으나, 이는 당 분야에 종사하는 기술자의 의도 또는 관례, 새로운 기술의 출현 등에 따라 달라질 수 있다. 또한, 특정한 경우는 출원인이 임의로 선정한 용어도 있으며, 이 경우 해당되는 발명의 설명 부분에서 상세히 그 의미를 기재할 것이다. 따라서 본 발명에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지는 의미와 본 발명의 전반에 걸친 내용을 토대로 정의되어야 한다.

[0012]

명세서 전체에서 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있음을 의미한다. 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.

[0013]

아래에서는 첨부한 도면을 참고하여 본 발명의 실시예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다.

[0014]

또한, 본 명세서에서 사용되는 '제1' 또는 '제2' 등과 같이 서수를 포함하는 용어는 다양한 구성 요소들을 설명하는데 사용할 수 있지만, 구성 요소들은 용어들에 의해 한정되어서는 안된다. 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로만 사용된다.

[0015]

이하에서는 도면을 참조하여 본 발명의 실시예들을 상세히 설명한다.

[0016]

유전자 발현량 예측 장치는 뉴럴 네트워크를 이용하여 유전자의 발현량을 예측할 수 있다. 유전자 발현량 예측 장치가 뉴럴 네트워크를 이용하여 유전자의 발현량을 예측하는 구체적인 방법에 관해서는 도 4 이하를 참조하여

후술하도록 한다. 이에 앞서, 뉴럴 네트워크에 관하여 도 1 내지 도 3을 참조하여 설명하도록 한다.

[0017] 도 1은 뉴럴 네트워크 아키텍처의 일 예를 설명하기 위한 도면이다.

[0018] 도 1에 도시된 뉴럴 네트워크(100)는 딥 뉴럴 네트워크(Deep Neural Network, 이하 'DNN'이라고 함) 아키텍처의 일 예일 수 있다. DNN은 컨볼루션 뉴럴 네트워크(Convolutional Neural Network, CNN), 리커런트 뉴럴 네트워크(Recurrent Neural Network, RNN), Deep Belief Network, Restricted Boltzman Machine, Fully Connected Network, LSTM(Long short-term memory), Attention Mechanism Network 등에 해당할 수 있으나 이에 한정되지 않는다.

[0019] 뉴럴 네트워크(100)는 입력 레이어(Layer 1), 2개의 히든 레이어들(Layer 2 및 Layer 3) 및 출력 레이어(Layer 4)를 포함하는 DNN일 수 있다. 예를 들어, 뉴럴 네트워크(100)가 컨볼루션 뉴럴 네트워크(CNN)를 표현하는 것인 경우, Layer 1 내지 Layer 4는 컨볼루션 뉴럴 네트워크 중 일부의 레이어들에 해당될 수 있으며, 이는 컨볼루션 레이어, 풀링 레이어(pooling layer), 풀리 커넥티드(fully connected) 레이어 등에 해당될 수 있다.

[0020] 뉴럴 네트워크(100)에 포함된 레이어들 각각은 "뉴런(neuron)", "프로세싱 엘리먼트(Processing element: PE)", "유닛(unit)" 또는 이와 유사한 용어들로 알려진, 복수의 인공 노드(artificial node)들을 포함할 수 있다. 예를 들어, 도 1에 도시된 바와 같이, Layer 1은 2개의 노드들, Layer 2는 3개의 노드들을 포함할 수 있다. 다만, 이는 예시에 불과할 뿐 뉴럴 네트워크(100)에 포함된 레이어들 각각은 다양한 개수의 노드들을 포함할 수 있다.

[0021] 뉴럴 네트워크(100)는 수신되는 입력 데이터(예를 들어, I_1 및 I_2)를 기초로 연산을 수행하고, 수행 결과를 기초로 출력 데이터(예를 들어, O_1 및 O_2)를 생성할 수 있다.

[0022] 뉴럴 네트워크(100)에 포함된 레이어들 각각에 포함된 노드들은 서로 연결되어 데이터를 교환할 수 있다. 예를 들어, 하나의 노드는 다른 노드들로부터 데이터를 수신하여 연산할 수 있고, 연산 결과를 또 다른 노드들로 출력할 수 있다.

[0023] 노드들 각각의 입력 및 출력 각각은 입력 액티베이션 및 출력 액티베이션이라고 지칭될 수 있다. 즉, 액티베이션은 한 노드의 출력임과 동시에, 다음 레이어에 포함된 노드들의 입력에 해당되는 파라미터일 수 있다. 한편, 노드들 각각은 이전 레이어에 포함된 노드들로부터 수신된 액티베이션들 및 웨이트 및 바이어스에 기초하여 자신의 액티베이션을 결정할 수 있다. 웨이트는 각 노드에서의 출력 액티베이션을 계산하기 위해 이용되는 파라미터로서, 노드들 간의 연결관계에 할당되는 값일 수 있다.

[0024] 노드들 각각은 입력을 수신하여 출력 액티베이션을 출력하는 연산 유닛(computational unit) 또는 프로세싱 엘리먼트(processing element)에 의해 처리될 수 있고, 노드들 각각의 입력-출력은 매핑될 수 있다. 예를 들어, f 는 액티베이션 함수(activation function)이고, w_{jk}^i 는 (i-1) 번째 레이어에 포함된 k 번째 노드로부터 i 번째 레이어에 포함된 j 번째 노드로의 웨이트며, b_j^i 는 i 번째 레이어에 포함된 j 번째 노드의 바이어스(bias)이고, a_j^i 는 i 번째 레이어의 j 번째 노드의 액티베이션이라고 할 때, 액티베이션 a_j^i 는 다음과 같은 수학적 식 1을 이용하여 계산될 수 있다.

수학적 식 1

$$a_j^i = \sigma\left(\sum_k (w_{jk}^i \times a_k^{i-1}) + b_j^i\right)$$

[0025]

[0026] 도 1에 도시된 바와 같이, 2번째 레이어(Layer 2)의 첫 번째 노드(CH 1)의 액티베이션은 a_1^2 로 표현될 수 있다. 또한, a_1^2 은 수학적 식 1에 따라 $a_1^2 = \sigma(w_{1,1}^2 \times a_1^1 + w_{1,2}^2 \times a_2^1 + b_1^2)$ 의 값을 가질 수 있다. 다만, 앞서 설명한 수학적 식 1은 뉴럴 네트워크(100)에서 데이터를 처리하기 위해 이용되는 액티베이션 및 웨이트 및 바이어스를 설명하기 위한 예시일 뿐, 이에 제한되지 않는다. 액티베이션은 이전 레이어로부터 수신된 액티베이션들의 가중치 합(weighted sum)을 sigmoid 함수나 Rectified Linear Unit(ReLU) 함수 등의 액티베이션 함수에 통과시킴으로써 획득된 값일 수 있다.

- [0027] 한편, 뉴럴 네트워크(100)가 DNN 아키텍처로 구현된 경우 유효한 정보를 처리할 수 있는 많은 레이어들을 포함할 수 있다. 따라서, 뉴럴 네트워크(100)는 싱글 레이어를 포함하는 뉴럴 네트워크보다 복잡한 데이터 집합들을 처리할 수 있다. 한편, 도 1에는 뉴럴 네트워크(100)가 4개의 레이어들을 포함하는 것으로 도시되어 있으나, 이는 예시에 불과할 뿐 뉴럴 네트워크(100)는 더 적거나 많은 레이어들을 포함할 수 있다. 즉, 뉴럴 네트워크(100)는 도 1에 도시된 것과는 다른, 다양한 구조의 레이어들을 포함할 수 있다.
- [0028] 도 2는 뉴럴 네트워크에서 수행되는 연산을 설명하기 위한 도면이다.
- [0029] 뉴럴 네트워크에서 입력 데이터는 웨이트맵(weight map)과의 컨벌루션 연산이 수행되고, 그 결과 피쳐맵들이 출력된다. 웨이트맵은 입력 데이터의 특징을 찾아내기 위한 파라미터들로, 커널(kernel) 또는 필터(filter)라고도 불린다. 이때 생성된 출력 피쳐맵들은 입력 피쳐맵들로서 다시 웨이트맵과의 컨벌루션 연산이 수행되고, 새로운 피쳐맵들이 출력된다. 이와 같은 컨벌루션 연산이 반복적으로 수행된 결과, 최종적으로는 뉴럴 네트워크를 통한 입력 데이터에 대한 연산 결과가 출력될 수 있다.
- [0030] 예를 들어, 도 2의 뉴럴 네트워크에 24x24 크기의 데이터가 입력된 경우, 입력 데이터는 웨이트맵과의 컨벌루션 연산을 통해 20x20 크기를 갖는 4채널의 피쳐맵들로 출력될 수 있다. 또한, 서브샘플링 과정을 통해 20x20 크기를 갖는 4채널의 피쳐맵의 픽셀 값들 중 일부만이 이용되어 10x10 크기를 갖는 4채널의 피쳐맵들이 출력될 수 있다. 서브샘플링 방식으로는 최대-풀링(max-pooling), 평균-풀링(average-pooling) 등의 방식 등이 적용될 수 있다.
- [0031] 이후에도, 10x10 피쳐맵들은 웨이트맵과의 반복적인 컨벌루션 연산 및 서브샘플링 연산을 통해 크기가 줄어들면서, 최종적으로는 글로벌(global)한 특징들이 출력될 수 있다. 뉴럴 네트워크는 여러 레이어들에서 컨벌루션 연산 및 서브샘플링(또는 풀링) 연산을 반복적으로 수행함으로써 입력 데이터로부터 피쳐맵들을 출력하고, 출력된 피쳐맵들이 풀리 커넥티드 레이어에 입력됨으로써 최종적으로 입력 데이터에 대한 연산 결과를 도출할 수 있다.
- [0032] 도 3은 뉴럴 네트워크에서 입력 피쳐맵 및 출력 피쳐맵의 관계의 일 예를 설명하기 위한 도면이다.
- [0033] 도 3을 참조하면, 뉴럴 네트워크의 어느 레이어에서, 제1 피쳐맵(FM1)은 입력 피쳐맵에 해당될 수 있고, 제2 피쳐맵(FM2)은 출력 피쳐맵에 해당될 수 있다. 피쳐맵은 입력 데이터의 다양한 특징들이 표현된 데이터 세트를 의미할 수 있다. 피쳐맵들(FM1, FM2)은 2차원 매트릭스의 엘리먼트들을 갖거나 또는 3차원 매트릭스의 엘리먼트들을 가질 수 있고, 각각의 엘리먼트에는 픽셀 값이 정의될 수 있다. 피쳐맵들(FM1, FM2)은 너비(W)(또는 칼럼이라고 함), 높이(H)(또는 로우라고 함) 및 깊이(D)를 가진다. 이때, 깊이(D)는 채널들의 개수에 해당될 수 있다.
- [0034] 제1 피쳐맵(FM1) 및 웨이트맵(weight map, WM)에 대한 컨벌루션 연산이 수행될 수 있고, 그 결과 제2 피쳐맵(FM2)이 생성될 수 있다. 웨이트맵은 각 엘리먼트에 정의된 웨이트 파라미터로 제1 피쳐맵(FM1)과 컨벌루션 연산을 수행함으로써 제1 피쳐맵(FM1)의 특징들을 필터링한다. 웨이트맵은 제1 피쳐맵(FM1)을 슬라이딩 윈도우 방식으로 스트라이드(stride) 만큼 시프트하면서 제1 피쳐맵(FM1)의 윈도우들(또는 타일)과 컨벌루션 연산을 수행한다. 각 시프트 동안, 웨이트맵에 포함된 웨이트 파라미터들 각각은 제1 피쳐맵(FM1) 내 중첩된 윈도우의 픽셀 값들 각각과 곱해지고 더해질 수 있다. 제1 피쳐맵(FM1)과 웨이트맵이 컨벌루션됨에 따라, 제2 피쳐맵(FM2)의 하나의 채널이 생성될 수 있다. 도 3에는 하나의 웨이트맵만이 도시되었으나, 실제로는 복수의 웨이트맵들이 제1 피쳐맵(FM1)과 각각 컨벌루션되어, 복수의 채널들의 제2 피쳐맵(FM2)이 생성될 수 있다.
- [0035] 한편, 제2 피쳐맵(FM2)은 다음 레이어의 입력 피쳐맵에 해당될 수 있다. 예를 들어, 제2 피쳐맵(FM2)은 풀링(또는 서브샘플링) 레이어의 입력 피쳐맵이 될 수 있다.
- [0036] 도 1 내지 도 3에서는 설명의 편의를 위하여 뉴럴 네트워크의 개략적인 아키텍처에 대해서만 도시되어 있다. 하지만, 뉴럴 네트워크는 도시된 바와 달리, 보다 많거나 적은 개수의 레이어들, 피쳐맵들, 웨이트맵들 등으로 구현될 수 있고, 그 크기들 또한 다양하게 변형될 수 있음을 당해 기술분야의 통상의 기술자라면 이해할 수 있다.
- [0037] 도 4는 일 실시예에 따른 유전자 발현량 예측 장치를 나타낸 블록도이다.
- [0038] 도 4를 참조하면, 유전자 발현량 예측 장치(200)는 프로세서(210) 및 메모리(220)를 포함할 수 있다. 유전자 발현량 예측 장치(200)는 뉴럴 네트워크(100)를 제어할 수 있다.
- [0039] 도 4에 도시된 유전자 발현량 예측 장치(200)에는 본 실시예들과 관련된 구성요소들만이 도시되어 있다. 따라서, 유전자 발현량 예측 장치(200)에는 도 4에 도시된 구성요소들 외에 다른 범용적인 구성요소들이 더 포함될 수 있음은 당업자에게 자명하다.

- [0040] 유전자 발현량 예측 장치(200)는 SNP 데이터를 이용하여 유전자의 발현량을 예측할 수 있다. SNP 데이터는 적어도 하나의 SNP 염기를 포함할 수 있다. 예를 들어, SNP 데이터는 적어도 하나의 SNP 염기 및 주변 염기들을 포함하며, 뉴럴 네트워크(100)에 입력되기에 적합하게 인코딩된 형태의 데이터일 수 있다.
- [0041] 인구집단에서 DNA의 염기서열은 대부분 일치하나 일부(예를 들어, 1%)에서 염기서열의 차이가 있는데, 이러한 염기서열이 발생하는 위치를 단일 핵산염기 다형성(SNP; Single Nucleotide polymorphism)이라 하고, 해당 위치에 표현된 염기를 SNP 염기라 한다. SNP로 인해 유전자 발현에 변이(variation)가 생기고 인간 간에 특성의 차이가 생길 수 있다. 또한, SNP로 인해 발현되는 유전자의 종류로부터 발병될 수 있는 질병의 종류가 결정되고, 유전자의 발현량으로부터 특정 질병의 발병 가능성이 결정될 수 있다. 유전자의 발현량은 유전자의 산물인 RNA 또는 단백질의 양에 대응될 수 있다.
- [0042] SNP 데이터에 포함된 적어도 하나의 SNP 염기로부터 SNP 데이터에 대응되는 유전자의 발현량이 결정될 수 있다. 예를 들어, SNP 데이터는 DNA에서 특정 유전자에 해당하는 영역의 전부 또는 일부를 포함하거나 영역을 초과하여 포함할 수도 있다. 해당 특정 유전자가 SNP 데이터에 대응되는 유전자일 수 있다.
- [0043] 프로세서(210)는 PC(personal computer), 서버 디바이스, 모바일 디바이스, 임베디드 디바이스, IoT(Internet of Things) 디바이스 등의 다양한 종류의 컴퓨팅 디바이스들에 구비된 프로세서(210)에 해당할 수 있다. 예를 들어, 프로세서(210)는 CPU(central processing unit), GPU(graphics processing unit), AP(application processor), NPU(neural processing unit) 등과 같은 프로세서에 해당할 수 있으나, 이에 제한되지 않는다.
- [0044] 프로세서(210)는 다수의 논리 게이트들의 어레이로 구현될 수도 있고, 범용적인 마이크로 프로세서와 이 마이크로 프로세서에서 실행될 수 있는 프로그램이 저장된 메모리(220)의 조합으로 구현될 수도 있다. 또한, 다른 형태의 하드웨어로 구현될 수도 있음을 본 실시예가 속하는 기술분야에서 통상의 지식을 가진 자라면 이해할 수 있다.
- [0045] 프로세서(210)는 프로세서(210)가 구비된 유전자 발현량 예측 장치(200)를 제어하기 위한 전반적인 기능들을 수행하는 역할을 한다. 프로세서(210)는 메모리(220)에 저장된 프로그램들을 실행함으로써, 유전자 발현량 예측 장치(200)를 전반적으로 제어할 수 있다.
- [0046] 메모리(220)는 프로세서(210) 내에서 처리되는 각종 데이터들을 저장하는 하드웨어로서, 예를 들어, 메모리(220)는 샘플, SNP 데이터 및 유전자의 발현량 등을 저장할 수 있다. 또한, 메모리(220)는 프로세서(210)에 의해 구동될 다양한 프로그램들 또는 애플리케이션들 등을 저장할 수 있다.
- [0047] 메모리(220)는 휘발성 메모리(volatile memory) 또는 불휘발성 메모리(nonvolatile memory) 중 적어도 하나를 포함할 수 있다. 불휘발성 메모리는 ROM(Read Only Memory), PROM(Programmable ROM), EPROM(Electrically Programmable ROM), EEPROM(Electrically Erasable and Programmable ROM), 플래시 메모리, PRAM(Phase-change RAM), MRAM(Magnetic RAM), RRAM(Resistive RAM), FeRAM(Ferroelectric RAM) 등을 포함한다. 휘발성 메모리는 DRAM(Dynamic RAM), SRAM(Static RAM), SDRAM(Synchronous DRAM), PRAM, MRAM(Magnetic RAM), RRAM(Resistive RAM) 등을 포함한다. 실시예에 있어서, 메모리(220)는 HDD(Hard Disk Drive), SSD(Solid State Drive), CF(compact flash), SD(secure digital), Micro-SD(micro secure digital), Mini-SD(mini secure digital), xD(extreme digital) 또는 Memory Stick 중 적어도 하나로 구현될 수 있다.
- [0048] 뉴럴 네트워크(100)는 유전자 발현량 예측 장치(200)의 외부에 구현된 별도의 독립적인 장치에 포함될 수 있다. 이 때, 뉴럴 네트워크(100)는 외부의 유전자 발현량 예측 장치(200)에 의해 데이터를 입력 받고, 연산을 수행한 결과를 출력할 수 있다. 다만, 도 4에 도시된 것과 달리 뉴럴 네트워크(100)는 유전자 발현량 예측 장치(200)에 포함될 수도 있다. 즉, 유전자 발현량 예측 장치(200)와 뉴럴 네트워크(100)의 구현 방식은 어느 하나의 실시예에 의해 제한되지 않는다.
- [0049] 프로세서(210)는 샘플로부터 SNP 데이터를 추출할 수 있다. 샘플은 개인(individual)으로부터 수집된 유전자들을 포함하는 것으로서, 한 개인으로부터 하나의 샘플이 수집될 수 있다. 예를 들어, 150명의 개인들 각각에 대해 특정 조직(tissue)에서 샘플을 수집함으로써 총 150개의 샘플들이 준비될 수 있다. 이러한 샘플은 메모리(220)에 저장될 수 있다. 프로세서(210)는 샘플로부터 획득한 데이터를 뉴럴 네트워크(100)에 입력되기에 적합한 형태로 변환 또는 조정함으로써 SNP 데이터를 생성할 수 있다. 프로세서(210)가 샘플로부터 SNP 데이터를 추출하는 구체적인 방법에 대해서는 도 6을 참조하여 후술하도록 한다.
- [0050] 프로세서(210)는 SNP 데이터를 입력 데이터로 하고 SNP 데이터에 대응되는 유전자의 발현량을 목표 데이터(또는

레이블 데이터)로 하여 뉴럴 네트워크(100)를 학습시킬 수 있다. 다시 말해, 프로세서(210)는 임의의 SNP 데이터를 입력했을 때 그에 대응되는 유전자의 발현량이 출력될 수 있도록 뉴럴 네트워크(100)를 학습시킬 수 있다. 목표 데이터인 유전자의 발현량은 기존에 측정된(observed) 데이터로서 해당 SNP 데이터를 제공한 샘플에서 미리 측정된 유전자의 발현량을 나타낼 수 있다. 프로세서(210)가 뉴럴 네트워크(100)를 학습시키는 구체적인 방법에 관해서는 도 7 이하를 참조하여 후술하도록 한다.

[0051] 프로세서(210)는 뉴럴 네트워크(100)의 학습이 완료된 후, SNP 데이터를 입력 데이터로 하여 학습된 뉴럴 네트워크(100)로부터 SNP 데이터에 대응되는 유전자의 발현량을 획득할 수 있다. 이와 같이, 유전자 발현량 예측 장치(200)는 학습된 뉴럴 네트워크(100)에 SNP 데이터를 입력함으로써 유전자의 발현량을 예측할 수 있다. 예측된 유전자의 발현량으로부터 질병의 발병 가능성이 예측되고 해당 질병의 예방이 가능해질 수 있다.

[0052] 도 5는 SNP를 설명하기 위한 염기 서열의 일 예를 나타낸 도면이다.

[0053] 도 5를 참조하면, 샘플 1(Sample 1) 및 샘플 2(Sample 2) 각각의 DNA에 포함되는 염기서열의 일부가 도시된다.

[0054] 샘플 1 및 샘플 2는 각기 다른 개인으로부터 수집된 샘플이다. 다른 개인이라도 DNA의 대부분은 서로 동일하기에 그 염기서열 또한 대부분 일치한다.

[0055] 도 5에 도시된 바와 같이, 샘플 1 및 샘플 2에서 염기서열 대부분은 서로 일치하나, 일부 불일치하는 염기가 존재하는데 이러한 염기가 SNP 염기(510)이며, SNP 염기(510)가 존재할 수 있는 영역은 SNP 영역(511)이다. 주변 염기들(520)은 SNP 염기(510) 주위의 SNP 염기가 아닌 염기들에 해당한다.

[0056] SNP 염기(510) 및 주변 염기들(520)은 SNP 데이터에서 하나의 단위 또는 그룹으로 그룹핑될 수 있다. 각 단위에는 SNP 염기(510)가 하나 또는 복수 개 포함될 수 있다. 도 5에는 주변 염기들(520)이 SNP 염기(510)를 중심으로 좌우 4개씩 표시되었지만 이는 예시에 불과하며, 주변 염기들(520)은 다양한 개수일 수 있고 좌우 개수가 상이할 수 있다.

[0057] 도 6은 샘플로부터 SNP 데이터를 출력하는 방법을 설명하기 위한 도면이다.

[0058] 도 6을 참조하면, 학습 샘플(610)로부터 제1 SNP 데이터(611)가 추출되고, 테스트 샘플(620)로부터 제2 SNP 데이터(621)가 추출될 수 있다.

[0059] 뉴럴 네트워크로부터 유전자 발현량을 획득하기 위해서, 프로세서는 뉴럴 네트워크를 학습시킬 수 있다. 학습 샘플(610)은 뉴럴 네트워크를 학습시키는 과정에서 사용될 수 있다. 테스트 샘플(620)은 학습 과정이 완료된 후, 학습된 뉴럴 네트워크로부터 유전자 발현량을 획득하는 과정에서 사용될 수 있다.

[0060] 학습 샘플(610)은 유전자 발현이 완료된 단계에서 수집된 샘플로서, 이로부터 미리 측정된 유전자 발현량은 뉴럴 네트워크의 학습 과정에서 목표 데이터로 사용될 수 있다. 프로세서는 학습 샘플(610)로부터 제1 SNP 데이터(611)를 추출하고 테스트 샘플(620)로부터 제2 SNP 데이터(621)를 추출할 수 있다. 제1 SNP 데이터(611)는 뉴럴 네트워크의 학습을 위해 입력되는 데이터이며, 제2 SNP 데이터(621)를 학습된 뉴럴 네트워크로부터 유전자 발현량을 획득하기 위해 입력되는 데이터이다.

[0061] 프로세서가 학습 샘플(610)로부터 제1 SNP 데이터(611)를 추출하는 방법 및 테스트 샘플(620)로부터 제2 SNP 데이터(621)를 추출하는 방법은 서로 대응될 수 있다. 따라서, 후술하는 샘플로부터 SNP 데이터를 추출하는 다양한 실시예들은 학습 샘플(610)로부터 제1 SNP 데이터(611)를 추출하는 과정 및 테스트 샘플(620)로부터 제2 SNP 데이터(621)를 추출하는 과정 모두에 적용될 수 있다.

[0062] 일 실시예에서, 프로세서는 샘플로부터 적어도 하나의 SNP 염기 및 주변 염기들을 포함하는 SNP 데이터를 추출할 수 있다. 유전자의 발현에는 SNP 염기 뿐 아니라 SNP 염기 주변의 주변 염기들도 영향을 미치므로, 프로세서는 SNP 염기 뿐 아니라 주변 염기들 또한 포함하여 SNP 데이터를 추출할 수 있다. 프로세서는 SNP 염기 및 주변 염기들에 웨이트를 상이하게 적용할 수 있다. 프로세서가 SNP 데이터에 웨이트를 적용하는 방법에 관해서는 도 7을 참조하여 후술하도록 한다.

[0063] 다른 실시예에서, 프로세서는 SNP 데이터를 연속되는 복수의 청크(chunk)들로 구분할 수 있다. SNP 데이터는 뉴럴 네트워크에서 각 청크가 구분되어 연산될 수 있다. 프로세서는 SNP 데이터에 대응되는 유전자에 포함되는 염기쌍의 개수에 비례하도록 청크들의 개수를 결정할 수 있다. 예를 들어, 유전자에 포함되는 염기쌍의 개수가 많을수록 SNP 데이터를 더 많은 개수의 청크들로 구분할 수 있다. 복수의 청크들이 연산되는 과정에 관해서는 도 9를 참조하여 후술하도록 한다.

- [0064] 또 다른 실시예에서, 프로세서는 SNP 염기가 SNP 영역에 표현될 수 있는 염기들 중 몇 번째로 흔하게 표현되는 염기인지에 따라 0, 1 또는 2로 레이블링(labeling)할 수 있다. 예를 들어, 인구집단에서 특정 SNP 영역에 A(Adenine)가 90% 비율로 표현되고 G(Guanine)가 9%의 비율로 표현되고 T(Thymine)가 1%의 비율로 표현되는 경우, 프로세서는 SNP 염기가 A이면 0으로 레이블링하고 G이면 1로 레이블링하고 T이면 2로 레이블링할 수 있다.
- [0065] 또 다른 실시예에서, 프로세서는 SNP 데이터에 대해 One-Hot 인코딩을 수행할 수 있다. 예를 들어, 프로세서는 염기들 각각에 0 또는 1의 인덱스를 부여함으로써 SNP 데이터에 대해 인코딩을 수행할 수 있다. 인코딩된 SNP 데이터는 뉴럴 네트워크에 입력되기에 적합한 형태일 수 있다.
- [0066] 도 7은 일 실시예에 따른 유전자 발현량 예측 장치의 동작 방법을 설명하기 위한 도면이다.
- [0067] 유전자 발현량 예측 장치의 동작 방법은 뉴럴 네트워크(100)를 학습시키는 과정 및 학습된 뉴럴 네트워크(110)를 이용하여 유전자의 발현량을 획득하는 과정을 포함할 수 있다.
- [0068] 뉴럴 네트워크(100)를 학습시키는 과정에서, 프로세서는 학습 샘플로부터 추출한 제1 SNP 데이터(611)를 뉴럴 네트워크(100)에 입력할 수 있다. 제1 SNP 데이터(611)의 입력에 응답하여 뉴럴 네트워크(100)로부터 예측 데이터(612)가 출력될 수 있다. 예측 데이터(612)는 뉴럴 네트워크(100)의 학습 과정에서 출력되어 목표 데이터와 비교됨으로써 뉴럴 네트워크(100)의 학습에 활용되는 데이터일 수 있다.
- [0069] 프로세서는 SNP 데이터를 입력했을 때 뉴럴 네트워크(100)로부터 유전자의 발현량이 출력되도록 뉴럴 네트워크(100)를 학습시킬 수 있다. 즉, 프로세서는 예측 데이터(612)가 제1 유전자의 발현량(613)과 유사해지는 방향으로 연산하도록 뉴럴 네트워크(100)를 학습시킬 수 있다. 예를 들어, 뉴럴 네트워크(100)에서 예측 데이터(612)와 제1 유전자의 발현량(613)을 비교함으로써 뉴럴 네트워크(100)의 파라미터들이 최적화하는 방법으로 학습 과정이 수행될 수 있다. 여기서, 제1 유전자는 제1 SNP 데이터(611)에 대응되는 유전자이다.
- [0070] 뉴럴 네트워크(100)를 학습시키는 과정이 완료된 후, 프로세서는 테스트 샘플로부터 추출한 제2 SNP 데이터(621)를 학습된 뉴럴 네트워크(110)에 입력할 수 있다. 제2 SNP 데이터(621)의 입력에 응답하여 학습된 뉴럴 네트워크(110)로부터 제2 유전자의 발현량(622)이 출력될 수 있다. 즉, 프로세서는 학습된 뉴럴 네트워크(110)로부터 제2 유전자에 대해 예측된 발현량을 획득할 수 있다. 여기서, 제2 유전자는 제2 SNP 데이터(621)에 대응되는 유전자이다.
- [0071] 한편, 뉴럴 네트워크(100) 학습 과정 및 유전자 발현량 획득 과정에서 뉴럴 네트워크(100)의 연산 방법은 유사하므로 두 과정 모두에서 후술하는 실시예들이 적용될 수 있다.
- [0072] 일 실시예에서, 프로세서는 SNP 데이터에 포함되는 SNP 염기 및 주변 염기들에 웨이트를 상이하게 적용할 수 있다. 프로세서는 SNP 염기가 유전자의 발현량에 미치는 영향에 기초하여 SNP 염기에 웨이트를 적용할 수 있다. 또한, 프로세서는 주변 염기들 각각이 유전자의 발현량에 미치는 영향 및 SNP 염기에 미치는 영향에 기초하여 주변 염기들 각각에 웨이트를 적용할 수 있다.
- [0073] 예를 들어, 프로세서는 학습 과정에서 SNP 데이터가 목표 데이터(유전자 발현량)에 이르는데 각 염기가 미친 영향을 연산하고, 그 영향이 클수록 해당 염기에 웨이트를 크게 적용하도록 뉴럴 네트워크(100)를 제어할 수 있다. 또한, 프로세서는 학습 과정에서 SNP 데이터가 목표 데이터(유전자 발현량)에 이르는데 SNP 염기가 미친 영향에 대해 주변 염기들 각각이 기여한 정도를 연산하고, 그 기여가 클수록 해당 주변 염기에 웨이트를 크게 적용하도록 뉴럴 네트워크(100)를 제어할 수 있다.
- [0074] 다른 실시예에서, 프로세서는 SNP 데이터에 포함되는 복수의 청크들 간의 상호 작용을 고려한 연산을 수행하도록 뉴럴 네트워크(100)를 제어할 수 있다. 복수의 청크들 간의 상호 작용을 고려한 연산에 관해서는 도 9를 참조하여 구체적으로 후술하도록 한다.
- [0075] 또 다른 실시예에서, 프로세서는 적어도 하나의 SNP 염기가 0, 1 또는 2로 레이블링 된 SNP 데이터에 대해 레이블링된 번호에 기초하여 연산을 수행하도록 뉴럴 네트워크(100)를 제어할 수 있다. 또는, 프로세서는 One-Hot 인코딩된 SNP 데이터를 입력 데이터로 하여 뉴럴 네트워크(100)를 학습시키거나 학습된 뉴럴 네트워크(110)로부터 유전자 발현량을 출력시킬 수 있다.
- [0076] 도 8은 SNP 데이터에 커널을 적용한 예시들을 도시한 도면이다.
- [0077] 도 8의 (a)에 따른 실시예에서, 프로세서는 SNP 데이터(810)에 커널(또는 필터)(820) 및 스트라이드를 적용할 수 있다. 프로세서는 커널(820)을 스트라이드만큼 시프트하면서 SNP 데이터(810) 전체에 대해 연산이 수행되도록

록 뉴럴 네트워크를 제어할 수 있다.

- [0078] 도 8의 (b)에 따른 실시예에서, 프로세서는 SNP 데이터(810)의 서로 구분되는 영역들 각각에 개별적인 커널(830)을 적용할 수 있다. 프로세서는 영역들 각각에 개별적인 커널(830)을 적용함으로써 뉴럴 네트워크에서 각 영역 별로, 중복되는 영역 없이 연산이 수행되도록 할 수 있다. 이 경우, 커널(830)이 시프트하지 않으므로 스트라이드가 존재하지 않는다. 커널(830)의 크기 및 SNP 데이터(810)의 각 영역의 크기는 서로 대응되며, 프로세서는 SNP 데이터(810)의 크기, SNP 데이터(810)에 포함되는 염기의 수, 뉴럴 네트워크의 레이어들의 수 및 구조 등에 기초하여 커널(830)의 크기 및 SNP 데이터(810)의 각 영역의 크기를 결정할 수 있다. SNP 데이터(810)에 개별적인 커널(830)이 적용되어 커널(830)의 시프트 및 중복 연산되는 영역 없이 연산이 수행됨에 따라, SNP 염기의 위치 정보가 효과적으로 고려될 수 있다.
- [0079] 도 9는 일 실시예에 따른 뉴럴 네트워크의 연산 과정을 도시한 도면이다.
- [0080] 도 9의 실시예에 따른 뉴럴 네트워크(100)는 LSTM에 해당할 수 있다. 뉴럴 네트워크(100)는 셀(cell)(920)을 복수 개 포함하며 셀(920)은 각각 작은 단위의 뉴럴 네트워크(100)에 해당할 수 있다. SNP 데이터(810)는 청크(910)를 복수 개 포함하고, 각 청크(910)는 각 셀(920)에 입력될 수 있다. 복수의 셀들은 서로 양방향으로(bidirectional) 연결되어 있으므로 뉴럴 네트워크(100)에서 복수의 청크들 간의 상호 작용이 고려되어 연산될 수 있다. 뉴럴 네트워크(100)가 각 셀(920)에서 출력된 데이터가 모두 종합하여 예측 데이터 또는 유전자 발현량(930)을 출력할 수 있다.
- [0081] 도 10은 다른 실시예에 따른 뉴럴 네트워크의 연산 과정을 도시한 도면이다.
- [0082] 도 10의 실시예에 따른 뉴럴 네트워크(100)는 Attention Mechanism Network에 해당할 수 있다.
- [0083] 뉴럴 네트워크(100)는 SNP 데이터(810)의 청크(1010)들 각각에서 Q(query), K(key) 및 V(value)를 추출하고, 추출된 요소들에 Attention 함수를 적용함으로써 Attention Value Matrix(1030)를 출력할 수 있다. Attention 함수는 Softmax 함수(1020)를 포함할 수 있다. 뉴럴 네트워크(100)는 Attention Value Matrix(1030)에 대해 Fully Connected Network 연산을 수행함으로써 예측 데이터 또는 유전자 발현량(1040)을 출력할 수 있다.
- [0084] 뉴럴 네트워크(100)에서 각 청크(1010)에 포함되는 요소들 모두를 고려한 함수 연산이 수행되므로 SNP 데이터(810)에서 복수의 청크들 간의 상호작용이 고려될 수 있다. 즉, 상이한 위치에 있는 SNP 염기들 간의 상호작용이 고려될 수 있다.
- [0085] 도 11은 선형 모델에 의한 유전자 발현량 예측 결과를 도시한 그래프이다.
- [0086] 도 11에는, 비선형 모델인 뉴럴 네트워크를 이용하여 유전자 발현량을 예측하는 유전자 발현량 예측 장치와의 비교를 위해 종래의 선형 모델로 유전자 발현량을 예측한 결과를 나타내는 그래프가 도시된다.
- [0087] 복수의 샘플들로부터 9916개의 유전자들에 대한 발현량을 예측하는 실험이 진행되었다. 그래프의 세로축은 피어슨 상관 계수(R; Pearson correlation coefficient 또는 Pearson's r)의 제곱을 나타내며, 가로축은 R^2 이 작은 것부터 큰 순서대로 나열하여 매긴 유전자 번호를 나타낸다.
- [0088] R^2 은 0와 1 사이의 값을 가지며, 1은 완벽한 선형 상관 관계, 0은 선형 상관 관계 없음을 의미한다. 특정 유전자에 대한 R^2 은 예측된 유전자 발현량을 가로축, 측정된 유전자 발현량을 세로축으로 하여 특정 유전자에 관해 복수의 샘플들로부터 획득된 좌표들을 종합하여 $y=x$ 축에 얼마나 가깝게 형성되는지 여부로 계산된다. 예측된 유전자 발현량과 측정된 유전자 발현량이 일치할수록 정확하게 예측된 것이므로 좌표들이 $y=x$ 축에 가깝게 형성될수록 R^2 이 1에 가깝게 나타난다.
- [0089] 도 11의 그래프를 참조하면, 9916번에 가까운 번호대의 유전자들은 R^2 이 크게 나타나 예측의 정확도가 높지만, 하위 번호 200개의 유전자들은 0에 가까운 R^2 이 나타난 바, 예측이 무의미해지는 결과가 도출되었다. 유전자 발현량 예측 장치가 해당 하위 번호 200개의 유전자들에 대해 발현량을 예측한 결과를 도 12 및 도 13을 참조하여 후술한다.
- [0090] 도 12는 일 실시예에 따른 유전자 발현량 예측 결과를 도시한 그래프들이다.
- [0091] 도 12를 참조하면, 하위 번호 200개의 유전자들에 대한 실험 결과를 나타낸 그래프(1210) 및 R^2 이 가장 높게 나

타난 유전자(1211)에 대한 그래프(1220)가 도시된다. 하위 번호 200개의 유전자들에 대한 그래프(1210)는 유전자 발현량 예측 장치에 의한 실험 결과 및 선형 모델에 의한 예측 결과를 나타낸다. 도 12의 실시예에 따른 뉴럴 네트워크는 Fully Connected Network에 해당할 수 있다.

[0092] 비교 대상인 도 11의 선형 모델에 의한 예측 결과는 200개의 유전자들에 대해 평균 R^2 이 4.9×10^{-6} 로 나타난 반면, 유전자 발현량 예측 장치에 의한 예측 결과는 평균 R^2 이 0.017로 나타나 유의미한 결과가 도출되었다. 또한, 일부 유전자(1211)에 대해서는 R^2 이 0.3에 가까운 값이 나타나는 현저한 효과가 도출되었다.

[0093] R^2 이 0.31327로 나타난 CCNA2 유전자(1211)에 대한 그래프(1220)를 참조하면, 하나의 좌표는 하나의 샘플로부터 획득된 좌표에 해당하고 좌표들이 선형성을 나타내는 것이 도시된다.

[0094] 도 13은 다른 실시예에 따른 유전자 발현량 예측 결과를 도시한 그래프들이다.

[0095] 도 13을 참조하면, 하위 번호 200개의 유전자들에 대한 실험 결과를 나타낸 그래프(1310) 및 R^2 이 가장 높게 나타난 유전자(1311)에 대한 그래프(1320)가 도시된다. 하위 번호 200개의 유전자들에 대한 그래프(1310)는 유전자 발현량 예측 장치에 의한 실험 결과 및 선형 모델에 의한 예측 결과를 나타낸다. 도 13의 실시예에 따른 뉴럴 네트워크는 Convolutional Neural Network에 해당할 수 있다.

[0096] 비교 대상인 도 11의 선형 모델에 의한 예측 결과는 200개의 유전자들에 대해 평균 R^2 이 4.9×10^{-6} 로 나타난 반면, 유전자 발현량 예측 장치에 의한 예측 결과는 평균 R^2 이 0.022로 나타나 유의미한 결과가 도출되었다. 또한, 일부 유전자(1311)에 대해서는 R^2 이 0.3에 가까운 값이 나타나는 현저한 효과가 도출되었다.

[0097] R^2 이 0.29626으로 나타난 SERINC5 유전자(1311)에 대한 그래프(1320)를 참조하면, 하나의 좌표는 하나의 샘플로부터 획득된 좌표에 해당하고 좌표들이 선형성을 나타내는 것이 도시된다.

[0098] 도 14는 상위 번호 유전자들에 대한 예측 결과를 도시한 그래프들이다.

[0099] 도 14를 참조하면, 도 11의 그래프(1410), 상위 번호 10개의 유전자들에 대한 실험 결과를 나타낸 그래프(1420) 및 R^2 이 가장 높게 나타난 유전자(1421)에 대한 그래프(1430)가 도시된다. 상위 번호 10개의 유전자들에 대한 그래프(1420)는 유전자 발현량 예측 장치에 의한 실험 결과 및 선형 모델에 의한 예측 결과를 나타낸다. 도 14의 실시예에 따른 뉴럴 네트워크는 Fully Connected Network에 해당할 수 있다.

[0100] 상위 번호 10개의 유전자들에 대해 유전자 발현량 예측 장치에 의한 예측 결과는 R^2 이 0.5560, 선형 모델에 의한 예측 결과는 R^2 이 0.5568로 유사한 결과가 도출되었다. 즉, 유전자 발현량 예측 장치는 하위 번호대의 유전자들에 대해서는 선형 모델 대비 현저한 효과를 나타내는 동시에, 선형 모델에서 R^2 이 상당히 높게 나타나는 상위 번호대의 유전자들에 대해서도 선형 모델 대비 유사한 효과를 나타낸다.

[0101] R^2 이 0.68857로 나타난 FAM118A 유전자(1421)에 대한 그래프(1430)를 참조하면, 하나의 좌표는 하나의 샘플로부터 획득된 좌표에 해당하고 좌표들이 선형성을 나타내는 것이 도시된다.

[0102] 도 15는 일 실시예에 따른 전자 시스템을 나타내는 블록도이다.

[0103] 도 15를 참고하면, 전자 시스템(1500)은 뉴럴 네트워크를 기초로 입력 데이터를 실시간으로 분석하여 유효한 정보를 추출하고, 추출된 정보를 기초로 상황 판단을 하거나 또는 전자 시스템(1500)이 탑재되는 전자 디바이스의 구성들을 제어할 수 있다. 예컨대 전자 시스템(1500)은 드론(drone), 첨단 운전자 보조 시스템(Advanced Drivers Assistance System; ADAS) 등과 같은 로봇 장치, 스마트 TV, 스마트폰, 의료 디바이스, 모바일 디바이스, 영상 표시 디바이스, 계측 디바이스, IoT 디바이스 등에 적용될 수 있으며, 이 외에도 다양한 종류의 전자 디바이스들 중 적어도 하나에 탑재될 수 있다.

[0104] 유전자 발현량 예측 장치는 CPU(1510), RAM(1520), 메모리(1540), 센서 모듈(1550) 및 통신 모듈(1560)을 포함하고, 뉴럴 네트워크 장치(1530)는 뉴럴 네트워크를 포함할 수 있다. 전자 시스템(1500)은 유전자 발현량 예측 장치 및 뉴럴 네트워크 장치를 포함할 수 있다. 도 15의 CPU는 도 4의 프로세서에 대응되고 도 15의 RAM 및 메모리는 도 4의 메모리에 대응될 수 있다. 따라서, 중복되는 설명은 생략한다.

- [0105] 유전자 발현량 예측 장치는 입출력 모듈, 보안 모듈, 전력 제어 장치 등을 더 포함할 수 있다. 전자 시스템(1500)의 하드웨어 구성들 중 일부는 적어도 하나의 반도체 칩에 탑재될 수 있다. 뉴럴 네트워크 장치(1530)는 뉴럴 네트워크 전용 하드웨어 가속기를 포함하는 장치일 수 있다.
- [0106] CPU(1510)는 전자 시스템(1500)의 전반적인 동작을 제어한다. CPU(1510)는 하나의 프로세서 코어(Single Core)를 포함하거나, 복수의 프로세서 코어들(Multi-Core)을 포함할 수 있다. CPU(1510)는 메모리(1540)에 저장된 프로그램들 및/또는 데이터를 처리 또는 실행할 수 있다. 일 실시예에 있어서, CPU(1510)는 메모리(1540)에 저장된 프로그램들을 실행함으로써, 뉴럴 네트워크 장치(1530)의 기능을 제어할 수 있다. CPU(1510)는 CPU, GPU, AP 등으로 구현될 수 있다.
- [0107] RAM(1520)은 프로그램들, 데이터, 또는 명령들(instructions)을 일시적으로 저장할 수 있다. 예컨대 메모리(1540)에 저장된 프로그램들 및/또는 데이터는 CPU(1510)의 제어 또는 부팅 코드에 따라 RAM(1520)에 일시적으로 저장될 수 있다. RAM(1520)은 DRAM 또는 SRAM 등의 메모리로 구현될 수 있다.
- [0108] 뉴럴 네트워크 장치(1530)는 수신되는 입력 데이터를 기초로 뉴럴 네트워크의 연산을 수행하고, 수행 결과를 기초로 정보 신호를 생성할 수 있다. 뉴럴 네트워크 장치(1530)는 뉴럴 네트워크를 이용하여 처리를 수행하는 하드웨어로서, 뉴럴 네트워크 전용 하드웨어 가속기에 해당될 수 있다.
- [0109] 정보 신호는 음성 인식 신호, 사물 인식 신호, 영상 인식 신호, 생체 정보 인식 신호 등과 같은 다양한 종류의 인식 신호 중 하나를 포함할 수 있다. 예를 들어, 뉴럴 네트워크 장치(1530)는 SNP 데이터를 입력 데이터로서 수신하고, SNP 데이터에 대응되는 유전자와 관련된 인식 신호를 생성할 수 있다. 그러나, 이에 제한되는 것은 아니며, 전자 시스템(1500)이 탑재된 전자 장치의 종류 또는 기능에 따라 뉴럴 네트워크 장치(1530)는 다양한 종류의 입력 데이터를 수신할 수 있고, 입력 데이터에 따른 인식 신호를 생성할 수 있다.
- [0110] 메모리(1540)는 데이터를 저장하기 위한 저장 장소로서, OS(Operating System), 각종 프로그램들, 및 각종 데이터를 저장할 수 있다. 실시예에 있어서, 메모리(1540)는 뉴럴 네트워크 장치(1530)의 연산 수행 과정에서 생성되는 중간 결과들, 예컨대 출력 피쳐 맵을 출력 피쳐 리스트 또는 출력 피쳐 매트릭스 형태로 저장할 수 있다. 실시예에 있어서, 메모리(1540)에는 압축된 출력 피쳐 맵이 저장될 수 있다. 또한, 메모리(1540)는 뉴럴 네트워크 장치(1530)에서 이용되는 양자화된 뉴럴 네트워크 데이터, 예컨대, 파라미터들, 웨이트 맵 또는 웨이트 리스트를 저장할 수 있다. 메모리(1540)는 휘발성 메모리 또는 비휘발성 메모리 중 적어도 하나를 포함할 수 있다.
- [0111] 센서 모듈(1550)은 전자 시스템(1500)이 탑재되는 전자 장치 주변의 정보를 수집할 수 있다. 센서 모듈(1550)은 전자 장치의 외부로부터 신호(예컨대 영상 신호, 음성 신호, 자기 신호, 생체 신호, 터치 신호 등)를 센싱 또는 수신하고, 센싱 또는 수신된 신호를 데이터로 변환할 수 있다. 이를 위해, 센서 모듈(1550)은 센싱 장치, 예컨대 마이크, 촬상 장치, 이미지 센서, 라이다(LIDAR; light detection and ranging) 센서, 초음파 센서, 적외선 센서, 바이오 센서, 및 터치 센서 등 다양한 종류의 센싱 장치 중 적어도 하나를 포함할 수 있다. 센서 모듈(1550)은 변환된 데이터를 뉴럴 네트워크 장치(1530)에 입력 데이터로서 제공할 수 있다. 센서 모듈(1550)은 다양한 종류의 데이터를 뉴럴 네트워크 장치(1530)에 제공할 수 있다.
- [0112] 통신 모듈(1560)은 외부 장치와 통신할 수 있는 다양한 유선 또는 무선 인터페이스를 구비할 수 있다. 예컨대 통신 모듈(1560)은 유선 근거리통신망(Local Area Network; LAN), Wi-fi(Wireless Fidelity)와 같은 무선 근거리 통신망(Wireless Local Area Network; WLAN), 블루투스(Bluetooth)와 같은 무선 개인 통신망(Wireless Personal Area Network; WPAN), 무선 USB(Wireless Universal Serial Bus), Zigbee, NFC(Near Field Communication), RFID(Radio-frequency identification), PLC(Power Line communication), 또는 3G(3rd Generation), 4G(4th Generation), LTE(Long Term Evolution) 등 이동 통신망(mobile cellular network)에 접속 가능한 통신 인터페이스 등을 포함할 수 있다.
- [0113] 도 16은 일 실시예에 따른 유전자 발현량 예측 방법을 나타낸 흐름도이다.
- [0114] 도 16을 참조하면, 유전자 발현량 예측 방법은 도 4에 도시된 유전자 발현량 예측 장치에서 시계열적으로 처리되는 단계들로 구성된다. 따라서, 이하에서 생략된 내용이라 하더라도 유전자 발현량 예측 방법에 관하여 전술된 내용은 도 16의 방법에도 적용됨을 알 수 있다.
- [0115] 한편, 단계 1610에서 SNP 데이터를 추출하는 방법은 단계 1630에도 적용될 수 있고, 단계 1620에서 뉴럴 네트워크의 연산 방법은 단계 1640에도 적용될 수 있다. 따라서, 단계 1630 및 단계 1640에서 중복되는 설명은 생략한다.

- [0116] 단계 1610에서, 유전자 발현량 예측 장치는 뉴럴 네트워크를 학습시키기 위한 학습 샘플로부터 제1 SNP 데이터를 추출할 수 있다.
- [0117] 일 실시예에서, 유전자 발현량 예측 장치는 학습 샘플로부터 SNP 염기 및 SNP 염기 주변의 염기들을 포함하는 제1 SNP 데이터를 추출할 수 있다.
- [0118] 다른 실시예에서, 유전자 발현량 예측 장치는 제1 SNP 데이터를 연속되는 복수의 청크들로 구분할 수 있다.
- [0119] 또 다른 실시예에서, 유전자 발현량 예측 장치는 제1 SNP 데이터에 포함된 SNP 영역에 표현된 SNP 염기가 SNP 영역에 표현될 수 있는 염기들 중 몇 번째로 흔하게 표현되는 염기인지에 따라 0, 1 또는 2로 레이블링할 수 있다.
- [0120] 또 다른 실시예에서, 유전자 발현량 예측 장치는 제1 SNP 데이터에 대해 One-Hot 인코딩을 수행할 수 있다.
- [0121] 단계 1620에서, 유전자 발현량 예측 장치는 제1 SNP 데이터를 입력 데이터로 하고 제1 SNP 데이터에 대응되는 제1 유전자의 발현량을 목표 데이터로 하여 뉴럴 네트워크를 학습시킬 수 있다.
- [0122] 일 실시예에서, 유전자 발현량 예측 장치는 SNP 염기 및 주변 염기들 각각에 웨이트를 적용할 수 있다. 유전자 발현량 예측 장치는 SNP 염기가 제1 유전자의 발현량에 미치는 영향에 기초하여 SNP 염기에 웨이트를 적용하고, 주변 염기들 각각이 제1 유전자의 발현량에 미치는 영향에 기초하여 주변 염기들 각각에 웨이트를 적용하고, 주변 염기들 각각이 SNP 염기에 미치는 영향에 기초하여 주변 염기들 각각에 웨이트를 적용할 수 있다.
- [0123] 다른 실시예에서, 유전자 발현량 예측 장치는 복수의 청크들 간의 상호 작용을 고려한 연산을 수행하도록 뉴럴 네트워크를 학습시킬 수 있다. 유전자 발현량 예측 장치는 제1 유전자에 포함되는 염기쌍의 개수에 비례하도록 청크들의 개수를 결정할 수 있다.
- [0124] 또 다른 실시예에서, 유전자 발현량 예측 장치는 SNP 염기에 레이블링된 번호에 기초하여 연산을 수행하도록 뉴럴 네트워크를 학습시킬 수 있다.
- [0125] 또 다른 실시예에서, 유전자 발현량 예측 장치는 One-Hot 인코딩된 제1 SNP 데이터를 입력 데이터로 하여 뉴럴 네트워크를 학습시킬 수 있다.
- [0126] 또 다른 실시예에서, 유전자 발현량 예측 장치는 제1 SNP 데이터의 서로 구분되는 영역들 각각에 개별적인 커널을 적용하고, 스트라이드 없이 연산을 수행하도록 뉴럴 네트워크를 학습시킬 수 있다.
- [0127] 단계 1630에서, 유전자 발현량 예측 장치는 유전자 발현량을 예측하기 위한 테스트 샘플로부터 제2 SNP 데이터를 추출할 수 있다.
- [0128] 단계 1640에서, 유전자 발현량 예측 장치는 제2 SNP 데이터를 입력 데이터로 하여 학습된 뉴럴 네트워크로부터 제2 SNP 데이터에 대응되는 제2 유전자의 발현량을 획득할 수 있다.
- [0129] 일 실시예에서, 뉴럴 네트워크는 FCN이며, 제2 유전자는 CCNA2 유전자일 수 있다. 다른 실시예에서, 뉴럴 네트워크는 CNN이며, 제2 유전자는 SERINC5 유전자일 수 있다.
- [0130] 본 실시예들은 전자 디바이스에 의해 실행 가능한 명령어 및 데이터를 저장하는 전자 디바이스로 읽을 수 있는 기록매체에 저장된 어플리케이션의 형태로 구현될 수 있다. 상기 명령어는 프로그램 코드의 형태로 저장될 수 있으며, 프로세서에 의해 실행되었을 때, 소정의 프로그램 모듈을 생성하여 소정의 동작을 수행할 수 있다. 또한, 상기 명령어는 프로세서에 의해 실행되었을 때, 개시된 실시예들의 소정의 동작들을 수행할 수 있다.
- [0131] 본 실시예들은 컴퓨터에 의해 실행되는 프로그램 모듈과 같은 컴퓨터에 의해 실행 가능한 명령어를 포함하는 기록 매체의 형태로도 구현될 수 있다. 컴퓨터 판독 가능 매체는 컴퓨터에 의해 액세스될 수 있는 임의의 가용 매체일 수 있고, 휘발성 및 비휘발성 매체, 분리형 및 비분리형 매체를 모두 포함한다. 또한, 컴퓨터 판독가능 매체는 컴퓨터 저장 매체 및 통신 매체를 모두 포함할 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보의 저장을 위한 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 모두 포함한다. 통신 매체는 전형적으로 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈과 같은 변조된 데이터 신호의 기타 데이터, 또는 기타 전송 매커니즘을 포함하며, 임의의 정보 전달 매체를 포함한다.
- [0132] 전술한 본 명세서의 설명은 예시를 위한 것이며, 본 명세서의 내용이 속하는 기술분야의 통상의 지식을 가진 자는 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는

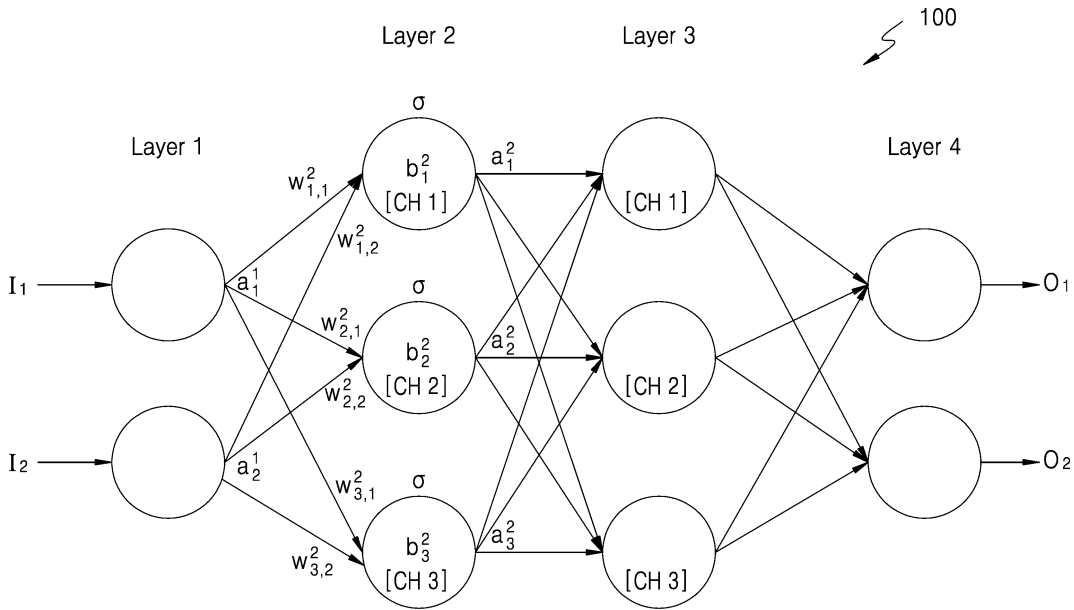
것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 예를 들어, 단일형으로 설명되어 있는 각 구성 요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명되어 있는 구성 요소들도 결합된 형태로 실시될 수 있다.

[0133]

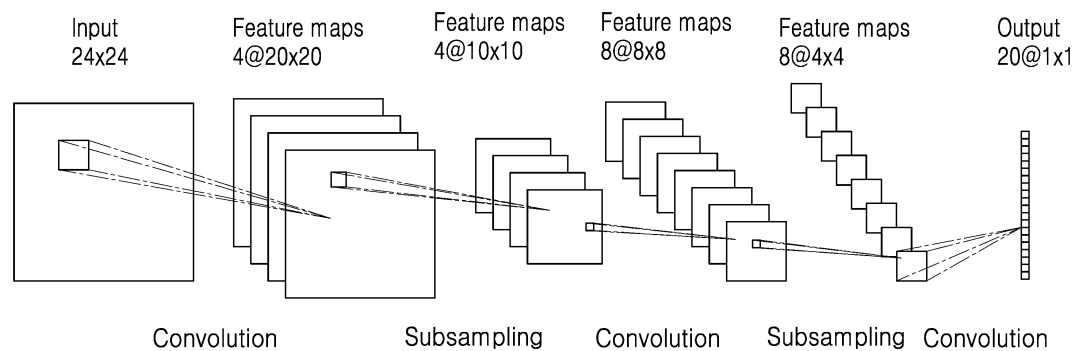
본 실시예의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 포함되는 것으로 해석되어야 한다.

도면

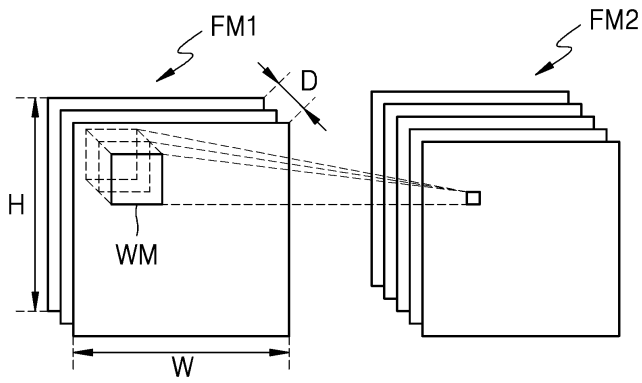
도면1



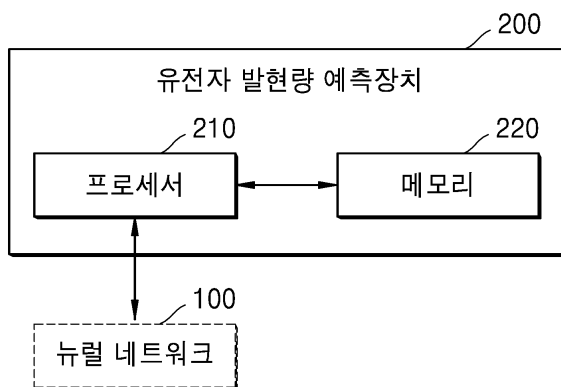
도면2



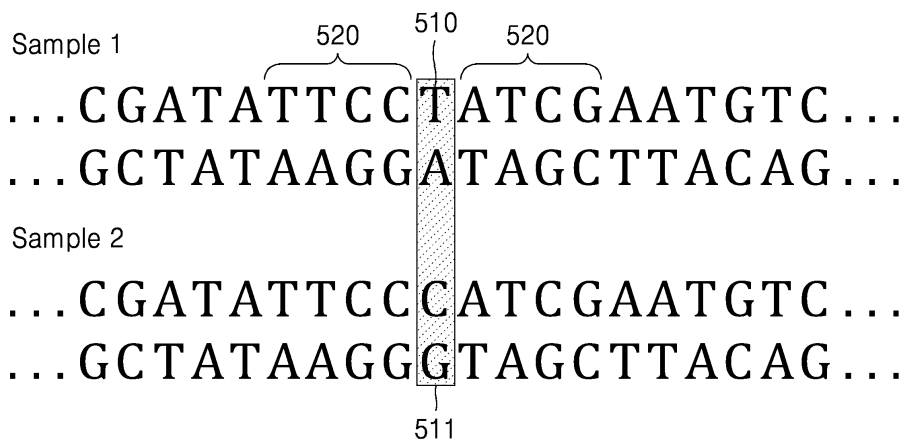
도면3



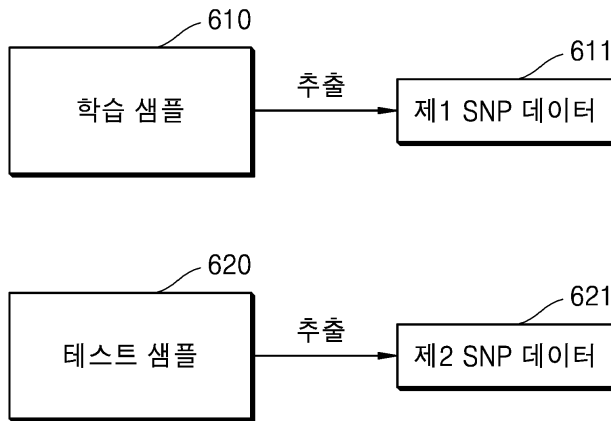
도면4



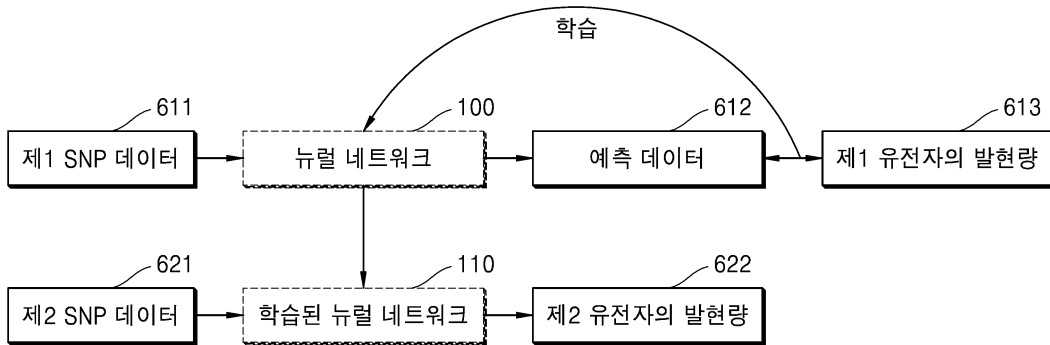
도면5



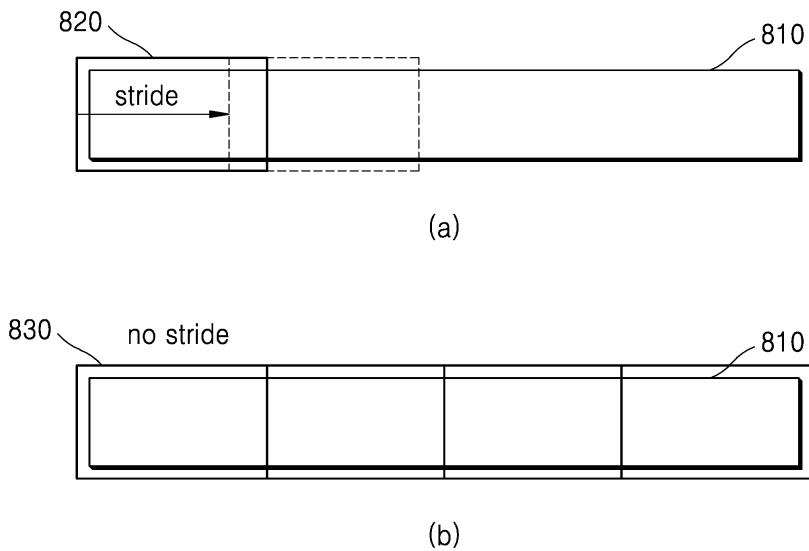
도면6



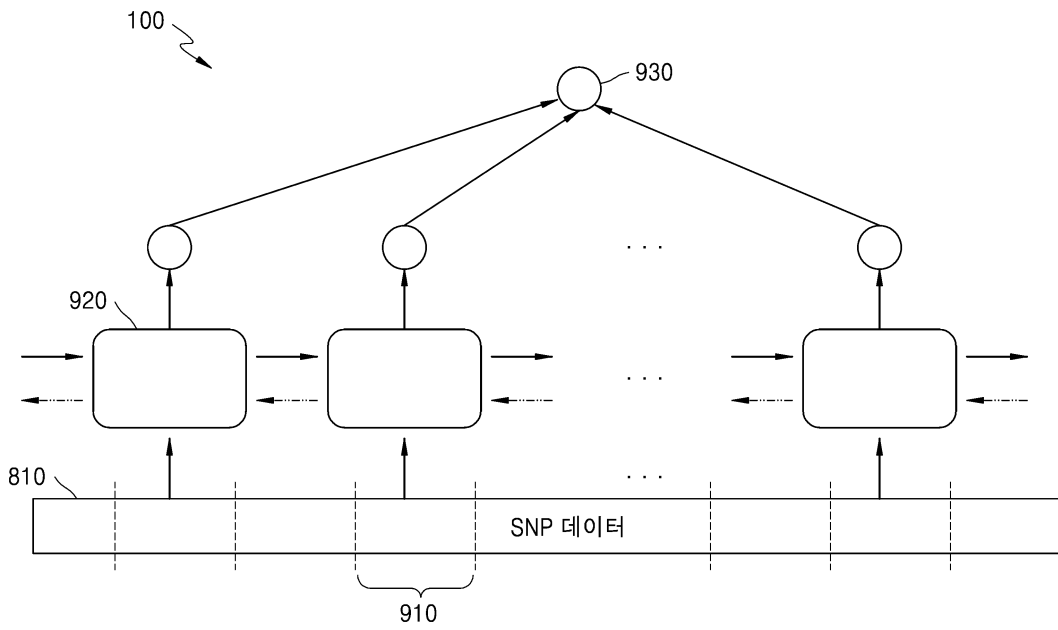
도면7



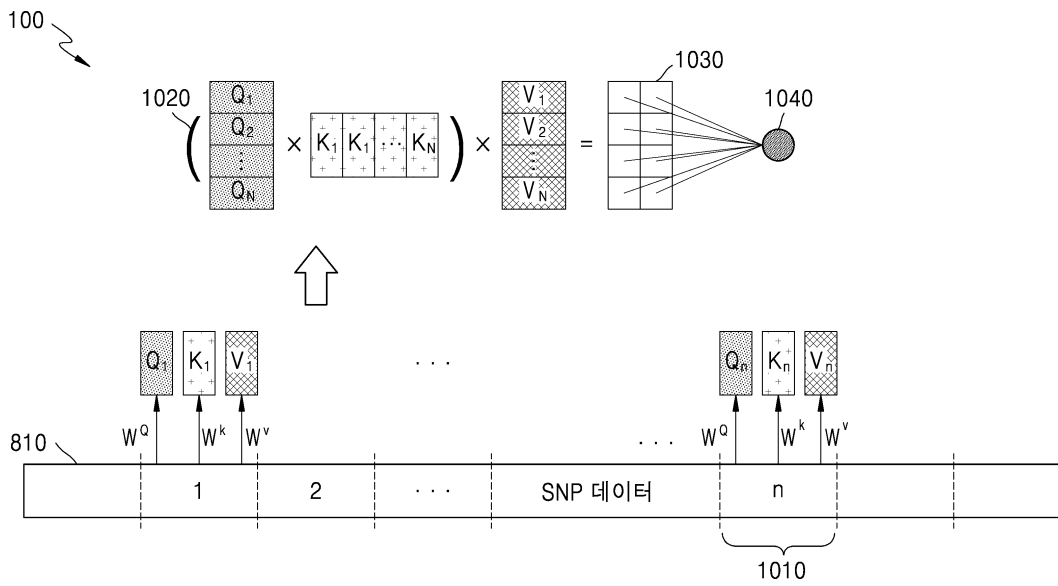
도면8



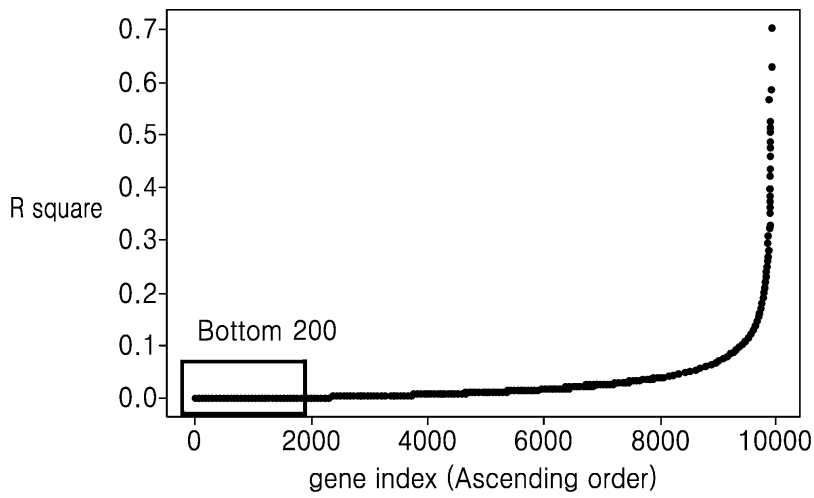
도면9



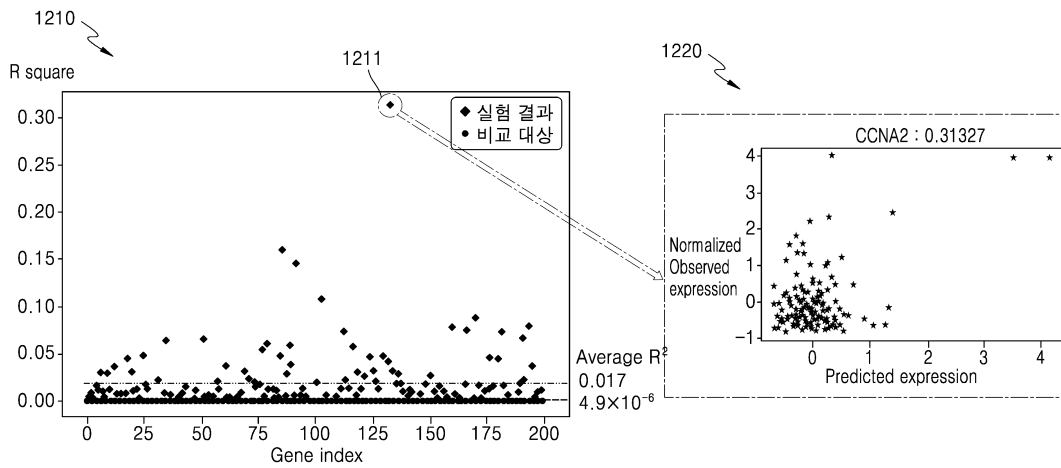
도면10



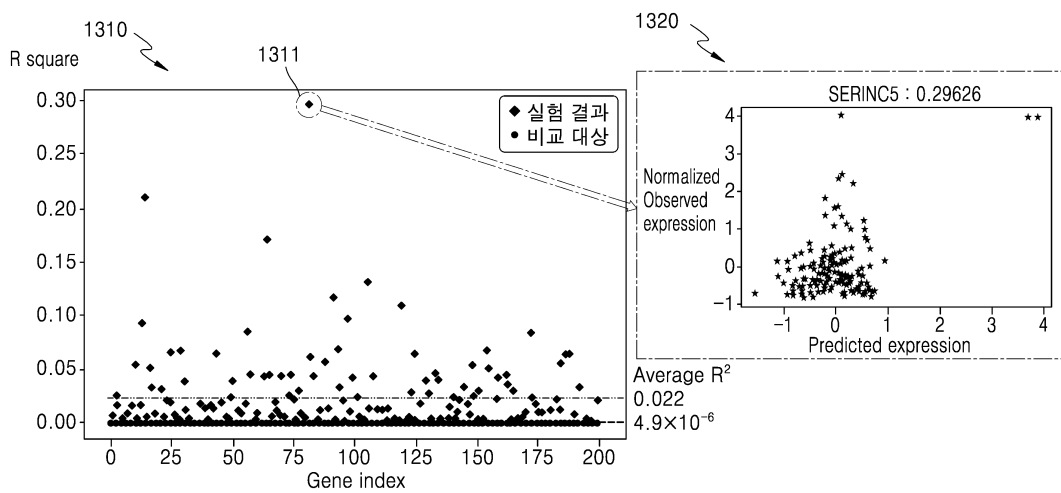
도면11



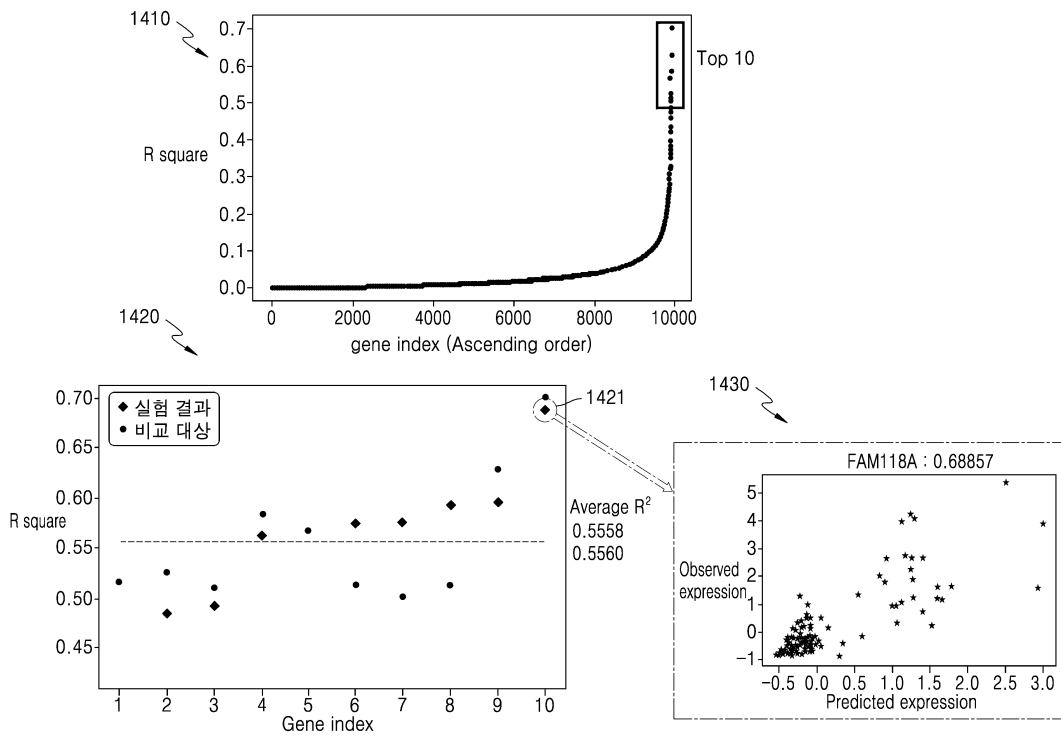
도면12



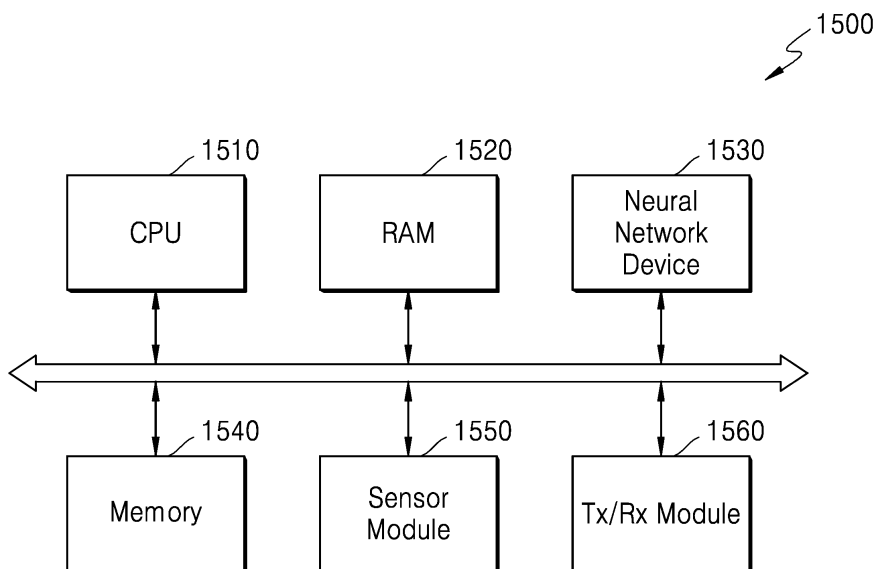
도면13



도면14



도면15



도면16

