(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0043072
(43) 공개일자 2023년03월30일

(51) 국제특허분류(Int. Cl.)
G16B 40/20 (2019.01) *G06N 3/04* (2023.01)
G06N 3/08 (2023.01) *G16B 20/20* (2019.01)
G16B 30/00 (2019.01) *G16B 45/00* (2019.01)
(52) CPC특허분류
G16B 40/20 (2019.02)
G06N 3/04 (2023.01)
(21) 출원번호 10-2022-7045561
(22) 출원일자(국제) 2021년07월21일
심사청구일자 없음
(85) 번역문제출일자 2022년12월23일
(86) 국제출원번호 PCT/US2021/042599
(87) 국제공개번호 WO 2022/020487
국제공개일자 2022년01월27일
(30) 우선권주장
63/055,724 2020년07월23일 미국(US)

(71) 출원인
일루미나, 인코포레이티드
미국 캘리포니아 92122 샌디에고 일루미나 웨이 5200
(72) 발명자
가오 홍
미국 캘리포니아주 92122 샌디에고 일루미나 웨이 5200
파르 카이-하우
미국 캘리포니아주 92122 샌디에고 일루미나 웨이 5200
맥레이 제레미 프란시스
미국 캘리포니아주 92122 샌디에고 일루미나 웨이 5200
(74) 대리인
특허법인아주김장리

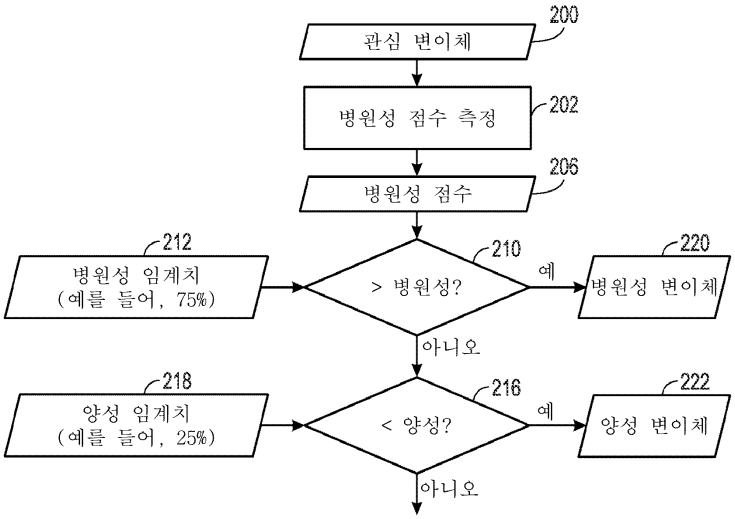
전체 청구항 수 : 총 21 항

(54) 발명의 명칭 **변이체 병원성 채점 및 분류 그리고 이의 사용**

(57) 요 약

유전자 변이체에 대한 병원성 점수(206)의 도출 및 사용이 본원에 설명된다. 병원성 채점 프로세스의 적용, 사용, 및 변형은 변이체를 병원성 또는 양성으로 특성화하기 위한 임계치(212, 218)의 도출 및 사용, 유전자 변이체와 연관된 선택 효과의 추정, 병원성 점수(206)를 사용한 유전병 유병률의 추정, 및 병원성 점수(206) 평가에 사용된 방법의 재보정을 포함하지만, 이에 제한되지 않는다.

대 표 도 - 도7



(52) CPC특허분류
G06N 3/08 (2023.01)
G16B 20/20 (2019.02)
G16B 30/00 (2019.02)
G16B 45/00 (2019.02)

명 세 서

청구범위

청구항 1

관심 변이체(200)를 분류하기 위한 방법으로서,

컴퓨터 시스템의 하나 이상의 프로세서에서, 상기 관심 변이체(200)를 분류하기 위한 저장된 명령어를 실행하는 단계를 포함하고, 상기 저장된 명령어는, 실행될 때, 상기 하나 이상의 프로세서가,

유전자의 관심 변이체(200)에 대한 병원성 점수(206)를 입력으로서 수신하는 단계;

상기 관심 변이체(200)에 대한 상기 병원성 점수(206)를 유전자-특이적 병원성 임계치(212)와 비교하는 단계; 및

상기 병원성 점수(206)가 상기 유전자-특이적 병원성 임계치(212)를 초과하는 것에 응답하여 상기 관심 변이체(200)를 병원성으로서 분류하는 단계(210)

를 포함하는 동작을 수행하게 하는, 관심 변이체(200)를 분류하기 위한 방법.

청구항 2

제1항에 있어서, 상기 병원성 점수(206)는 상기 관심 변이체(200)에 대한 인간 및 비인간 영장류에서 정제 선택의 정도를 기반으로 하는, 방법.

청구항 3

제1항에 있어서,

병원성 변이체에 대한 평균 병원성 점수와 병원성 점수의 백분위수 사이의 상관 관계를 결정하는 단계를 더 포함하고,

상기 병원성 변이체에 대한 평균 병원성 점수와 상기 병원성 점수의 백분위수 사이의 상관 관계는 상기 유전자-특이적 병원성 임계치(212)를 정의하는 데 사용되는, 방법.

청구항 4

제1항에 있어서, 상기 병원성 점수(206)는 아미노산 서열로부터 상기 병원성 점수(206)를 생성하도록 훈련된 신경망(102)에 의해 처리된 아미노산 서열에 기초하여 생성되는, 방법.

청구항 5

제4항에 있어서, 상기 아미노산 서열의 중심 아미노산은 상기 관심 변이체(200)에 대응하는, 방법.

청구항 6

제4항에 있어서, 상기 신경망(102)은 인간 서열과 비인간 서열 둘 모두를 사용하여 훈련되는, 방법.

청구항 7

제1항에 있어서, 상기 유전자-특이적 병원성 임계치(212)는 51번째 백분위수 내지 99번째 백분위수에 의해 정의되고 이를 포함하는 범위에 있는, 방법.

청구항 8

제1항에 있어서, 상기 유전자-특이적 병원성 임계치(212)는 75번째 백분위수 내지 99번째 백분위수에 의해 정의되고 이를 포함하는 범위에 있는, 방법.

청구항 9

제1항에 있어서, 상기 저장된 명령어는, 실행될 때, 상기 하나 이상의 프로세서가,

상기 관심 변이체(200)에 대한 상기 병원성 점수(206)를 유전자-특이적 양성 임계치(218)와 비교하는 단계; 및

상기 병원성 점수(206)가 상기 유전자-특이적 양성 임계치(218) 미만인 것에 응답하여, 상기 관심 변이체(200)를 양성으로 분류하는 단계

를 포함하는 추가 동작을 수행하게 하는, 방법.

청구항 10

제9항에 있어서,

양성 변이체에 대한 평균 병원성 점수와 병원성 점수의 백분위수 사이의 상관 관계를 결정하는 단계를 더 포함하고,

상기 양성 변이체에 대한 평균 병원성 점수와 상기 병원성 점수의 백분위수 사이의 상관 관계는 상기 유전자-특이적 양성 임계치(218)를 정의하는 데 사용되는, 방법.

청구항 11

제9항에 있어서, 상기 유전자-특이적 양성 임계치(218)는 1번째 백분위수 내지 49번째 백분위수에 의해 정의되고 이를 포함하는 범위에 있는, 방법.

청구항 12

제9항에 있어서, 상기 유전자-특이적 양성 임계치(218)는 1번째 백분위수 내지 25번째 백분위수에 의해 정의되고 이를 포함하는 범위에 있는, 방법.

청구항 13

프로세서 실행 가능 명령어를 저장하는 비일시적 컴퓨터 판독 가능 매체로서, 상기 명령어는, 하나 이상의 프로세서에 의해 실행될 때, 상기 하나 이상의 프로세서가,

병원성 채점 신경망(102)을 사용하여 유전자에 대한 관심 변이체(200)를 처리해서 상기 관심 변이체(200)에 대한 병원성 점수(206)를 생성하는 단계;

상기 관심 변이체(200)에 대한 상기 병원성 점수(206)를 상기 유전자에 특이적인 하나 이상의 임계치와 비교하는 단계; 및

상기 유전자에 특이적인 하나 이상의 임계치와 상기 병원성 점수(206)의 비교에 기초하여 상기 관심 변이체(206)를 병원성, 양성, 또는 병원성도 양성도 아닌 것으로 분류하는 단계

를 포함하는 단계를 수행하게 하는, 프로세서 실행 가능 명령어를 저장하는 비일시적 컴퓨터 판독 가능 매체.

청구항 14

제13항에 있어서, 상기 하나 이상의 임계치는 유전자-특이적 병원성 임계치(212)를 포함하고, 상기 관심 변이체(200)는 상기 병원성 점수(206)가 상기 유전자-특이적 병원성 임계치(212)를 초과하는 것에 응답하여 병원성으로 분류되는, 비일시적 컴퓨터 판독 가능 매체.

청구항 15

제13항에 있어서, 상기 하나 이상의 임계치는 유전자-특이적 양성 임계치(218)를 포함하고, 상기 관심 변이체(200)는 상기 병원성 점수(206)가 상기 유전자-특이적 양성 임계치(218) 미만인 것에 응답하여 양성으로 분류되는, 비일시적 컴퓨터 판독 가능 매체.

청구항 16

제13항에 있어서, 상기 유전자에 특이적인 하나 이상의 임계치는 유전자-특이적 병원성 임계치(212) 및 유전자-특이적 양성 임계치(218)를 포함하고, 상기 관심 변이체(200)는 상기 병원성 점수(206)가 상기 유전자-특이적 양성 임계치(218)와 상기 유전자-특이적 병원성 임계치(212) 사이에 있는 것에 응답하여 병원성도 양성도 아닌

것으로 분류되는, 비일시적 컴퓨터 판독 가능 매체.

청구항 17

제13항에 있어서, 상기 관심 변이체(200)에 대한 상기 병원성 점수(206)를 상기 유전자에 특이적인 하나 이상의 임계치와 비교하는 단계는 상기 관심 변이체(200)에 대한 상기 병원성 점수(206)를 제2 임계치 이전의 제1 임계치와 비교하는 단계를 포함하고, 상기 제2 임계치와의 비교는 상기 제1 임계치와의 비교 결과에 따라 좌우되는, 비일시적 컴퓨터 판독 가능 매체.

청구항 18

제13항에 있어서, 상기 관심 변이체(200)에 대한 상기 병원성 점수(206)를 상기 유전자에 특이적인 하나 이상의 임계치와 비교하는 단계는 상기 관심 변이체(200)에 대한 상기 병원성 점수(206)를 제1 임계치 및 제2 임계치와 병렬로 비교하는 단계를 포함하는, 비일시적 컴퓨터 판독 가능 매체.

청구항 19

변이체 분류 시스템으로서,

프로세서 실행 가능 명령어를 인코딩하는 하나 이상의 메모리 구조; 및

상기 하나 이상의 메모리 구조와 통신하는 하나 이상의 프로세서를 포함하고, 상기 하나 이상의 프로세서는 상기 명령어를 실행하여,

유전자의 관심 변이체(200)에 대한 병원성 점수(206)를 입력으로서 수신하는 단계;

상기 관심 변이체(200)에 대한 상기 병원성 점수(206)를 상기 유전자에 특이적인 하나 이상의 임계치와 비교하는 단계; 및

상기 유전자에 특이적인 하나 이상의 임계치와 상기 병원성 점수(206)의 비교에 기초하여 상기 관심 변이체(200)를 병원성, 양성, 또는 병원성도 양성도 아닌 것으로 분류하는 단계

를 포함하는 작업을 수행하게 하는, 변이체 분류 시스템.

청구항 20

제19항에 있어서, 상기 하나 이상의 임계치는 유전자-특이적 병원성 임계치(212)를 포함하고, 상기 관심 변이체(200)는 상기 병원성 점수(206)가 상기 유전자-특이적 병원성 임계치(212)를 초과하는 것에 응답하여 병원성으로 분류되는, 변이체 분류 시스템.

청구항 21

제19항에 있어서, 상기 하나 이상의 임계치는 유전자-특이적 양성 임계치(218)를 포함하고, 상기 관심 변이체(200)는 상기 병원성 점수(206)가 상기 유전자-특이적 양성 임계치(218) 미만인 것에 응답하여 양성으로 분류되는, 변이체 분류 시스템.

발명의 설명

기 술 분 야

[0001] 관련 출원의 교차 참조

[0002] 본 출원은 2020년 7월 23일자로 출원된 미국 가특허 출원 제63/055,724호에 대한 우선권을 주장하며, 이는 모든 목적을 위해 그 전체가 본원에 참조로서 원용된다.

[0003] 기술분야

[0004] 개시된 기술은 생물학적 서열 변이체의 병원성을 평가하고 병원성 평가를 사용하여 다른 병원성 관련 데이터를 도출할 목적으로 컴퓨터 및 디지털 데이터 처리 시스템에서 구현되는, 인공 지능으로 지칭될 수 있는, 기계 학습 기술의 사용에 관한 것이다. 이러한 접근법은 지능(즉, 지식 기반 시스템, 추론 시스템, 및 지식 획득 시스템)의 에뮬레이션을 위한 대응하는 데이터 처리 방법 및 제품 및/또는 불확실성이 있는 추론을 위한 시스템(예

를 들어, 퍼지 논리 시스템), 적응 시스템, 기계 학습 시스템, 및 인공 신경망을 포함하거나 이용한다. 특히, 개시된 기술은 병원성 평가뿐만 아니라 이러한 병원성 정보의 사용 또는 개선을 위한 심층 컨벌루션 신경망 훈련을 위한 딥 러닝 기반 기술의 사용에 관한 것이다.

배 경 기 술

[0005] 본 섹션에서 논의된 기술 요지는 단순히 이 섹션에서 언급된 결과로 선행 기술로 가정해서는 안 된다. 유사하게, 본 섹션에서 언급되거나 배경 기술로서 제공된 기술 요지와 연관된 문제는 종래 기술에서 이전에 인식된 것으로 가정해서는 안 된다. 본 섹션에서의 기술 요지는 단지 서론 다른 접근법을 나타낼 뿐이며, 그 자체로도 청구되는 기술의 구현예에 대응할 수 있다.

[0006] 유전적 변이는 많은 질병을 설명하는 데 도움이 될 수 있다. 모든 인간은 고유한 유전자 코드를 가지고 있으며 개체 그룹 내에는 많은 유전적 변이체가 있다. 유해한 많은 또는 대부분의 유전적 변이체는 자연 선택에 의해 게놈에서 고갈되었다. 그러나, 어떤 유전적 변이가 임상적 관심을 가질 가능성이 있는지 식별하는 것은 여전히 어렵다.

[0007] 더 나아가, 변이체의 특성 및 기능적 효과(예를 들어, 병원성)를 모델링하는 것은 유전체학 분야에서 어려운 작업이다. 기능적 게놈 시퀀싱 기술의 급속한 발전에도 불구하고, 변이체의 기능적 결과에 대한 해석은 세포 유형-특이적 전사 조절 시스템의 복잡성으로 인해 여전히 어려운 문제이다.

발명의 내용

[0008] 변이체 병원성 분류기를 구성하고 이러한 병원성 분류기 정보를 사용하거나 개선하기 위한 시스템, 방법, 및 제조품을 설명한다. 이러한 구현예는 본원에 설명된 시스템 및 방법론의 동작을 수행하기 위해 프로세서에 의해 실행 가능한 명령어를 저장하는 비일시적 컴퓨터 판독 가능 저장 매체를 포함하거나 이용할 수 있다. 일 구현예의 하나 이상의 특징은 명시적으로 나열되거나 설명되지 않은 경우에도 기본 구현예 또는 다른 구현예와 조합될 수 있다. 더 나아가, 일 구현예의 하나 이상의 특징이 다른 구현예와 조합될 수 있도록, 상호 배타적이지 않은 구현예는 조합 가능한 것으로 교시된다. 본 개시내용은 이러한 옵션을 사용자에게 주기적으로 상기시킬 수 있다. 그러나, 이러한 옵션을 반복하는 설명이 일부 구현예로부터 누락된 것은 다음 섹션에서 교시되는 잠재적인 조합을 제한하는 것으로 취해져서는 안 된다. 대신, 이러한 설명은 참조로서 다음의 구현예 각각에 원용된다.

[0009] 이러한 시스템 구현예 및 개시된 다른 시스템은 본원에서 논의된 바와 같은 특징의 일부 또는 전부를 선택적으로 포함한다. 시스템은 개시된 방법과 관련하여 설명된 특징을 포함할 수도 있다. 간결함을 위해, 시스템 특징의 대안적 조합은 개별적으로 열거되지 않는다. 더 나아가, 시스템, 방법, 및 제조품에 적용 가능한 특징은 기본 특징의 각 법정 클래스 세트에 대해 반복되지 않는다. 독자는 식별된 특징이 다른 법정 클래스의 기본 특징과 어떻게 쉽게 조합될 수 있는지 이해할 것이다.

[0010] 논의된 기술 요지의 일 양태에서, 메모리에 결합된 수많은 프로세서에서 실행되는 컨벌루션 신경망 기반 변이체 병원성 분류기를 훈련시키는 방법론 및 시스템이 설명된다. 대안적으로, 다른 시스템 구현예에서, 훈련되거나 적합하게 매개화된 통계 모델 또는 기술 및/또는 다른 기계 학습 접근법이 신경망 기반 분류기에 추가로 또는 이의 대안으로 이용될 수 있다. 시스템은 양성 변이체 및 병원성 변이체로부터 생성된 단백질 서열 쌍의 양성 훈련 예시 및 병원성 훈련 예시를 사용한다. 양성 변이체는 부합하는 참조 코돈 서열을 인간과 공유하는 대체 비인간 영장류 코돈 서열에서 발생하는 일반적인 인간 미스센스 변이체 및 비인간 영장류 미스센스 변이체를 포함한다. 샘플링된 인간은 아프리카/아프리카계 미국인(약칭 AFR), 미국인(약칭 AMR), 애슈케나지 유대인(약칭 ASJ), 동아시아인(약칭 EAS), 핀란드인(약칭 FIN), 비핀란드 유럽인(약칭 NFE), 남아시아인(약칭 SAS) 및 기타(약칭 OTH)를 포함하거나 이로 특징지어질 수 있는 서로 다른 인간 부분모집단에 속할 수 있다. 비인간 영장류 미스센스 변이체는 침팬지, 보노보, 고릴라, B. 오랑우탄, S. 오랑우탄, 레수스, 및 마모셋을 포함하나 반드시 이에 제한되지 않는 복수의 비인간 영장류 종으로부터의 미스센스 변이체를 포함한다.

[0011] 본원에서 논의된 바와 같이, 수많은 프로세서에서 실행되는 심층 컨벌루션 신경망은 변이체 아미노산 서열을 양성 또는 병원성으로 분류하도록 훈련될 수 있다. 따라서, 이러한 심층 컨벌루션 신경망의 출력은 변이체 아미노산 서열에 대한 병원성 점수 또는 분류를 포함할 수 있지만 이에 제한되지 않는다. 이해할 수 있는 바와 같이, 특정 구현예에서, 적합하게 매개화된 통계 모델 또는 기술 및/또는 다른 기계 학습 접근법이 신경망 기반 접근법에 추가로 또는 이의 대안으로 이용될 수 있다.

[0012] 본원에서 논의된 특정 실시예에서, 병원성 처리 및/또는 채점 작업은 추가 특징 또는 양태를 포함할 수 있다. 예를 들어, 변이체를 양성 또는 병원성으로 평가하거나 채점하는 것과 같은 감정 또는 평가 프로세스의 일부로서 다양한 병원성 채점 임계치가 이용될 수 있다. 예를 들어, 특정 구현예에서, 가능성 있는 병원성 변이체에 대한 임계치로서 사용하기 위한 유전자당 병원성 점수의 적합한 백분위수는 51% 내지 99% 범위, 예를 들어 51번째, 55번째, 65번째, 70번째, 75번째, 80번째, 85번째, 90번째, 95번째, 또는 99번째 백분위수일 수 있으나, 이에 제한되지 않는다. 반대로, 가능성 있는 양성 변이체에 대한 임계치로서 사용하기 위한 유전자당 병원성 점수의 적합한 백분위수는 1% 내지 49% 범위, 예를 들어 1번째, 5번째, 10번째, 15번째, 20번째, 25번째, 30번째, 35번째, 40번째, 또는 45번째 백분위수일 수 있으나, 이에 제한되지 않는다.

[0013] 추가 실시예에서, 병원성 처리 및/또는 채점 작업은 선택 효과가 추정되게 하는 추가 특징 또는 양태를 포함할 수 있다. 이러한 실시예에서, 돌연변이율 및/또는 선택을 특징짓는 적합한 입력을 사용하여, 주어진 모집단 내의 대립유전자 빈도의 순방향 시간 시뮬레이션을 이용해 관심 유전자에서 대립유전자 빈도 스펙트럼을 생성할 수 있다. 그런 다음, 예를 들어 선택이 있거나 없는 대립유전자 빈도 스펙트럼을 비교하고 대응하는 선택-고갈 함수를 피팅시키거나 특성화함으로써 관심 변이체에 대해 고갈 메트릭을 계산할 수 있다. 주어진 병원성 점수 및 이러한 선택-고갈 함수에 기초하여, 선택 계수는 변이체에 대해 생성된 병원성 점수에 기초하여 주어진 변이체에 대해 결정될 수 있다.

[0014] 추가 양태에서, 병원성 처리 및/또는 채점 작업은 유전병 유병률이 병원성 점수를 사용하여 추정되게 하는 추가 특징 또는 양태를 포함할 수 있다. 각 유전자에 대한 유전병 유병률 메트릭의 계산과 관련하여, 제1 방법론에서, 유해 변이체 세트의 트리뉴클레오티드 컨텍스트 구성이 초기에 얻어진다. 이러한 세트에서 각 트리뉴클레오티드 컨텍스트에 대해, 특정 선택 계수(예를 들어, 0.01)를 가정하는 순방향 시간 시뮬레이션을 수행하여 해당 트리뉴클레오티드 컨텍스트에 대한 예상된 대립유전자 빈도 스펙트럼(AFS)을 생성한다. 유전자에서 트리뉴클레오티드의 빈도에 의해 가중된 트리뉴클레오티드에 걸친 AFS를 합산하면 유전자에 대해 예상된 AFS가 생성된다. 이러한 접근법에 따른 유전병 유병률 메트릭은 해당 유전자에 대한 임계치를 초과하는 병원성 점수를 갖는 변이체의 예상된 누적 대립유전자 빈도로 정의될 수 있다.

[0015] 추가 양태에서, 병원성 처리 및/또는 채점 작업은 병원성 채점을 재보정하기 위한 특징 또는 방법론을 포함할 수 있다. 이러한 재보정과 관련하여, 하나의 예시적인 실시예에서, 재보정 접근법은 변이체의 병원성 점수의 백분위수에 초점을 맞출 수 있는데, 이는 더 강력하고 전체 유전자에 가해지는 선택 압력에 의해 덜 영향을 받을 수 있기 때문이다. 일 구현예에 따르면, 병원성 점수의 각 백분위수에 대한 생존 확률이 계산되며, 이는 병원성 점수의 백분위수가 높을수록 변이체가 정제 선택에서 생존할 기회가 적다는 것을 암시하는 생존 확률 보정 계수를 구성한다. 생존 확률 보정 계수는 미스센스 변이체에서 선택 계수의 추정에 대한 노이즈의 영향을 완화하는 데 도움이 되도록 재보정을 수행하는 데 이용될 수 있다.

[0016] 전술한 설명은 개시된 기술의 제작 및 사용을 가능하게 하기 위해 제시된다. 개시된 구현예에 대한 다양한 변형예는 명백할 것이며, 본원에서 정의된 일반적인 원리는 개시된 기술의 사상 및 범주로부터 벗어남이 없이 다른 구현예 및 적용 분야에 적용될 수 있다. 따라서, 개시된 기술은 도시된 구현예로 제한되도록 의도된 것이 아니라, 본원에 개시된 원리 및 특징과 일치하는 가장 넓은 범주에 부합되어야 한다. 개시된 기술의 범위는 첨부된 청구범위에 의해 정의된다.

도면의 간단한 설명

[0017] 본 발명의 이러한 및 다른 특징, 양태, 및 이점은 첨부 도면을 참조하여 다음의 상세한 설명을 판독할 때 더 잘 이해될 것이며, 도면 전체에서 유사한 문자는 유사한 부분을 나타낸다.

도 1은 개시된 기술의 일 구현예에 따른 컨벌루션 신경망을 훈련시키는 양태의 블록도이고;

도 2는 개시된 기술의 일 구현예에 따른 단백질의 2차 구조 및 용매 접근성을 예측하기 위해 사용되는 딥 러닝 네트워크 아키텍처를 도시하고 있고;

도 3은 개시된 기술의 일 구현예에 따른 병원성 예측을 위한 심층 잔차 네트워크의 예시적인 아키텍처를 도시하고 있고;

도 4는 개시된 기술의 일 구현예에 따른 병원성 점수 분포를 도시하고 있고;

도 5는 개시된 기술의 일 구현예에 따른 ClinVar 병원성 변이체에 대한 평균 병원성 점수와 해당 유전자 내의

모든 미스센스 변이체의 75번째 백분위수에서의 병원성 점수의 상관 관계의 플롯을 도시하고 있고;

도 6은 개시된 기술의 일 구현예에 따른 ClinVar 양성 변이체에 대한 평균 병원성 점수와 해당 유전자 내의 모든 미스센스 변이체의 25번째 백분위수에서의 병원성 점수의 상관 관계의 플롯을 도시하고 있고;

도 7은 개시된 기술의 일 구현예에 따른 병원성 점수에 기초하여 변이체를 양성 또는 병원성 범주로 특성화하기 위해 임계치가 사용될 수 있는 샘플 프로세스 흐름을 도시하고 있고;

도 8은 개시된 기술의 일 구현예에 따른 최적의 순방향 시간 모델 파라미터가 도출될 수 있는 샘플 프로세스 흐름을 도시하고 있고;

도 9는 개시된 기술의 일 구현예에 따른 상이한 성장률을 갖는 4개의 기하급수적 확장 단계로 단순화된 인간 모집단의 진화 역사를 도시하고 있고;

도 10은 본 접근법에 따라 도출된 돌연변이율의 추정치와 다른 문헌에서 도출된 돌연변이율 사이의 상관 관계를 도시하고 있고;

도 11은 본 개시내용의 양태에 따른 CpG 돌연변이의 예상된 수 대 메틸화 수준에 대한 관측된 수의 비율을 도시하고 있고;

도 12a, 도 12b, 도 12c, 도 12d, 및 도 12e는 본 개시내용의 양태에 따른 순방향 시간 시뮬레이션 모델의 구현을 위한 최적의 파라미터 조합을 나타낸 피어슨의 카이 제곱 통계의 히트맵을 도시하고 있고;

도 13은 일례에서, 본 접근법에 따라 결정된 최적의 모델 파라미터를 사용하여 도출된 모의 대립유전자 빈도 스펙트럼이 관측된 대립유전자 빈도 스펙트럼에 대응함을 도시하고 있고;

도 14는 개시된 기술의 일 구현예에 따른 순방향 시간 시뮬레이션의 컨텍스트에서 선택 효과가 통합되는 샘플 프로세스 흐름을 도시하고 있고;

도 15는 본 개시내용의 양태에 따른 선택-고갈 곡선의 일례를 도시하고 있고;

도 16은 개시된 기술의 일 구현예에 따른 관심 변이체에 대한 선택 계수가 도출될 수 있는 샘플 프로세스 흐름을 도시하고 있고;

도 17은 개시된 기술의 일 구현예에 따른 병원성-고갈 관계가 도출될 수 있는 샘플 프로세스 흐름을 도시하고 있고;

도 18은 본 개시내용의 양태에 따른 BRCA1 유전자에 대한 병원성 점수 대 고갈의 플롯을 도시하고 있고;

도 19는 본 개시내용의 양태에 따른 LDLR 유전자에 대한 병원성 점수 대 고갈의 플롯을 도시하고 있고;

도 20은 개시된 기술의 일 구현예에 따른 누적 대립유전자 빈도가 도출될 수 있는 샘플 프로세스 흐름을 도시하고 있고;

도 21은 개시된 기술의 일 구현예에 따른 예상된 누적 대립유전자 빈도가 도출될 수 있는 일반화된 샘플 프로세스 흐름을 도시하고 있고;

도 22는 본 개시내용의 양태에 따른 예상된 누적 대립유전자 빈도 대 관측된 누적 대립유전자 빈도의 플롯을 도시하고 있고;

도 23은 본 개시내용의 양태에 따른 예상된 누적 대립유전자 빈도 대 질병 유병률의 플롯을 도시하고 있고;

도 24는 개시된 기술의 일 구현예에 따른 예상된 누적 대립유전자 빈도가 도출될 수 있는 제1 샘플 프로세스 흐름을 도시하고 있고;

도 25는 개시된 기술의 일 구현예에 따른 예상된 누적 대립유전자 빈도가 도출될 수 있는 제2 샘플 프로세스 흐름을 도시하고 있고;

도 26은 본 개시내용의 양태에 따른 예상된 누적 대립유전자 빈도 대 관측된 누적 대립유전자 빈도의 플롯을 도시하고 있고;

도 27은 본 개시내용의 양태에 따른 예상된 누적 대립유전자 빈도 대 질병 유병률의 플롯을 도시하고 있고;

도 28은 개시된 기술의 일 구현예에 따른 병원성 채점 프로세스에 대한 재보정 접근법의 양태와 관련된 샘플 프

로세스 흐름을 도시하고 있고;

도 29는 본 개시내용의 양태에 따른 병원성 점수 백분위수 대 확률의 분포를 도시하고 있고;

도 30은 본 개시내용의 양태에 따른 가우시안 노이즈로 오버레이된 관측된 병원성 점수 백분위수의 이산 균일 분포의 밀도 플롯을 도시하고 있고;

도 31은 본 개시내용의 양태에 따른 가우시안 노이즈로 오버레이된 관측된 병원성 점수 백분위수의 이산 균일 분포의 누적 분포 함수를 도시하고 있고;

도 32는 본 개시내용의 양태에 따른 실제 병원성 점수 백분위수(x축)를 갖는 변이체가 관측된 병원성 점수 백분위수 간격(y축)에 속할 확률을 히트맵을 통해 도시하고 있고;

도 33은 개시된 기술의 일 구현예에 따른 보정 계수를 결정하는 단계의 샘플 프로세스 흐름을 도시하고 있고;

도 34는 본 개시내용의 양태에 따른 SCN2A 유전자의 미스센스 변이체에 대한 10개의 빈의 백분위수에 걸친 고갈 확률을 도시하고 있고;

도 35는 본 개시내용의 양태에 따른 SCN2A 유전자의 미스센스 변이체에 대한 10개의 빈의 백분위수에 걸친 생존 확률을 도시하고 있고;

도 36은 개시된 기술의 일 구현예에 따른 교정된 고갈 메트릭을 결정하는 단계의 샘플 프로세스 흐름을 도시하고 있고;

도 37은 본 개시내용의 양태에 따른 실제 병원성 점수 백분위수(x축)를 갖는 변이체가 관측된 병원성 점수 백분위수 간격(y축)에 속할 확률을 전달하는 교정 또는 재보정된 히트맵을 도시하고 있고;

도 38은 본 개시내용의 양태에 따른 각 병원성 점수 백분위수 빈에 대한 교정된 고갈 메트릭의 플롯을 도시하고 있고;

도 39는 본 개시내용의 양태에 따른 다수의 계층을 갖는 피드-포워드 신경망의 일 구현예를 도시하고 있고;

도 40은 본 개시내용의 양태에 따른 컨벌루션 신경망의 일 구현예의 일례를 도시하고 있고;

도 41은 본 개시내용의 양태에 따른 특징 맵 추가를 통해 하류에 사전 정보를 재주입하는 잔차 연결을 도시하고 있고;

도 42는 개시된 기술이 작동될 수 있는 예시적인 컴퓨팅 환경을 도시하고 있고;

도 43은 개시된 기술을 구현하는 데 사용될 수 있는 컴퓨터 시스템의 단순화된 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0018] 다음 논의는 어느 당업자라도 개시된 기술을 제조하고 사용할 수 있도록 제시되며, 특정 적용 분야 및 이의 요건과 관련하여 제공된다. 개시된 구현예에 대한 다양한 변형은 당업자에게 용이하게 명백할 것이며, 본원에서 정의된 일반적인 원리는 개시된 기술의 사상 및 범주로부터 벗어남이 없이 다른 구현예 및 적용 분야에 적용될 수 있다. 따라서, 개시된 기술은 도시된 구현예로 제한되도록 의도된 것이 아니라, 본원에 개시된 원리 및 특징과 일치하는 가장 넓은 범주에 부합되어야 한다.

[0019] **I.서론**

[0020] 다음 논의는 변이체 병원성 점수 또는 분류의 생성 및 이러한 병원성 점수 또는 분류에 기초한 유용한 임상 분석 또는 메트릭의 도출과 같은 아래에서 논의되는 특정 분석을 구현하는 데 사용될 수 있는 컨벌루션 신경망을 포함하는 신경망의 훈련 및 사용과 관련된 양태를 포함한다. 이를 염두에 두고, 이러한 신경망의 특정 양태 및 특징이 본 기술을 설명하는 데 언급되거나 참조될 수 있다. 논의를 간소화하기 위해, 본 기술에 대한 설명은 이러한 신경망에 대한 기본 지식이 가정된다. 그러나, 관련 신경망 개념에 대한 추가 정보 및 설명은 관련 신경망 개념에 대한 추가적인 설명을 원하는 사람들을 위해 설명의 말미에 제공된다. 더 나아가, 본원에서는 유용한 예를 제공하고 설명을 용이하게 하기 위해 신경망이 주로 논의되지만, 훈련되거나 적합하게 매개화된 통계 모델 또는 기술 및/또는 다른 기계 학습 접근법을 포함하나 이에 제한되지 않는 다른 구현예가 신경망 접근법 대신에 또는 이에 추가하여 이용될 수 있다.

[0021] 특히, 다음 논의는 특정 관심 게놈 데이터를 분석하는 데 사용되는 구현예에서 신경망(예를 들어, 컨벌루션 신

경망)과 관련된 특정 개념을 이용할 수 있다. 이를 염두에 두고, 기본적인 생물학적 및 유전적 관심 문제의 특정 양태가 본원에서 설명되어 본원에서 논의된 신경망 기술이 이용될 수 있는 문제에 대한 유용한 컨텍스트를 제공한다.

[0022] 유전적 변이는 많은 질병을 설명하는 데 도움이 될 수 있다. 모든 인간은 고유한 유전자 코드를 가지고 있으며 개체 그룹 내에는 많은 유전적 변이체가 있다. 유해한 많은 또는 대부분의 유전적 변이체는 자연 선택에 의해 게놈에서 고갈되었다. 그러나, 어떤 유전적 변이가 병원성이거나 유해할 가능성이 있는지 식별하는 것이 여전히 바람직하다. 특히, 이러한 지식은 연구자가 병원성 유전 변이체에 집중하고 많은 질병의 진단 및 치료 속도를 가속화하는 데 도움이 될 수 있다.

[0023] 변이체의 특성 및 기능적 효과(예를 들어, 병원성)를 모델링하는 것은 유전체학 분야에서 중요하지만 어려운 작업이다. 기능적 게놈 시퀀싱 기술의 급속한 발전에도 불구하고, 변이체의 기능적 결과에 대한 해석은 세포 유형-특이적 전사 조절 시스템의 복잡성으로 인해 여전히 어려운 문제이다. 그러므로, 변이체의 병원성을 예측하기 위한 강력한 계산 모델은 기초 과학과 중개 연구 둘 모두에 상당한 이점을 가질 수 있다.

[0024] 더 나아가, 지난 수십 년간 생화학 기술의 발전으로 인해 이전보다 훨씬 저렴한 비용으로 신속하게 게놈 데이터를 생성하는 차세대 시퀀싱(NGS) 플랫폼이 등장하여 점점 더 많은 양의 게놈 데이터가 생성된다. 이렇게 압도적으로 많은 양의 시퀀싱된 DNA는 주석을 달기가 어렵다. 지도 기계 학습 알고리즘은 통상적으로 많은 양의 레이블링된 데이터를 이용할 수 있을 때 잘 수행된다. 그러나, 생물 정보학 및 기타 여러 데이터가 풍부한 분야에서, 인스턴스를 레이블링하는 프로세스는 비용이 많이 든다. 반대로, 레이블링되지 않은 인스턴스는 저렴하고 쉽게 이용 가능하다. 레이블링된 데이터의 양이 상대적으로 적고 레이블링되지 않은 데이터의 양이 상당히 큰 시나리오의 경우, 준지도 학습이 수동 레이블링에 대한 비용 효율적인 대안을 나타낸다. 그러므로, 이는 변이체의 병원성을 정확하게 예측하는 딥 러닝 기반 병원성 분류기를 구성하기 위해 준지도 알고리즘을 사용할 수 있는 기회를 제공한다. 인간의 확인 편향이 없는 병원성 변이체의 데이터베이스가 생성될 수 있다.

[0025] 기계 학습 기반 병원성 분류기에 관하여, 심층 신경망은 다수의 비선형 및 복잡한 변환 계층을 사용하여 높은 수준의 특징을 연속적으로 모델링하는 일종의 인공 신경망이다. 심층 신경망은 파라미터를 조정하기 위해 관측된 출력과 예측된 출력 간의 차이를 전달하는 역전파를 통해 피드백을 제공한다. 심층 신경망은 대규모 훈련 데이터세트의 가용성, 병렬 및 분산 컴퓨팅의 파워, 및 정교한 훈련 알고리즘으로 진화되었다.

[0026] 컨벌루션 신경망(CNN) 및 순환 신경망(RNN)은 심층 신경망의 구성요소이다. 컨벌루션 신경망은 컨벌루션 계층, 비선형 계층, 및 풀링 계층을 포함하는 아키텍처를 가질 수 있다. 순환 신경망은 퍼셉트론, 장단기 메모리 유닛, 및 게이트형 순환 유닛과 같은 빌딩 블록 사이에서 주기적 연결을 통해 입력 데이터의 순차적 정보를 이용하도록 설계된다. 또한, 심층 시공간 신경망, 다차원 순환 신경망, 및 컨벌루션 자동 인코더와 같은 많은 다른 신생의 심층 신경망이 제한된 컨텍스트에 대해 제안되었다.

[0027] 서열 데이터가 다차원 및 고차원임을 고려하면, 심층 신경망은 광범위한 적용 가능성 및 향상된 예측력으로 인해 생물 정보학 연구에 유망하다. 컨벌루션 신경망은 모티프 발견, 병원성 변이체 식별, 및 유전자 발현 추론과 같은 유전체학에서의 서열 기반 문제를 해결하도록 구성되었다. 컨벌루션 신경망은 DNA를 연구하는 데 유용한 가중치 공유 전략을 사용하는데, 이는 컨벌루션 신경망이, 중요한 생물학적 기능을 가진 것으로 추정되는, DNA에서 짧고 순환되는 로컬 패턴인 서열 모티프를 포착할 수 있기 때문이다. 컨벌루션 신경망의 특질은 컨벌루션 필터의 사용이다. 정교하게 설계되고 수동으로 제작된 특징에 기초하는 전통적인 분류 접근법과는 달리, 컨벌루션 필터는 원시 입력 데이터를 유익한 지식 표현에 매핑하는 프로세스와 유사하게 특징의 적응식 학습을 수행한다. 이러한 의미에서, 컨벌루션 필터는 일련의 모티프 스캐너로서 역할을 하는데, 이는 이러한 필터 세트가 훈련 절차 동안 입력에서 관련 패턴을 인식하고 그 자체를 업데이트할 수 있기 때문이다. 순환 신경망은 단백질 또는 DNA 서열과 같은 다양한 길이의 순차적 데이터에서 장거리 의존성을 포착할 수 있다.

[0028] 도 1에 개략적으로 도시된 바와 같이, 심층 신경망의 훈련은 각 계층에서 가중치 파라미터를 최적화하는 것을 포함하며, 이는 데이터로부터 가장 적합한 계층적 표현을 학습할 수 있도록 간단한 특징을 복잡한 특징으로 점진적으로 조합한다. 최적화 프로세스의 단일 사이클은 다음과 같이 조직화된다. 먼저, 훈련 데이터세트(예를 들어, 이러한 예에서는 입력 데이터(100))가 주어지면, 순방향 패스는 각 계층의 출력을 순차적으로 계산하고 기능 신호를 신경망(102)을 통해 전방으로 전파한다. 최종 출력 계층에서, 목적 손실 함수(비교 단계 106)는 추론 출력(110)과 주어진 레이블(112) 사이의 오류(104)를 측정한다. 훈련 오류를 최소화하기 위해, 역방향 패스는 연쇄 법칙을 사용하여 오류 신호를 역전파하고(단계 114) 신경망(102) 전체에 걸쳐 모든 가중치에 대한 구배를 계산한다. 마지막으로, 가중치 파라미터는 확률적 경사 하강법 또는 다른 적합한 접근법에 기반한 최적화

알고리즘을 사용하여 업데이트된다(단계 120). 배치 경사 하강법은 각각의 완전한 데이터세트에 대한 파라미터 업데이트를 수행하는 반면, 확률적 경사 하강법은 각각의 작은 데이터 예시 세트에 대한 업데이트를 수행함으로써 확률적 근사치를 제공한다. 여러 최적화 알고리즘은 확률적 경사 하강법으로부터 유래한다. 예를 들어, Adagrad 및 Adam 훈련 알고리즘은 각 파라미터에 대한 구배의 모멘트 및 업데이트 빈도에 기초하여 학습률을 적응적으로 수정하면서 확률적 경사 하강법을 수행한다.

[0029] 심층 신경망의 훈련에서의 다른 요소는 규칙화인데, 이는 오비피팅을 방지하여 양호한 일반화 성능을 달성하도록 의도된 전략을 지칭한다. 예를 들어, 가중치 감쇠는 가중치 파라미터가 더 작은 절대값으로 수렴하도록 목적 손실 함수에 패널티 항을 추가한다. 드롭아웃은 훈련 동안 신경망으로부터 은닉 유닛을 무작위로 제거하며, 가능한 서브네트워크의 앙상블로 간주될 수 있다. 더 나아가, 배치 정규화는 미니 배치 내에서 각 활성화를 위한 스칼라 특징의 정규화 및 각 평균과 분산을 파라미터로서 학습을 통해 새로운 규칙화 방법을 제공한다.

[0030] 현재 설명된 기술과 관련하여 이전의 높은 수준의 개요를 염두에 두고, 현재 설명된 기술은 많은 수의 인간 공학 특징 및 메타 분류기를 이용하는 이전의 병원성 분류 모델과 상이하다. 대조적으로, 본원에 설명된 기술의 특정 실시예에서, 관심 변이체 측면에 있는 아미노산 서열만을 입력으로 취하고 다른 종에서 이종상동성 서열 정렬을 취하는 간단한 딥 러닝 잔차 네트워크가 이용될 수 있다. 특정 구현예에서, 네트워크에 단백질 구조에 대한 정보를 제공하기 위해, 2개의 별도 네트워크가 서열 단독으로부터 각각 2차 구조 및 용매 접근성을 학습하도록 훈련될 수 있다. 이들은 단백질 구조에 대한 영향을 예측하기 위해 더 큰 딥 러닝 네트워크에서 서브네트워크로 통합될 수 있다. 서열을 시작점으로 사용하면 불완전하게 확인되거나 일관성 없이 적용될 수 있는 단백질 구조 및 기능적 도메인 주석에서 잠재적 편향을 방지할 수 있다.

[0031] 딥 러닝 분류기의 정확도는 훈련 데이터세트의 크기에 따라 조정되며, 6종의 영장류 종 각각으로부터의 변이 데이터는 독립적으로 분류기의 정확도를 높이는 데 기여한다. 단백질 변형 변이체에 대한 선택적 압력이 영장류 혈통 내에서 대체로 일치한다는 증거와 함께, 현존하는 비인간 영장류의 많은 수와 다양성은 체계적인 영장류 모집단 시퀀싱이 현재 임상 게놈 해석을 제한하는 의미를 알 수 없는 수백만 개의 인간 변이체를 분류하는 효과적인 전략임을 시사한다.

[0032] 더 나아가, 일반적인 영장류 변이는 메타 분류기의 확산으로 인해 객관적으로 평가하기 어려웠던 이전에 사용된 훈련 데이터와 완전히 독립적인 기존 방법을 평가하기 위한 명확한 검증 데이터세트를 제공한다. 본원에 설명된 본 모델의 성능은 10,000개의 유지된 영장류 일반 변이체를 사용하여 4개의 다른 대중적인 분류 알고리즘(Sift, Polyphen2, CADD, M-CAP)과 함께 평가되었다. 모든 인간의 미스센스 변이체의 대략 50%가 일반적인 대립유전자 빈도에서 자연 선택에 의해 제거되기 때문에, 50번째 백분위수 점수는 돌연변이율에 의해 10,000개의 유지된 영장류 일반 변이체와 부합되는 무작위로 선택된 미스센스 변이체 세트의 각 분류기에 대해 계산되었으며, 해당 임계치는 유지된 영장류 일반 변이체를 평가하는 데 사용되었다. 본 개시된 딥 러닝 모델의 정확도는 인간의 일반 변이체에 대해서만 훈련된 딥 러닝 네트워크를 사용하거나 또는 인간의 일반 변이체와 영장류 변이체 둘 모두를 사용하여, 이러한 독립적인 검증 데이터세트에 대해 다른 분류기보다 훨씬 뛰어났다.

[0033] 이전 내용을 염두에 두고, 요약하면, 본원에 설명된 방법론은 다양한 방식으로 변이체의 병원성을 예측하기 위한 기존 방법과 상이하다. 첫째, 현재 설명된 접근법은 준지도 심층 컨벌루션 신경망의 새로운 아키텍처를 채택한다. 둘째, 신뢰 가능한 양성 변이체는 인간의 일반 변이체(예를 들어, gnomAD) 및 영장류 변이체로부터 얻어지고, 매우 확신적인 병원성 훈련 세트는 동일한 인간 선별 변이체 데이터베이스를 사용하여 모델의 순환 훈련 및 테스트를 피하기 위해 반복적인 균형 잡힌 샘플링 및 훈련을 통해 생성된다. 셋째, 2차 구조 및 용매 접근성에 대한 딥 러닝 모델은 병원성 모델의 아키텍처에 통합된다. 구조 및 용매 모델로부터 얻은 정보는 특정 아미노산 잔기에 대한 레이블 예측으로 제한되지 않는다. 오히려, 판독 계층은 구조 및 용매 모델로부터 제거되고, 사전 훈련된 모델은 병원성 모델과 병합된다. 병원성 모델을 훈련하는 동안, 구조 및 용매 사전 훈련된 계층도 오류를 최소화하기 위해 역전파된다. 이는 사전 훈련된 구조 및 용매 모델이 병원성 예측 문제에 집중하는 데 도움이 된다.

[0034] 또한 본원에서 논의된 바와 같이, 본원에 기재된 바와 같이 훈련되고 사용된 모델의 출력(예를 들어, 병원성 점수 및/또는 분류)은 임상적으로 중요한 변이체 범위에 대한 선택 효과 추정 및 유전병 유병률 추정과 같은 가치 있는 추가 데이터 또는 진단을 생성하는 데 사용될 수 있다. 모델 출력의 재보정 및 병원성과 양성 변이체를 특성화하기 위한 임계값의 생성과 사용과 같은, 다른 관련 개념도 설명된다.

[0035] **II.용어/정의**

[0036] 본원에서 사용된 바와 같이:

[0037] 염기는 뉴클레오티드 염기 또는 뉴클레오티드, A(아데닌), C(시토신), T(티민), 또는 G(구아닌)를 지칭한다.

[0038] "단백질" 및 "변환된 서열"이란 용어는 상호 교환적으로 사용될 수 있다.

[0039] "코돈" 및 "염기 삼중체"란 용어는 상호 교환적으로 사용될 수 있다.

[0040] "아미노산" 및 "변환된 유닛"이란 용어는 상호 교환적으로 사용될 수 있다.

[0041] "변이체 병원성 분류기", "변이체 분류를 위한 컨벌루션 신경망 기반 분류기", 및 "변이체 분류를 위한 심층 컨벌루션 신경망 기반 분류기"라는 어구는 상호 교환적으로 사용될 수 있다.

[0042] **III.병원성 분류 신경망**

[0043] A. 훈련 및 입력

[0044] 예시적인 구현예로 돌아가면, 변이체 병원성 분류(예를 들어, 병원성 또는 양성) 및/또는 병원성 또는 병원성 부족을 수치적으로 특징짓는 정량적 메트릭(예를 들어, 병원성 점수)의 생성을 위해 사용될 수 있는 딥 러닝 네트워크가 본원에 설명된다. 양성 레이블이 있는 변이체만 사용하여 분류기를 훈련하는 하나의 컨텍스트에서, 예측 문제는 주어진 돌연변이가 모집단에서 일반 변이체로 관찰될 가능성이 있는지 여부로 구성되었다. 여러 요인이 높은 대립유전자 빈도에서 변이체를 관찰할 확률에 영향을 미치지만, 유해성은 본 논의 및 설명의 주요 초점이다. 다른 요인은 돌연변이율, 시퀀싱 커버리지와 같은 기술적 인공물, 및 중립적 유전적 드리프트(예를 들어, 유전자 전환)에 영향을 미치는 요인을 포함하지만, 이에 제한되지 않는다.

[0045] 딥 러닝 네트워크의 훈련과 관련하여, 임상 응용에 대한 변이체 분류의 중요성은 문제를 해결하기 위해 지도 기계 학습을 사용하려는 수많은 시도에 영감을 주었지만, 이러한 노력은 훈련을 위해 확실하게 레이블링된 양성 및 병원성 변이체를 포함하는 적절한 크기의 실측 데이터세트가 부족하여 방해를 받았다.

[0046] 인간 전문가 선별 변이체의 기존 데이터베이스는 전체 게놈을 나타내지 않으며, ClinVar 데이터베이스에서 변이체의 약 50%는 200개의 유전자(인간 단백질 코딩 유전자의 약 1%)에서만 나온다. 더욱이, 체계적인 연구는 많은 인간 전문가 주석이 의심스러운 지원 증거를 가지고 있음을 확인하여 단일 환자에서만 관찰될 수 있는 희귀 변이체를 해석하는 것이 어렵다는 점을 강조한다. 인간 전문가 해석이 점점 더 엄격해지고 있지만, 분류 지침은 대체로 합의 관행을 중심으로 공식화되며 기존 경향을 강화할 위험이 있다. 인간의 해석 편향을 줄이기 위해, 최근의 분류기는 일반적인 인간 다형성 또는 고정된 인간-침팬지 대체에 대해 훈련되었지만, 이러한 분류기는 또한 인간 선별 데이터베이스에서 훈련된 이전 분류기의 예측 점수를 입력으로 사용한다. 이러한 다양한 방법의 성능에 대한 객관적인 벤치마킹은 독립적이고 편향되지 않은 실측 데이터세트가 없기 때문에 파악하기 어렵다.

[0047] 이러한 문제를 해결하기 위해, 현재 설명된 기술은 일반적인 인간 변이와 겹치지 않고 정제 선택의 체를 통과한 양성 결과의 일반적인 변이체를 주로 나타내는 300,000개 초과의 고유한 미스센스 변이체를 제공하는 비인간 영장류(예를 들어, 침팬지, 보노보, 고릴라, 오랑우탄, 레수스, 및 마모셋)로부터의 변이를 활용한다. 이는 기계 학습 접근법에 이용할 수 있는 훈련 데이터세트를 크게 확장한다. 평균적으로, 각 영장류 종은 ClinVar 데이터베이스 전체보다 더 많은 변이체를 제공한다(2017년 11월 현재 약 42,000개의 미스센스 변이체, 불확실한 의미의 변이체 및 주석이 충돌하는 변이체를 제외). 추가적으로, 이러한 컨텐츠는 인간의 해석의 편견에서 자유롭다.

[0048] 본 기술에 따라 사용하기 위한 양성 훈련 데이터세트를 생성하는 것과 관련하여, 이러한 하나의 데이터세트는 기계 학습을 위한 인간 및 비인간 영장류로부터의 대체로 일반적인 양성 미스센스 변이체로 구성되었다. 데이터세트는 일반적인 인간 변이체(> 0.1% 대립유전자 빈도; 83,546개 변이체), 및 침팬지, 보노보, 고릴라, 오랑우탄, 레수스, 및 마모셋으로부터의 변이체(301,690개의 고유한 영장류 변이체)를 포함하였다.

[0049] 일반적인 인간 변이체(대립유전자 빈도(AF) > 0.1%) 및 영장류 변이를 포함하는 이러한 하나의 데이터세트를 사용하여, 본원에서 PrimateAI 또는 pAI로 지칭되는 심층 잔차 네트워크를 훈련했다. 네트워크는 관심 변이체 측면에 있는 아미노산 서열 및 다른 종의 직교 서열 정렬을 입력으로 수용하도록 훈련되었다. 인간 공학 특징을 이용하는 기존 분류기와 달리, 현재 설명된 딥 러닝 네트워크는 1차 서열로부터 직접 특징을 추출하도록 훈련되었다. 특정 구현예에서, 단백질 구조에 대한 정보를 통합하기 위해, 더 자세히 후술될 바와 같이, 별도의 네트워크가 서열 단독으로부터 2차 구조 및 용매 접근성을 예측하도록 훈련되었고, 이들은 전체 모델에서 서브네트

워크로 포함되었다. 성공적으로 결정화된 제한된 수의 인간 단백질을 고려하면, 1차 서열로부터 구조를 추론하는 것은 불완전한 단백질 구조 및 기능적 도메인 주석으로 인한 편향을 방지할 수 있는 이점이 있다. 단백질 구조가 포함된 네트워크의 일 구현예의 총 깊이는 대략 400,000개의 훈련 가능한 파라미터를 포함하는 36개의 컨벌루션 계층이였다.

[0050] B. 단백질 구조 2차 구조 및 용매 접근성 서브네트워크

[0051] 일 구현예의 한 예에서, 병원성 예측을 위한 딥 러닝 네트워크는, 2차 구조 및 용매 접근성 예측 서브네트워크에 대한 19개의 컨벌루션 계층 및 2차 구조 및 용매 접근성 서브네트워크의 결과를 입력으로 취하는 주요 병원성 예측 네트워크에 대한 17개를 포함하여, 총 36개의 컨벌루션 계층을 포함한다. 특히, 대부분의 인간 단백질의 결정 구조가 알려져 있지 않기 때문에, 2차 구조 네트워크 및 용매 접근성 예측 네트워크는 네트워크가 1차 서열로부터 단백질 구조를 학습할 수 있도록 훈련되었다.

[0052] 이러한 일 구현예에서 2차 구조 및 용매 접근성 예측 네트워크는 동일한 아키텍처 및 입력 데이터를 갖지만 예측 상태에 대해서는 다르다. 예를 들어, 이러한 일 구현예에서, 2차 구조 및 용매 접근성 네트워크에 대한 입력은 인간과 99종의 다른 척추동물의 다중 서열 정렬로부터의 보존 정보를 인코딩하는 적합한 차원의 아미노산 위치 빈도 매트릭스(PFM)(예를 들어, 51 길이 × 20 아미노산 PFM)이다.

[0053] 일 실시예에서, 도 2를 참조하면, 병원성 예측을 위한 딥 러닝 네트워크 및 2차 구조 및 용매 접근성 예측을 위한 딥 러닝 네트워크는 잔차 블록(140)의 아키텍처를 채택했다. 잔차 블록(140)은 이전 계층으로부터의 정보가 잔차 블록(140)을 건너뛰게 하는 스킵 연결(142)이 산재된 반복 컨벌루션 유닛을 포함한다. 각 잔차 블록(140)에서, 입력 계층이 먼저 배치 정규화된 다음, 정류 선형 유닛(ReLU)을 사용하는 활성화 계층이 이어진다. 그런 다음, 활성화는 1D 컨벌루션 계층을 통해 전달된다. 1D 컨벌루션 계층으로부터의 이러한 중간 출력은 다시 배치 정규화되고 ReLU가 활성화된 다음, 다른 1D 컨벌루션 계층이 이어진다. 제2 1D 컨벌루션의 끝에서, 그 출력은 원래의 입력 정보가 잔차 블록(140)을 우회하게 하여 스킵 연결(142)로 작용하는 잔차 블록으로의 원래 입력과 합산되었다(단계 146). 심층 잔차 학습 네트워크로 지칭될 수 있는 이러한 아키텍처에서, 입력은 원래 상태로 보존되고 잔차 연결은 모델로부터의 비선형 활성화 없이 유지되어, 더 깊은 네트워크의 효과적인 훈련을 가능하게 한다. 2차 구조 네트워크(130)와 용매 접근성 네트워크(132) 둘 모두의 컨텍스트에서의 상세한 아키텍처는 PWM 보존 데이터(150)가 초기 입력으로서 예시된 표 1 및 2(아래에서 논의됨) 및 도 2에 제공된다. 도시된 예에서, 모델에 대한 입력(150)은 (Protein Data Bank 서열에 대한 훈련용) RaptorX 소프트웨어 또는 (인간 단백질 서열에 대한 훈련 및 추론용) 99-척추동물 정렬에 의해 생성된 보존을 사용하는 위치 가중 매트릭스(PWM)일 수 있다.

[0054] 잔차 블록(140) 다음에, 소프트맥스 계층(154)이 각 아미노산에 대한 세 상태의 확률을 계산하며, 이 중에서 가장 큰 소프트맥스 확률이 아미노산의 상태를 결정한다. 이러한 일 구현예에서 모델은 ADAM 최적화기를 사용하여 전체 단백질 서열에 대해 누적된 범주 교차 엔트로피 손실 함수로 훈련된다. 예시된 일 구현예에서, 네트워크가 2차 구조 및 용매 접근성에 대해 사전 훈련되면, 병원성 예측 네트워크(160)에 대한 입력으로서 네트워크의 출력을 직접 취하는 대신에, 더 많은 정보가 병원성 예측 네트워크(160)로 전달되도록 소프트맥스 계층(154) 이전의 계층을 대신 취하였다. 일례에서, 소프트맥스 계층(154) 이전 계층의 출력은 적합한 길이(예를 들어, 51개의 아미노산 길이)의 아미노산 서열이고 병원성 분류를 위한 딥 러닝 네트워크에 대한 입력이 된다.

[0055] 이전 내용을 염두에 두고, 2차 구조 네트워크는 (1) 알파 나선(H), (2) 베타 시트(B), 또는 (3) 코일(C)의 3상태 2차 구조를 예측하도록 훈련된다. 용매 접근성 네트워크는 (1) 매립(B), (2) 중간(I), 또는 (3) 노출(E)의 3상태 용매 접근성을 예측하도록 훈련된다. 전술한 바와 같이, 둘 모두의 네트워크는 입력(150)으로 1차 서열만 사용하고 Protein DataBank에서 알려진 결정 구조로부터의 레이블을 사용하여 훈련되었다. 각 모델 모델은 각 아미노산 잔기에 대해 하나의 각각의 상태를 예측한다.

[0056] 이전 내용을 염두에 두고, 예시적인 구현예의 추가 예시를 통해, 입력 데이터세트(150) 내의 각 아미노산 위치에 대해, 위치 빈도 매트릭스로부터의 윈도우는 측면 아미노산(예를 들어, 측면 51개 아미노산)에 대응하게 취해지고, 이는 길이 아미노산 서열의 중심에 있는 아미노산에 대한 2차 구조 또는 용매 접근성에 대한 레이블을 예측하는 데 사용되었다. 2차 구조 및 상대 용매 접근성에 대한 레이블은 DSSP 소프트웨어를 사용하여 단백질의 알려진 3D 결정 구조로부터 직접 얻었고 1차 서열로부터 예측할 필요가 없었다. 병원성 예측 네트워크(160)의 일부로서 2차 구조 네트워크 및 용매 접근성 네트워크를 통합하기 위해, 인간 기반 99 척추동물 다중 서열 정렬로부터 위치 빈도 매트릭스를 계산하였다. 이러한 두 방법에서 생성된 보존 매트릭스는 일반적으로 유사하지만, 역전파는 파라미터 가중치의 미세 조정이 가능하도록 병원성 예측을 위한 훈련 중에 2차 구조 및 용매 접

근성 모델을 통해 가능해졌다.

[0057] 예를 들어, 표 1은 3상태 2차 구조 예측 딥 러닝(DL) 모델에 대한 예시적인 모델 아키텍처 세부사항을 보여준다. 형상은 모델의 각 계층에서 출력 텐서의 형상을 지정하고 활성화는 계층의 뉴런에 주어진 활성화이다. 모델에 대한 입력은 변이체 주변의 측면 아미노산 서열에 대한 적합한 차원(예를 들어, 51 아미노산 길이, 20 깊이)의 위치-특이적 빈도 매트릭스였다.

[0058] 유사하게, 표 2에 예시된 모델 아키텍처는 3상태 용매 접근성 예측 딥 러닝 모델에 대한 예시적인 모델 아키텍처 세부사항을 보여주며, 이는 본원에 언급된 바와 같이 아키텍처에서 2차 구조 예측 DL 모델과 동일할 수 있다. 형상은 모델의 각 계층에서 출력 텐서의 형상을 지정하고 활성화는 계층의 뉴런에 주어진 활성화이다. 모델에 대한 입력은 변이체 주변의 측면 아미노산 서열에 대한 적합한 차원(예를 들어, 51 아미노산 길이, 20 깊이)의 위치-특이적 빈도 매트릭스였다.

[0059] 3상태 2차 구조 예측 모델에 대한 최고의 테스트 정확도는 유사한 훈련 데이터세트에서 DeepCNF 모델이 예측한 최신 정확도와 유사한 80.32%였다. 3상태 용매 접근성 예측 모델에 대한 최고의 테스트 정확도는 유사한 훈련 데이터세트에서 RaptorX가 예측한 현재 최고의 정확도와 유사한 64.83%였다.

[0060] **예시적인 구현예 - 모델 아키텍처 및 훈련**

[0061] 표 1 및 표 2(아래에 재현됨) 및 도 2를 참조하고 일 구현예의 예시를 제공하는 방식으로, 단백질의 3상태 2차 구조 및 3상태 용매 접근성을 각각 예측하기 위해 2개의 종단간 심층 컨벌루션 신경망 모델을 훈련했다. 두 모델은 2개의 입력 채널(하나는 단백질 서열용이고 다른 하나는 단백질 보존 프로파일용임)을 포함하여 유사한 구성을 가졌다. 각 입력 채널의 차원은 L x 20이며, 여기서 L은 단백질의 길이를 나타낸다.

[0062] 각 입력 채널은 40개의 커널과 선형 활성화가 있는 1D 컨벌루션 계층(계층 1a 및 1b)을 통해 전달되었다. 이러한 계층은 입력 차원을 20에서 40으로 업샘플링하는 데 사용되었다. 모델 전체에 걸쳐 다른 모든 계층은 40개의 커널을 사용했다. 두 계층(1a 및 1b) 활성화는 40개 차원 각각에 걸친 값을 합산하여 함께 병합되었다(즉, 병합 모드 = '합계'). 병합 노드의 출력은 1D 컨벌루션의 단일 계층(계층 2)을 통해 전달된 후 선형 활성화되었다.

[0063] 계층 2로부터의 활성화는 일련의 9개의 잔차 블록(계층 3 내지 11)을 통해 전달되었다. 계층 3의 활성화는 계층 4에 공급되고 계층 4의 활성화는 계층 5에 공급되는 식으로 계속된다. 세 번째 잔차 블록(계층 5, 8 및 11)마다의 출력을 직접 합산하는 스킵 연결도 있었다. 그런 다음, 병합된 활성화는 ReLU 활성화와 함께 2개의 1D 컨벌루션(계층 12 및 13)에 공급되었다. 계층 13으로부터의 활성화는 소프트맥스 판독 계층에 제공되었다. 소프트맥스는 주어진 입력에 대한 3클래스 출력의 확률을 계산했다.

[0064] 추가적으로, 2차 구조 모델의 일 구현예에서, 1D 컨벌루션은 1의 아트러스 레이트(atrous rate)를 가졌다. 용매 접근성 모델의 구현예에서, 마지막 3개의 잔차 블록(계층 9, 10 및 11)은 커널의 커버리지를 증가시키기 위해 2의 아트러스 레이트를 가졌다. 이러한 양태와 관련하여, 아트러스/다이레이티드(atrous/dilated) 컨벌루션은 입력 값을 일정 단계로 건너뛰어 길이보다 큰 영역에 커널을 적용하는 컨벌루션으로, 아트러스 컨벌루션 레이트 또는 다이레이션 팩터(dilation factor)라고도 한다. 아트러스/다이레이티드 컨벌루션은 컨벌루션 필터/커널의 요소 사이에 간격을 추가하여 컨벌루션 작업을 수행할 때 더 큰 간격의 인접 입력 항목(예를 들어, 뉴클레오티드, 아미노산)을 고려한다. 이를 통해 입력에 장기적인 컨텍스트 종속성을 통합할 수 있다. 아트러스 컨벌루션은 인접한 뉴클레오티드가 처리될 때 재사용을 위해 부분적인 컨벌루션 계산을 보존한다. 아트러스/다이레이티드 컨벌루션을 통해 훈련 가능한 파라미터가 거의 없는 큰 수용 필드가 가능하다. 단백질의 2차 구조는 근접한 아미노산의 상호 작용에 따라 크게 달라진다. 따라서, 커널 커버리지가 더 높은 모델은 성능을 약간 향상시켰다. 반대로, 용매 접근성은 아미노산 간의 장거리 상호 작용에 의해 영향을 받는다. 그러므로, 아트러스 컨벌루션을 사용하는 커널의 커버리지가 높은 모델의 경우, 그 정확도는 짧은 커버리지 모델보다 2% 이상 더 높았다.

[0065] [표 1]

3 상태 2 차 구조 예측 모델의 예

계층	유형	커널 수, 윈도우 크기	형상	아트러스 레이트	활성화
입력 서열(계층 1a)	컨벌루션 1D	40,1	(L,40)	1	선형
입력 PSSM(계층 1b)	컨벌루션 1D	40,1	(L,40)	1	선형
병합 서열 + PSSM	병합(모드 = 연결)	-	(L,80)	-	-
계층 2	컨벌루션 1D	40,5	(L,40)	1	선형
계층 3	컨벌루션 1D	40,5	(L,40)	1	ReLU
계층 4	컨벌루션 1D	40,5	(L,40)	1	ReLU
계층 5	컨벌루션 1D	40,5	(L,40)	1	ReLU
계층 6	컨벌루션 1D	40,5	(L,40)	1	ReLU
계층 7	컨벌루션 1D	40,5	(L,40)	1	ReLU
계층 8	컨벌루션 1D	40,5	(L,40)	1	ReLU
계층 9	컨벌루션 1D	40,5	(L,40)	1	ReLU
계층 10	컨벌루션 1D	40,5	(L,40)	1	ReLU
계층 11	컨벌루션 1D	40,5	(L,40)	1	ReLU
활성화 병합	병합 - 계층 5, 8 및 11,모드='합계'	-	(L,40)	-	-
계층 12	컨벌루션 1D	40,5	(L,40)	1	ReLU
계층 13	컨벌루션 1D	40,5	(L,40)	1	ReLU
출력 계층	컨벌루션 1D	1,3	(L,3)	-	소프트맥스

[0066]

[0067] [표 2]

3 상태 용매 접근성 모델의 예

계층	유형	커널 수, 윈도우 크기	형상	아트러스 레이트	활성화
입력 서열(계층 1a)	컨벌루션 1D	40,1	(L,40)	1	선형
입력 PSSM(계층 1b)	컨벌루션 1D	40,1	(L,40)	1	선형
병합 서열 + PSSM	병합(모드 = 연결)	-	(L,80)	-	-
계층 2	컨벌루션 1D	40,5	(L,40)	1	선형
계층 3	컨벌루션 1D	40,5	(L,40)	1	ReLU
계층 4	컨벌루션 1D	40,5	(L,40)	1	ReLU
계층 5	컨벌루션 1D	40,5	(L,40)	1	ReLU
계층 6	컨벌루션 1D	40,5	(L,40)	1	ReLU
계층 7	컨벌루션 1D	40,5	(L,40)	1	ReLU
계층 8	컨벌루션 1D	40,5	(L,40)	1	ReLU
계층 9	컨벌루션 1D	40,5	(L,40)	2	ReLU
계층 10	컨벌루션 1D	40,5	(L,40)	2	ReLU
계층 11	컨벌루션 1D	40,5	(L,40)	2	ReLU
활성화 병합	병합 - 계층 5, 8 및 11,모드='합계'	-	(L,40)	-	-
계층 12	컨벌루션 1D	40,5	(L,40)	1	ReLU
계층 13	컨벌루션 1D	40,5	(L,40)	1	ReLU
출력 계층	컨벌루션 1D	1,3	(L,3)	-	소프트맥스

[0068]

[0069] C. 병원성 예측 네트워크 아키텍처

[0070] 병원성 예측 모델과 관련하여, 변이체의 병원성을 예측하기 위해 준지도 심층 컨벌루션 신경망(CNN) 모델을 개

발하였다. 모델에 대한 입력 특징으로는 특정 유전자 영역에서 변이체 측면에 있는 단백질 서열 및 보존 프로파일 및 미스센스 변이체의 고갈을 포함한다. 2차 구조 및 용매 접근성에 대한 변이체로 인한 변화도 딥 러닝 모델에 의해 예측되었으며 이는 병원성 예측 모델에 통합되었다. 이러한 일 구현예에서 예측된 병원성은 0(양성)에서 1(병원성)까지의 규모이다.

[0071] 이러한 하나의 병원성 분류 신경망(예를 들어, PrimateAI)에 대한 아키텍처는 도 3 및 보다 상세한 예에서는 표 3(아래)에 개략적으로 설명되어 있다. 도 3에 도시된 예에서, 1D는 1차원 컨벌루션 계층을 지칭한다. 다른 구현예에서, 모델은 2D 컨벌루션, 3D 컨벌루션, 다이레이티드 또는 아트러스 컨벌루션, 트랜스포즈형 컨벌루션, 분리식 컨벌루션, 깊이별 분리식 컨벌루션 등과 같은 다양한 유형의 컨벌루션을 사용할 수 있다. 더 나아가, 전술한 바와 같이, 병원성 예측을 위한 딥 러닝 네트워크(예를 들어, PrimateAI 또는 pAI)와 2차 구조 및 용매 접근성을 예측하기 위한 딥 러닝 네트워크 둘 모두의 특정 구현예는 잔차 블록의 아키텍처를 채택했다.

[0072] 특정 실시예에서, 심층 잔차 네트워크의 일부 또는 모든 계층은 ReLU 활성화 함수를 사용하는데, 이는 시그모이드 또는 하이퍼볼릭 탄젠트와 같은 포화 비선형성과 비교하여 확률적 경사 하강법의 수렴을 크게 가속화시킨다. 개시된 기술에 의해 사용될 수 있는 활성화 함수의 다른 예로는 파라메트릭 ReLU, 리키 ReLU, 및 지수 선형 유닛(ELU)을 포함한다.

[0073] 본원에 설명된 바와 같이, 일부 또는 모든 계층은 훈련 중에 컨벌루션 신경망(CNN)의 각 계층의 분포가 변경되고 계층마다 달라지는 배치 정규화를 이용할 수도 있다. 이는 최적화 알고리즘의 수렴 속도를 감소시킨다.

[0074] [표 3]

일 구현예에 따른 병원성 예측 모델

계층	유형	커널 수, 윈도우 크기	형상	아트러스 레이트	활성화
참조 서열(1a)	컨벌루션 1D	40,1	(51,40)	1	선형
대체 서열(1b)	컨벌루션 1D	40,1	(51,40)	1	선형
영장류 보존(1c)	컨벌루션 1D	40,1	(51,40)	1	선형
포유류 보존(1d)	컨벌루션 1D	40,1	(51,40)	1	선형
척추동물 보존(1e)	컨벌루션 1D	40,1	(51, 40)	1	선형
참조 서열 기반 2 차 구조(1f)	입력 계층	-	(51,40)	-	-
대체 서열 기반 2 차 구조(1g)	입력 계층	-	(51,40)	-	-
참조 서열 기반 용매 접근성(1h)	입력 계층	-	(51,40)	-	-
대체 서열 기반 용매 접근성(1i)	입력 계층	-	(51,40)	-	-
참조 서열 병합(2a)	병합 (모드 = 합계) (1a,1c,1d,1e)	-	(51,40)	-	-
대체 서열 병합(2b)	병합 (모드 = 합계) (1b,1c,1d,1e)	-	(51,40)	-	-
3a	컨벌루션 1D (2a)	40,5	(51,40)	1	ReLU
3b	컨벌루션 1D (2b)	40,5	(51,40)	1	ReLU
4	병합(모드 = 연결) (3a,3b,1f,1g,1h,1i)	-	(51,240)	-	-
5	컨벌루션 1D	40,5	(51,40)	1	ReLU
6	컨벌루션 1D	40,5	(51,40)	1	ReLU
7	컨벌루션 1D	40,5	(51,40)	1	ReLU
8	컨벌루션 1D	40,5	(51,40)	1	ReLU
9	컨벌루션 1D	40,5	(51,40)	2	ReLU
10	컨벌루션 1D	40,5	(51,40)	2	ReLU
11	컨벌루션 1D	40,1	(51,1)	2	시그모이드
12	전역 최대 풀링	-	1	-	-
출력 계층	활성화 계층	-	1	-	시그모이드

[0075]

[0076] **예시적인 구현예 - 모델 아키텍처**

[0077] 이전 내용을 염두에 두고, 도 3 및 표 3을 참조하면, 일 구현에서, 병원성 예측 네트워크는 5개의 직접 입력 및 4개의 간접 입력을 수신한다. 이러한 예에서 5개의 직접 입력은 적합한 차원의 아미노산 서열(예를 들어, 51-길이 아미노산 서열 × 20-깊이)(20개의 상이한 아미노산을 인코딩)을 포함할 수 있고, 변이체가 없는 참조 인간 아미노산 서열(1a), 변이체가 치환된 대체 인간 아미노산 서열(1b), 영장류 종의 다중 서열 정렬로부터의 위치-특이적 빈도 메트릭스(PFM)(1c), 포유류 종의 다중 서열 정렬로부터의 PFM(1d), 및 더 먼 척추동물 종의 다중 서열 정렬로부터의 PFM(1e)을 포함할 수 있다. 간접 입력으로는 참조 서열 기반 2차 구조(1f), 대체 서열 기반 2차 구조(1g), 참조 서열 기반 용매 접근성(1h), 및 대체 서열 기반 용매 접근성(1i)을 포함한다.

[0078] 간접 입력 1f 및 1g의 경우, 소프트맥스 계층을 제외하고 2차 구조 예측 모델의 사전 훈련된 계층이 로딩된다. 입력 1f의 경우, 사전 훈련된 계층은 변이체에 대해 PSI-BLAST에 의해 생성된 PSSM과 함께 변이체에 대한 인간 참조 서열에 기초한다. 마찬가지로, 입력 1g의 경우, 2차 구조 예측 모델의 사전 훈련된 계층은 PSSM 매트릭스와 함께 입력으로 인간 대체 서열에 기초한다. 입력 1h 및 1i는 각각 변이체의 참조 및 대체 서열에 대한 용매 접근성 정보를 포함하는 유사한 사전 훈련된 채널에 대응한다.

[0079] 이러한 예에서, 5개의 직접 입력 채널은 선형 활성화가 있는 40개 커널의 업샘플링 컨벌루션 계층을 통해 전달된다. 계층 1a, 1c, 1d 및 1e는 계층 2a를 생성하기 위해 40개의 특징 차원에 걸쳐 합산된 값과 병합된다. 다시 말해서, 참조 서열의 특징 맵은 세 가지 유형의 보존 특징 맵과 병합된다. 유사하게, 1b, 1c, 1d 및 1e는 계층 2b를 생성하기 위해 40개의 특징 차원에 걸쳐 합산된 값과 병합된다, 즉 대체 서열의 특징이 세 가지 유형의 보존 특징과 병합된다.

[0080] 계층 2a 및 2b는 ReLU의 활성화로 배치 정규화되고 각각은 필터 크기 40(3a 및 3b)의 1D 컨벌루션 계층을 통과한다. 계층 3a 및 3b의 출력은 서로 연결된 특징 맵과 함께 1f, 1g, 1h 및 1i와 병합된다. 다시 말해서, 보존 프로파일이 있는 참조 서열의 특징 맵 및 보존 프로파일이 있는 대체 서열은 참조 및 대체 서열의 2차 구조 특징 맵 및 참조 및 대체 서열의 용매 접근성 특징 맵과 병합된다(계층 4).

[0081] 계층 4의 출력은 6개의 잔차 블록(계층 5, 6, 7, 8, 9, 10)을 통해 전달된다. 마지막 3개의 잔차 블록은 커널에 더 높은 커버리지를 제공하도록 1D 컨벌루션에 대해 2의 아트러스 레이트를 갖는다. 계층 10의 출력은 필터 크기 1 및 활성화 시그모이드(계층 11)의 1D 컨벌루션을 통해 전달된다. 계층 11의 출력은 변이체에 대한 단일 값을 선택하는 전역 최대 풀을 통해 전달된다. 이 값은 변이체의 병원성을 나타낸다. 병원성 예측 모델의 일 구현예에 대한 세부사항은 표 3에 나와 있다.

[0082] D. 훈련(준지도) 및 데이터 분포

[0083] 준지도 학습 접근법과 관련하여, 이러한 기술은 네트워크(들)를 훈련하기 위해 레이블링된 데이터와 레이블링되지 않은 데이터 둘 모두를 이용 가능하게 한다. 준지도 학습을 선택하는 동기는 인간 선별 변이 데이터베이스가 신뢰할 수 없고 노이즈가 많으며, 특히 신뢰할 수 있는 병원성 변이체가 부족하다는 점이다. 준지도 학습 알고리즘은 훈련 과정에서 레이블링된 인스턴스와 레이블링되지 않은 인스턴스 둘 모두를 사용하기 때문에, 훈련에 이용할 수 있는 소량의 레이블링된 데이터만을 갖는 완전 지도 학습 알고리즘보다 더 나은 성능을 달성하는 분류기를 생성할 수 있다. 준지도 학습의 기본 원리는 레이블링된 인스턴스만 사용하는 지도 모델의 예측 역량을 강화하기 위해 레이블링되지 않은 데이터 내의 고유한 지식을 활용하여 준지도 학습에 잠재적인 이점을 제공할 수 있다는 점이다. 소량의 레이블링된 데이터로부터 지도 분류기에 의해 학습된 모델 파라미터는 레이블링되지 않은 데이터에 의해 보다 현실적인 분포(테스트 데이터의 분포와 더 유사함)로 조향될 수 있다.

[0084] 생물 정보학에서 널리 퍼진 또 다른 문제는 데이터 불균형 문제이다. 데이터 불균형 현상은 예측될 클래스 중 하나에 속하는 인스턴스가 드물거나(주목할 만한 경우) 얻기 어렵기 때문에 해당 클래스가 데이터에서 과소 표현될 때 발생한다. 소수 클래스는 통상적으로 특별한 경우와 연관될 수 있기 때문에 학습에 가장 중요하다.

[0085] 불균형 데이터 분포를 처리하는 알고리즘 접근법은 분류기의 앙상블에 기초한다. 제한된 양의 레이블링된 데이터는 자연스럽게 약한 분류기로 이어지지만, 약한 분류기의 앙상블은 단일 구성 분류기의 성능을 능가하는 경향이 있다. 더욱이, 앙상블은 통상적으로 다수의 모델 학습과 연관된 노력과 비용을 검증하는 요인으로 단일 분류기에서 얻은 예측 정확도를 향상시킨다. 개별 분류기의 높은 변동성을 평균화하면 분류기의 오버피팅도 평균화되므로, 직관적으로 여러 분류기를 집계하면 다 나은 오버피팅 제어를 발생시킨다.

[0086] **IV.유전자-특이적 병원성 점수 임계치**

[0087] 이전 내용은 신경망으로 구현된 병원성 분류기의 훈련 및 검증에 관한 것이지만, 다음 섹션은 이러한 네트워크

를 사용하여 병원성 분류를 추가로 개선 및/또는 활용하기 위한 다양한 구현예-특이적 및 사용 사례 시나리오와 관련된다. 제1 양태에서, 임계치 채점 및 이러한 점수 임계치의 사용에 대한 논의를 설명한다.

[0088] 본원에 언급된 바와 같이, 본원에 설명된 PrimateAI 또는 pAI 분류기와 같은(그러나, 이에 제한되지 않음) 본 개시된 병원성 분류 네트워크는 유전자 내의 양성 변이체로부터 병원성 변이체를 구별하거나 가려내는 데 유용한 병원성 점수를 생성하기 위해 사용될 수 있다. 본원에 설명된 바와 같은 병원성 채점은 인간 및 비인간 영장류에서 정제 선택의 정도에 기초하기 때문에, 병원성 및 양성 변이체와 연관된 병원성 점수는 강력한 정제 선택 하에 있는 유전자에서 더 높을 것으로 예상된다. 반면에, 중립 진화 또는 약한 선택 하에 있는 유전자의 경우, 병원성 변이체에 대한 병원성 점수가 더 낮은 경향이 있다. 이러한 개념은 변이체에 대한 병원성 점수(206)가 각각의 유전자에 대한 점수 분포 내에 예시되어 있는 도 4에 시각적으로 도시되어 있다. 도 4를 참조하여 알 수 있는 바와 같이, 실제로 병원성 또는 양성일 가능성이 있는 변이체를 식별하기 위한 대략적인 유전자-특이적 임계치(들)를 갖는 것이 유용할 수 있다.

[0089] 병원성 점수를 평가하는 데 유용할 수 있는 가능한 임계치를 평가하기 위해, ClinVar에서 적어도 10개의 양성/가능성이 있는 양성 변이체 및 적어도 10개의 병원성 및 가능성 있는 병원성 변이체를 포함하는 84개의 유전자를 사용하여 잠재적 점수 임계치를 연구했다. 이러한 유전자는 각 유전자에 대한 적합한 병원성 점수 임계치를 평가하는 데 도움을 주기 위해 사용되었다. 이러한 유전자 각각의 경우, ClinVar에서 양성 및 병원성 변이체에 대해 평균 병원성 점수를 측정했다.

[0090] 일 구현예에서, 병원성 변이체에 대한 유전자-특이적 PrimateAI 임계치를 나타낸 도 5 및 양성 변이체에 대한 유전자-특이적 PrimateAI 임계치를 나타낸 도 6에 그래프로 도시된 바와 같이, 각 유전자에서 병원성 및 양성 ClinVar 변이체에 대한 평균 병원성 점수는 해당 유전자에서 병원성 점수(여기서는 PrimateAI 또는 pAI 점수)의 75번째 및 25번째 백분위수와 매우 상관 관계가 있는 것으로 관찰되었다. 두 도면에서, 각 유전자는 위의 유전자 심볼 레이블이 있는 점으로 표시된다. 이러한 예에서, ClinVar 병원성 변이체에 대한 평균 PrimateAI 점수는 해당 유전자에서 모든 미스센스 변이체의 75번째 백분위수에서 PrimateAI 점수와 밀접한 상관 관계가 있었다(스피어만 상관 관계 = 0.8521, 도 5). 마찬가지로, ClinVar 양성 변이체에 대한 평균 PrimateAI 점수는 해당 유전자에서 모든 미스센스 변이체의 25번째 백분위수에서 PrimateAI 점수와 밀접한 상관 관계가 있었다(스피어만 상관 관계 = 0.8703, 도 6).

[0091] 본 접근법을 고려할 때, 가능성 있는 병원성 변이체에 대한 컷오프로서 사용하기 위한 유전자당 병원성 점수의 적합한 백분위수는 51번째 백분위수 내지 99번째 백분위수(예를 들어, 65번째, 70번째, 75번째, 80번째, 또는 85번째 백분위수)에 의해 정의되고 이를 포함하는 범위에 있을 수 있다. 반대로, 가능성이 있는 양성 변이체에 대한 컷오프로 사용하기 위한 유전자당 병원성 점수의 적합한 백분위수는 1번째 백분위수 내지 49번째 백분위수(예를 들어, 15번째, 20번째, 25번째, 30번째, 또는 35번째 백분위수)에 의해 정의되고 이를 포함하는 범위에 있을 수 있다.

[0092] 이러한 임계치의 사용과 관련하여, 도 7은 병원성 점수(206)에 기초하여 변이체를 양성 또는 병원성 범주로 분류하기 위해 이러한 임계치가 사용될 수 있는 샘플 프로세스 흐름을 도시하고 있다. 이러한 예에서, 관심 변이체(200)는 관심 변이체(200)에 대한 병원성 점수(206)를 도출하기 위해 본원에 설명된 바와 같은 병원성 채점 신경망을 사용하여 처리될 수 있다(단계 202). 도시된 예에서, 병원성 점수는 유전자-특이적 병원성 임계치(212)(예를 들어, 75%)와 비교되고(결정 블록 210), 병원성으로 결정되지 않으면 유전자-특이적 양성 임계치(218)와 비교된다(결정 블록 216). 이 예에서 비교 프로세스는 단순화를 위해 직렬로 발생하는 것으로 도시되어 있지만, 실제로 비교는 단일 단계에서 병렬로 수행될 수 있거나 또는 대안적으로 비교 중 하나만이 수행될 수 있다(예를 들어, 변이체가 병원성인지 여부를 결정). 병원성 임계치(212)가 초과되면, 관심 변이체(200)는 병원성 변이체(220)로 간주될 수 있는 반면, 반대로 병원성 점수(206)가 양성 임계치(212) 미만이면, 관심 변이체(200)는 양성 변이체(222)로 간주될 수 있다. 임계치 기준이 모두 충족되지 않으면, 관심 변이체는 병원성도 양성도 아닌 것으로 취급될 수 있다. 한 연구에서, ClinVar 데이터베이스 내에서 17,948개의 고유한 유전자에 대해 본원에 설명된 접근법을 사용하여 유전자-특이적 임계치 및 메트릭을 도출하고 평가했다.

[0093] **V.순방향 시간 시뮬레이션을 사용하여 병원성 점수를 기반으로 모든 인간 변이체에 대한 선택 효과 추정**

[0094] 임상 연구 및 환자 치료는 PrimateAI와 같은 병원성 분류 네트워크를 이용하여 유전자 내 양성 변이체로부터 병원성 변이체를 분류 및/또는 분리하는 사용 사례 시나리오의 예이다. 특히, 임상 게놈 시퀀싱은 희귀 유전병 환자를 위한 표준 치료가 되었다. 희귀 유전병은, 주로는 아니지만, 매우 유해한 희귀 돌연변이에 의해 발생하는 경우가 많으며, 이는 일반적으로 중증도로 인해 발견하기가 더 쉽다. 그러나, 일반적인 유전병의 기저를 이

루는 희귀 돌연변이는 약한 영향과 많은 수로 인해 대체로 특성화되지 않은 상태로 남아 있다.

[0095] 이를 염두에 두고, 희귀 돌연변이와 일반적인 질병 사이의 메커니즘을 이해하고 특히 본원에서 논의된 바와 같이 변이체의 병원성 채점과 관련하여 인간 돌연변이의 진화 역학을 연구하는 것이 바람직할 수 있다. 인간 모집단의 진화 과정에서, 새로운 변이체가 *드노보* 돌연변이에 의해 끊임없이 생성되었지만, 그 중 일부는 자연 선택으로 인해 또한 제거되었다. 인간 모집단 규모가 일정하다면, 두 힘에 의해 영향을 받는 변이체의 대립유전자 빈도는 궁극적으로 평형에 도달할 것이다. 이를 염두에 두고, 관측된 대립유전자 빈도를 사용하여 임의의 변이체에 대한 자연 선택의 중증도를 결정하는 것이 바람직할 수 있다.

[0096] 그러나, 인간 모집단은 어느 순간에도 안정적 상태에 있지 않고 대신 농업의 출현 이후 기하급수적으로 증가하고 있다. 그러므로, 본원에서 논의된 특정 접근법에 따르면, 전방 시간 시뮬레이션은 변이체의 대립유전자 빈도 분포에 대한 두 힘의 효과를 조사하기 위한 도구로 사용될 수 있다. 이러한 접근법의 양태는 최적의 순방향 시간 모델 파라미터를 도출하는 것으로 참조되고 반환될 수 있는 도 8에 도시된 단계와 관련하여 설명되며 논의된다.

[0097] 이를 염두에 두고, *드노보* 돌연변이율(280)을 사용하는 중립 진화의 순방향 시간 시뮬레이션은 시간 경과에 따른 변이체의 대립유전자 빈도 분포를 모델링(단계 282)하는 일부로서 이용될 수 있다. 기준선으로서, 중립 진화를 가정하여 순방향 시간 모집단 모델을 시뮬레이션할 수 있다. 모델 파라미터(300)는 시뮬레이션된 대립유전자 빈도 스펙트럼(AFS)(304)을 인간 게놈에서 관측된 동의 돌연변이(동의 AFS(308))에 피팅함으로써(단계 302) 도출되었다. 그런 다음, 최적의 모델 파라미터(300) 세트(즉, 최상의 피팅에 해당하는 파라미터)를 사용하여 생성된 시뮬레이션된 AFS(304)는 유용한 임상 정보를 도출하기 위해 변이체 병원성 채점과 같은 본원에서 논의된 다른 개념과 함께 사용될 수 있다.

[0098] 희귀 변이체의 분포가 주요 관심사이므로, 본 예시적인 구현예에서 인간 모집단의 진화 이력은 단순화된 인간 모집단 확장 모델(즉, 단순화된 진화 이력(278))의 개략도인 도 9에 도시된 바와 같이 이러한 시뮬레이션에서 상이한 성장률을 갖는 4개의 기하급수적 확장 단계로 단순화된다. 이 예에서, 인구 조사 모집단 규모와 유효 모집단 규모 간의 비율은 *r*로 표시될 수 있고 초기 유효 모집단 규모는 *Ne0* = 10,000으로 표시될 수 있다. 각 세대는 약 30년이 걸린다고 가정할 수 있다.

[0099] 이 예에서, 유효 모집단 규모의 변화가 적은 긴 번인 기간(약 3,500세대)이 제1 단계에서 이용되었다. 모집단 규모 변화는 *n*으로 표시될 수 있다. 번인 후 시간은 알 수 없으므로, 이 시간은 *T1*으로 표시될 수 있고 유효 모집단 규모는 *T1*에서 10,000*n으로 표시될 수 있다. 번인 동안 성장률(284)은 g1=n^(1/3,500)이다.

[0100] 서기 1400년에, 전 세계의 인구 조사 모집단 규모는 약 3억 6천만 명으로 추정된다. 서기 1700년에, 인구 조사 모집단 규모는 약 6억 2천만 명으로 증가했으며, 서기 2000년에는 62억 명이다. 이러한 추정치에 기초하여, 각 단계에서의 성장률(284)은 표 4에 나타낸 바와 같이 도출될 수 있다:

[0101] [표 4]

시뮬레이션 계획

시점	과거 세대 수	유효 모집단 규모	인구 조사 모집단 규모	성장률
T1	대규모 번인 (예를 들어, 3500)	10,000 * n	10,000 * r * n	n^(1/3500)
T2: 1400 AD	(T2-T1)/30	36000 / r	360m	(3.6/r/n) ^ ((T2-T1)/30)
T3: 1700 AD	10	62000 / r	620m	(6.2/3.6)^(1/10)
T4: 2000 AD	10	620000 / r	6.2b	10^(1/10)

[0102]

[0103] 세대 *j*(286)의 경우, N_j 염색체는 이전 세대로부터 무작위로 샘플링되어 새로운 세대 모집단을 형성하며, 여기서 $N_j = g_j * N_{j-1}$이고, g_j는 세대 *j*에서의 성장률(284)이다. 대부분의 돌연변이는 염색체 샘플링 중에 이전 세대로부터 유전된다. 그런 다음, *드노보* 돌연변이가 *드노보* 돌연변이율(μ)(280)에 따라 이러한 염색체에 적용된다.

[0104] *드노보* 돌연변이율(280)과 관련하여, 특정 구현에 따르면, 이들은 다음 접근법 또는 동등한 접근법에 따라 도출

될 수 있다. 특히, 이러한 일 구현예에서, 문헌 자료(Halldorsson 세트(2976개 트리오), Goldmann 세트(1291개 트리오) 및 Sanders 세트(3804개 트리오))로부터 전체 게놈 시퀀싱으로 총 8,071개 트리오에 이르는 3개의 큰 부모-자손 트리오 데이터세트를 얻었다. 이러한 8,071개 트리오를 병합하여, 유전자간 영역에 매핑된 드노보 돌연변이가 얻어졌고 192개의 트리뉴클레오티드 컨텍스트 구성 각각에 대해 *드노보* 돌연변이율(280)이 도출되었다.

[0105] 이러한 돌연변이율의 추정치를 도 10에 도시된 바와 같이 다른 문헌 돌연변이율(1,000개 게놈 프로젝트의 유전자간 영역으로부터 도출된 Kaitlin의 돌연변이율)과 비교하였다. 상관 관계는 0.9991이었으며, 현재 추정치는 표 5(CpGTi = CpG 부위에서의 전이 돌연변이, 비-CpGTi = 비-CpG 부위에서의 전이 돌연변이, Tv = 전환 돌연변이)에 나타내 바와 같이 일반적으로 케이틀린의 돌연변이율보다 낮다.

[0106] [표 5]

돌연변이율 비교

돌연변이율	Kaitlin 의 돌연변이율			드노보 돌연변이		
	가중 평균	평균	중앙값	가중 평균	평균	중앙값
게놈 전체	1.2e-8			1.048189e-08		
NonCpGTi (56 개 트리누크)	6.634517e-09	6.700049e-09	6.449612e-09	5.493724e-09	5.552238e-09	5.339421e-09
Tv	2.101242e-09	2.345956e-09	1.818607e-09	1.922048e-09	2.034197e-09	1.663312e-09
CpGTi	9.33749e-08	9.486598e-08	8.917675e-08	9.040611e-08	9.174335e-08	8.593169e-08
CpGTi(고메틸화)					1.011751e-07	
CpGTi(저메틸화)					2.264356e-08	

[0107]

[0108] CpG 섬에서의 돌연변이율과 관련하여, CpG 부위에서의 메틸화 수준은 돌연변이율에 상당한 영향을 미친다. CpGTi 돌연변이율을 정확하게 계산하기 위해, 해당 부위에서의 메틸화 수준을 고려해야 한다. 이를 염두에 두고, 예시적인 구현예에서, 돌연변이율 및 CpG 섬은 다음 접근법에 따라 계산될 수 있다.

[0109] 먼저, CpG 돌연변이율에 대한 메틸화 수준의 영향을 전체 게놈 중아황산염 시퀀싱 데이터(Roadmap Epigenomics 프로젝트에서 얻음)를 사용하여 평가했다. 각 CpG 섬에 대한 메틸화 데이터를 추출하여 10개의 배아 줄기 세포(ESC) 샘플에 걸쳐 평균화했다. 그런 다음, 도 11에 도시된 바와 같이 해당 CpG 섬을 10개의 정의된 메틸화 수준에 기초하여 10개의 빈으로 분리하였다. 각각 유전자간 영역과 엑손 영역 둘 모두에서 각 메틸화 빈에 속하는 CpG 부위의 수 및 관측된 CpG 전이 변이체의 수를 카운팅했다. 각 메틸화 빈의 CpG 부위에서 예상되는 전이 변이체 수는 CpGTi 변이체의 총 수에 해당 메틸화 빈의 CpG 부위의 분율을 곱한 값으로 계산되었다. 도 11에 도시된 바와 같이, CpG 돌연변이의 예상된 수에 대한 관측된 수의 비율이 메틸화 수준에 따라 증가하였고 고메틸화 수준과 저메틸화 수준 사이에서 CpGTi 돌연변이의 관측된/예상된 수의 비율에서 약 5배의 변화가 있었던 것으로 관찰되었다.

[0110] CpG 부위는 두 유형으로 분류되었다: (1) 고메틸화(평균 메틸화 수준 > 0.5인 경우); 및 (2) 저메틸화(평균 메틸화 수준 ≤ 0.5인 경우). 8개의 CpGTi 트리-뉴클레오티드 컨텍스트 각각에 대한 *드노보* 돌연변이율은 고메틸화 수준과 저메틸화 수준에 대해 각각 계산되었다. 8개의 CpGTi 트리-뉴클레오티드 컨텍스트에 걸쳐 평균화하여, CpGTi 돌연변이율을 얻었다: 표 6에 나타낸 바와 같이 고메틸화의 경우 1.01e-07 및 저메틸화의 경우 2.264e-08.

[0111] 그런 다음, 엑솜 시퀀싱 데이터의 대립유전자 빈도 스펙트럼(AFS)이 피팅되었다. 이러한 일 샘플 구현예에서, 시뮬레이션은 100,000개의 독립 부위를 가정하고 *T1*, *r* 및 *n*의 다양한 파라미터 조합을 사용하여 수행되었으며, 여기서 *T1* ∈(330, 350, 370, 400, 430, 450, 470, 500, 530, 550), r ∈(20, 25, 30,…, 100, 105, 110), 및 n ∈ (1.0, 2.0, 3.0, 4.0, 5.0)을 고려하였다.

[0112] 돌연변이의 세 가지 주요 클래스 각각은 서로 다른 *드노보* 돌연변이율을, 즉 CpGTi, non-CpGTi 및 Tv(표 6에 나타낸 바와 같음)를, 사용하여 별도로 시뮬레이션되었다. CpGTi의 경우, 고메틸화 수준과 저메틸화 수준을 별도로 시뮬레이션하고 두 AFS를 병합하여 고메틸화 부위 또는 저메틸화 부위의 비율을 가중치로 적용했다.

[0113] 파라미터의 각 조합 및 각 돌연변이율에 대해, 현재까지 인간 모집단을 시뮬레이션했다. 그런 다음, gnomAD 엑솜의 샘플 크기에 대응하는 약 246,000개 염색체의 1000개 세트를 (예를 들어, 목표 또는 최종 세대(290)로부터) 무작위로 샘플링하였다(단계 288). 그런 다음, 시뮬레이션된 AFS(304)는 1000개의 각각의 샘플 세트(294)에 걸쳐 평균화(단계 292)하여 생성되었다.

[0114] 검증 측면에서, 전 세계의 8개 하위 모집단으로부터 123,136명 개체의 전체 엑솜 시퀀싱(WES) 데이터를 수집한 게놈 집계 데이터베이스(gnomAD) v2.1.1로부터 인간 엑솜 다형성 데이터를 획득했다(http://gnomad.broadinstitute .org/).필터를 통과하지 않았거나, 중앙값 커버리지 < 15이거나, 또는 낮은 복잡도 영역이나 세그먼트 복제 영역에 속하는 변이체는 제외되었으며, 영역 경계는 https://storage.googleapis.com/gnomad-public/release/2.0.2/README.txt)로부터 다운로딩된 파일에서 정의되었다. hg19 빌드에 대한 UCSC 게놈 브라우저에 의해 정의된 표준 코딩 시퀀스에 매핑된 변이체를 유지했다.

[0115] gnomAD의 동의 대립유전자 빈도 스펙트럼(308)은 싱글톤, 더블톤, 3 ≤ 대립유전자 카운트(AC) ≤ 4, … 및 33 ≤ AC ≤ 64를 포함한 7개의 대립유전자 빈도 범주에서 동의 변이체의 수를 카운팅하여 생성되었다(단계 306). 희귀 변이체에 초점이 맞춰졌기 때문에 AC > 64인 변이체는 폐기되었다.

[0116] 3개의 돌연변이 클래스에 걸친 희귀 동의 변이체의 gnomAD AFS(즉, 동의 AFS(308))에 대한 시뮬레이션된 AFS(304)의 피팅을 평가(단계 302)하기 위해 우도비 테스트가 적용되었다. 피어슨의 카이 제곱 통계(-2*로그 우도비에 해당)의 히트맵은 도 12a 내지 12e에 도시된 바와 같이 이 예에서 최적의 파라미터 조합(즉, 최적 모델 파라미터(300))이 T1=530, r=30 및 n=2.0에서 발생함을 보여준다. 도 13은 이러한 파라미터 조합을 갖는 시뮬레이션된 AFS(304)가 관측된 gnomAD AFS(즉, 동의 AFS(308))를 모사함을 도시하고 있다. 추정된 T1=530 세대는 농업이 광범위하게 채택된 시기가 약 12,000년 전(즉, 신석기 시대의 시작)까지 거슬러 올라가는 고고학에 동의한다. 인구 조사와 유효 인간 모집단 규모 사이의 비율이 예상보다 낮아, 인간 모집단의 다양성이 실제로 상당히 높다는 것을 암시한다.

[0117] 하나의 예시적인 구현예에서, 도 14를 참조하면, 순방향 시간 시뮬레이션의 컨텍스트에서 선택 효과를 다루기 위해, 이전 시뮬레이션 결과에서 인간 확장 이력의 가장 가능성 있는 인구 통계학적 모델을 검색했다. 이러한 모델에 기초하여, 선택은 {0, 0.0001, 0.0002,…, 0.8, 0.9}로부터 선택된 선택 계수(들)(320)의 상이한 값을 갖는 시뮬레이션에 통합되었다. 각 세대(286)의 경우, 부모 모집단으로부터 돌연변이를 물려받고 *드노보* 돌연변이를 적용한 후, 작은 분율의 돌연변이를 선택 계수(320)에 따라 무작위로 제거하였다.

[0118] 시뮬레이션의 정밀도를 향상시키기 위해, 8071개의 트리오(즉, 부모-자손 트리오)로부터 도출된 특정 *드노보* 돌연변이율(230)을 사용하여 192개의 트리-뉴클레오티드 컨텍스트 각각에 대해 별도의 시뮬레이션이 적용되었다(단계 282). 각 선택 계수(320) 값 및 각 돌연변이율(280) 하에서, 약 20,000개 염색체의 초기 크기를 가진 인간 모집단이 현재까지 확장되는 것으로 시뮬레이션되었다. 결과 모집단(즉, 목표 또는 최종 세대(290))으로부터 1000개 세트(294)가 무작위로 샘플링되었다(단계 288). 각 세트는 gnomAD+Topmed+UK 바이오뱅크의 샘플 크기에 해당하는 약 500,000개의 염색체를 포함하였다. 8개의 CpGTi 트리-뉴클레오티드 컨텍스트 각각의 경우, 고메틸화 수준과 저메틸화 수준을 별도로 시뮬레이션했다. 두 AFS는 고메틸화 부위 또는 저메틸화 부위의 비율을 가중치로 적용하여 병합되었다.

[0119] 192개의 트리-뉴클레오티드 컨텍스트에 대한 AFS를 얻은 후, 이러한 AFS는 엑솜의 AFS를 생성하기 위해 엑솜에서 192개의 트리-뉴클레오티드 컨텍스트의 빈도에 의해 가중되었다. 이 절차는 36개의 선택 계수 각각에 대해 반복되었다(단계 306).

[0120] 그런 다음, 선택-고갈 곡선(312)이 도출되었다. 특히, 돌연변이에 대한 선택적 압력(들)이 증가함에 따라, 변이체는 점진적으로 고갈될 것으로 예상된다. 다양한 선택 수준 하에서 시뮬레이션된 AFS(304)로, "고갈"로 특성화되는 메트릭을 정의하여 중립 진화(즉, 선택 없음) 하의 시나리오와 비교하여 정제 선택에 의해 제거된 변이체의 비율을 측정하였다. 이 예에서, 고갈은 다음과 같이 특징지을 수 있다:

[0121] $$고갈 = 1 - \frac{선택을\ 갖는\ 변이체\ 수}{선택이\ 없는\ 변이체\ 수} \quad (1)$$

[0122] 고갈 값(316)은 36개의 선택 계수 각각에 대해 생성되어(단계 314) 도 15에 도시된 선택-고갈 곡선(312)을 그렸다. 이 곡선에서, 보간을 적용하여 고갈 값과 연관된 추정 선택 계수를 얻을 수 있다.

[0123] 순방향 시간 시뮬레이션을 사용한 선택 및 고갈 특성화에 관한 이전 논의를 염두에 두고, 이러한 요인을 사용하여 병원성(예를 들어, PrimateAI 또는 pAI) 점수를 기반으로 미스센스 변이체에 대한 선택 계수를 추정할 수 있다.

[0124] 예를 들어, 도 16을 참조하면, 한 연구에서, 122,439개의 gnomAD 엑솜과 13,304개의 gnomAD 전체 게놈 시퀀싱(WGS) 샘플(Topmed 샘플 제거 후), 약 65K Topmed WGS 샘플, 및 약 50K 영국 바이오뱅크 WGS 샘플을 포함하여,

대략 200,000명의 개체으로부터 변이체 대립유전자 빈도 데이터를 획득하여 데이터를 생성했다(단계 350). 각 데이터세트에서 희귀 변이체(AF <0.1%)에 초점을 맞추었다. 모든 변이체는 필터를 통과하고 gnomAD 엑솜 커버리지에 따라 중간 깊이 ≥ 1을 가져야 했다. 근교 계수 < -0.3으로 정의된 과도한 이형 접합체를 나타내는 변이체를 제외했다. 전체 엑솜 시퀀싱의 경우, 랜덤 포레스트 모델에 의해 생성된 확률이 ≥ 0.1인 경우 변이체가 제외되었다. WGS 샘플의 경우, 랜덤 포레스트 모델에 의해 생성된 확률이 ≥ 0.4인 경우 변이체가 제외되었다. 단백질 절단형 변이체(PTV)(넌센스 돌연변이 포함), 스플라이스 변이체(즉, 스플라이싱 기증자 또는 수용자 부위에서 발생하는 해당 변이체), 및 프레임시프트의 경우, 추가 필터가, 즉 기능 상실 전사체 효과 추정기(LOFTEE) 알고리즘에 의해 추정된 낮은 신뢰도에 기초한 필터링이, 적용되었다.

[0125] 3개의 데이터세트 간에 변이체를 병합하여 최종 데이터세트(352)를 형성해서 고갈 메트릭을 계산했다. 대립유전자 빈도는 다음에 따라 3개의 데이터세트에 걸쳐 평균화되도록 재계산되었다:

[0126] (2) $$AF = \frac{sum(AC_i)}{sum(AN_i)}$$

[0127] 여기서, i는 데이터세트 인덱스를 나타낸다. 변이체가 하나의 데이터세트에 나타나지 않으면, 해당 데이터세트에 대해 AC에 제로(0)가 할당되었다.

[0128] 넌센스 돌연변이 및 스플라이스 변이체의 고갈과 관련하여, 각 유전자에서 예상된 수와 비교하여 정제 선택에 의해 고갈된 PTV의 분율을 계산할 수 있다. 유전자에서 예상된 프레임시프트 수 계산의 어려움으로 인해, 대신 기능 상실 돌연변이(LOF)로 표시되는 스플라이스 변이체와 넌센스 돌연변이에 초점을 맞추었다.

[0129] 넌센스 돌연변이 및 스플라이스 변이체의 수를 병합된 데이터세트의 각 유전자에서 카운팅하여(단계 356) 관측된 LOF 수(360)(아래 식의 분자)를 얻었다.예상된 LOF 수(364)를 계산하기 위해(단계 362), 제약 메트릭을 포함하는 파일(https://storage.googleapis.com/gnomad-public/release/2.1.1/constraint/gnomad.v2.1.1.lof_metrics.by_gene.txt.bgz)은 gnomAD 데이터베이스 웹사이트에서 다운로드되었다. 병합된 데이터세트에서 관측된 동의 변이체는 기준선으로 사용되고 gnomAD로부터 예상된 LOF 수와 예상된 동의 변이체 수의 비율을 곱하여 예상된 LOF 수(364)로 전환되었다. 그런 다음, 고갈 메트릭(380)이 계산되었고(단계 378) [0,1] 내에 있는 것으로 확인되었다. 0보다 작으면, 0이 할당되고 그 반대의 경우도 마찬가지이다. 위의 내용은 다음과 같이 표현될 수 있다:

[0130] (3) $$고갈_{LOF} = 1 - \frac{\#obs\ LOF}{\#\ exp\ LOF}$$

[0131] 여기서, $$\#expLOF = \#obs\ Syn * \frac{\#exp\ LOF\ of\ gnomAD}{\#exp\ Syn\ of\ gnomAD}$$

[0132] LOF의 고갈 메트릭(380)에 기초하여, 각각의 선택-고갈 곡선(312)을 사용하여 각 유전자에 대한 LOF(390)의 선택 계수의 추정치가 도출될 수 있다(단계 388).

[0133] 미스센스 변이체에서 고갈(380)의 계산(단계 378)과 관련하여, 유전자 데이터세트(418)에 대해 도출된 예측된 병원성 점수(예를 들어, PrimateAI 또는 pAI 점수)의 구현 백분위수(420)의 일 예(도 17에 도시됨)는 각 유전자 내에서 가능한 모든 미스센스 변이체를 나타내는 데 사용되었다. 본원에 설명된 바와 같은 병원성 점수는 변이체의 상대적 피팅도를 측정하기 때문에, 미스센스 변이체의 병원성 점수는 강한 음성 선택 하에서 유전자에서 더 높은 경향이 있을 것으로 예상할 수 있다. 반대로, 중간 정도의 선택을 가진 유전자에서 점수는 더 낮을 것으로 예상할 수 있다. 따라서, 유전자에 대한 전반적인 영향을 피하기 위해 병원성 점수(예를 들어, pAI 점수)의 백분위수(420)를 사용하는 것이 적절하다.

[0134] 각 유전자에 대해, 병원성 점수 백분위수(420)(본 예에서는 pAI 백분위수)를 10개의 빈(예를 들어, (0.0, 0.1], (0.1, 0.2], …, (0.9, 1.0])으로 분할했고(단계 424), 각 빈에 속하는 관측된 미스센스 변이체의 수(428)를 카운팅했다(단계 426). 고갈 메트릭(380)은, 각각의 고갈 메트릭(380)이 10개의 빈 각각에 대해 계산되었다는 점을 제외하고는, LOF의 것과 유사하게 계산된다(430). 본원에 설명된 LOF 고갈 계산에 사용된 것과 유사하게, gnomAD로부터의 미스센스/동의 변이체의 보정 계수를 각 빈에서 예상된 미스센스 변이체 수에 적용했다. 위의 내용은 다음과 같이 표현될 수 있다:

[0135] (4) $$고갈_{Mis} = 1 - \frac{\#obs\ Mis\ in\ each\ bin}{\#\ exp\ Mis/10}$$

[0136] 여기서, $$\#expMis = \#obs\ Syn * \frac{\#exp\ Mis\ of\ gnomAD}{\#exp\ Syn\ of\ gnomAD}$$

[0137] 각 유전자 내의 10개 빈의 고갈에 기초하여, 병원성 점수의 백분위수(420)와 고갈 메트릭(들)(380) 사이의 관계(436)가 도출될 수 있다(단계 434). 일례에서, 각 빈의 중간 백분위수를 결정하고 매끄러운 스플라인을 10개 빈의 중간 지점에 피팅시켰다. 이에 대한 예는, 고갈 메트릭이 병원성 점수 백분위수와 실질적인 선형 방식으로 증가함을 나타낸, 각각 BRCA1 및 LDLR 유전자의 두 예와 관련하여 도 18 및 19에 도시되어 있다.

[0138] 이러한 방법론에 기초하여, 가능한 모든 미스센스 변이체에 대해, 그의 고갈 메트릭(380)은 유전자-특이적 피팅된 스플라인을 사용하여 병원성 백분위수 점수(420)에 기초하여 예측될 수 있다. 그런 다음, 이러한 미스센스 변이체의 선택 계수(320)는 선택-고갈 관계(예를 들어, 선택-고갈 곡선(312) 또는 다른 피팅된 함수)를 사용하여 추정될 수 있다.

[0139] 또한, 예상된 유해 희귀 미스센스 변이체 및 PTV의 수를 개별적으로 추정할 수 있다. 예를 들어, 정상적인 개체가 코딩 게놈에 반입할 수 있는 유해 희귀 변이체의 수를 평균적으로 추정하는 것이 관심 대상일 수 있다. 이러한 구현예의 예시에서, AF < 0.01%인 희귀 변이체에 초점을 맞추었다. 개체당 유해 희귀 PTV의 예상된 수를 계산하려면, 다음과 같이 특정 임계치를 초과하는 선택 계수(320)를 가진 PTV의 대립유전자 빈도를 합산하는 것과 같다:

[0140] (5) $$E\left[\#PTV \mid s > k\right] = \sum AF \mid s > k$$

[0141] PTV는 넌센스 돌연변이, 스플라이스 변이체, 및 프레임시프트를 포함하고 있으므로, 각 범주에 대해 별도로 계산이 이루어졌다. 아래 표 6에 나타낸 결과로부터, 각 개체는 s > 0.01(BRCA1 돌연변이와 같거나 그보다 나쁨)인 약 1.9개의 희귀 PTV를 갖고 있는 것으로 관찰될 수 있다.

[0142] [표 6]

[0143] 상이한 선택 계수 컷오프에서 예상된 유해 희귀 PTV 수

	s > 0.2	s > 0.1	s > 0.05	s > 0.02	s > 0.01	s > 0.001
E[#넌센스 변이체]	0.0037	0.0209	0.0754	0.3078	0.6459	0.9959
E[#스플라이스 변이체]	0.0032	0.0153	0.0471	0.1639	0.3361	0.5100
E[#프레임시프트]	0.0234	0.0635	0.1574	0.4944	0.9338	1.3745
E[#PTV]	0.0303	0.0996	0.2799	0.9662	1.9159	2.8804

[0144] 또한, 예상된 유해 희귀 미스센스 변이체 수는 상이한 임계치를 초과하는 선택 계수(320)를 갖는 희귀 미스센스의 대립 유전자 빈도를 합산하여 계산되었다:

[0145] (6) $$E\left[\#missense \mid s > k\right] = \sum AF \mid s > k$$

[0146] 아래 표 7에 나타낸 결과로부터, s > 0.01인 PTV보다 약 4배 많은 미스센스 변이체가 있다.

[0147] [표 7]

[0148] 상이한 선택 계수 컷오프에서 예상된 유해 희귀 미스센스 변이체 수

E[#미스센스 변이체]	s > 0.2	s > 0.1	s > 0.05	s > 0.02	s > 0.01	s > 0.001
교정된 gnomAD+Topmed	0.0022	0.0178	0.208	2.846	10.107	31.990
교정된 gnomAD+Topmed + UK 바이오뱅크	0.00067	0.010	0.146	2.336	8.989	30.868
교정 없는 gnomAD+Topmed + UK 바이오뱅크	0.0002	0.0035	0.0806	1.8327	8.2948	33.0164

[0149] ## VI.병원성 점수를 사용한 유전병 유병률 추정

[0150] 임상 환경에서 미스센스 변이체의 병원성 점수의 채택 및 사용을 촉진하기 위해, 임상 관심 유전자 중에서 병원성 점수와 임상 질병 유병률 사이의 관계를 조사했다. 특히, 병원성 점수에 기초한 다양한 메트릭을 사용하여 유전병 유병률을 추정하기 위한 방법론이 개발되었다. 유전병 유병률을 예측하기 위한 병원성 점수에 기초한

이러한 두 방법론의 비제한적인 예가 본원에 설명되어 있다.

[0151] 이 연구에 이용된 데이터의 관점에서 예비적 컨텍스트를 통해, 본원에서 참조된 DiscovEHR 데이터는 게이징거의 마이코드 커뮤니티 헬스 이니셔티브(MyCode Community Health Initiative)에서 50,726명의 성인 참가자의 종적 전자 건강 기록(EHR)으로부터 임상 표현형과 전체 엑솜 시퀀싱을 통합하여 정밀 의학을 촉진하는 것을 목표로 하는 리제너론 유전학 센터(Regeneron Genetics Center)와 게이징거 헬스 시스템(Geisinger Health System) 간의 협업이다. 이를 염두에 두고, 도 20을 참조하면, 임상적으로 실행 가능한 유전적 소견의 식별 및 보고를 위한 미국 의학 유전학 및 유전체학 대학(ACMG) 권장사항에서 식별된 56개의 유전자 및 25개의 의학적 상태를 포함한 76개의 유전자(G76)가 정의되었다(즉, 유전자 데이터세트(450)). G76 유전자 내에서 ClinVar "병원성" 분류뿐만 아니라 알려지고 예측된 기능 상실 변이체를 포함하는 이러한 잠재적 병원성 변이체(456)의 유병률을 평가했다. 각 유전자에서 해당 ClinVar 병원성 변이체(456)의 누적 대립유전자 빈도(CAF)(466)는 본원에서 논의된 바와 같이 도출되었다(단계 460). 대부분의 76개 유전자에 대한 대략적인 유전병 유병률은 문헌 자료로부터 얻었다.

[0152] 이러한 컨텍스트를 염두에 두고, 유전병 유병률을 예측하기 위한 병원성 점수(206)(예를 들어, PrimateAI 또는 pAI 점수)에 기반한 접근법의 두 가지 예가 개발되었다. 이러한 방법론에서, 도 21을 참조하면, 유전자의 미스센스 변이체(200)가 병원성인지(즉, 병원성 변이체(220)) 그렇지 않은지(즉, 비병원성 변이체(476))를 결정하기 위해 유전자-특이적 병원성 점수 임계치(212)가 이용된다(결정 블록 210). 일례에서, 병원성 점수(206)가 특정 유전자에서 병원성 점수의 75번째 백분위수보다 큰 경우 예측된 유해 변이체에 대한 컷오프가 정의되었지만, 다른 컷오프가 적절하게 이용될 수 있다. 유전병 유병률 메트릭은 단계 478에서 도출된 바와 같이 예측된 유해 미스센스 변이체의 예상된 누적 대립유전자 빈도(CAF)(480)로 정의되었다. 도 22에 도시된 바와 같이, Clinvar 병원성 변이체의 DiscovEHR 누적 AF와 이 메트릭의 스피어만 상관 관계는 0.5272이다. 유사하게, 도 23은 질병 유병률과 이러한 메트릭의 스피어만 상관 관계가 0.5954로서, 양호한 상관 관계를 암시하고 있음을 예시하고 있다. 따라서, 유전병 유병률 메트릭(즉, 예측된 유해 미스센스 변이체의 예상된 누적 대립유전자 빈도(CAF))는 유전병 유병률의 예측자 역할을 할 수 있다.

[0153] 도 21의 단계 478로 나타낸 각 유전자에 대한 유전병 유병률 메트릭을 계산하는 것과 관련하여, 두 가지 서로 다른 접근법을 평가했다. 제1 방법론에서, 도 24를 참조하면, 유해 미스센스 변이체(220) 목록의 트리뉴클레오티드 컨텍스트 구성(500)이 초기에 획득된다(단계 502). 본 컨텍스트에서, 이는 모든 가능한 미스센스 변이체를 얻는 것에 대응할 수 있으며, 이러한 병원성 변이체(220)는 해당 유전자에서 75번째 백분위수 임계치(또는 다른 적합한 컷오프)를 초과하는 병원성 점수(206)를 갖는 변이체이다.

[0154] 각 트리뉴클레오티드 컨텍스트(500)에 대해, 본원에 설명된 바와 같은 순방향 시간 시뮬레이션이 수행되어(단계 502) 0.01과 같은 선택 계수(320)를 가정하고 해당 트리뉴클레오티드 컨텍스트에 대한 *드노보* 돌연변이율(280)을 사용하여 예상된(즉, 시뮬레이션된) 대립유전자 빈도 스펙트럼(AFS)(304)을 생성한다. 이 방법론의 일 구현예에서, 시뮬레이션은 약 400K 염색체(약 200K 샘플) 중에서 100,000개의 독립적인 부위를 시뮬레이션했다. 따라서, 이러한 컨텍스트에서 특정 트리뉴클레오티드 컨텍스트(500)에 대해 예상된 AFS(304)는 유해 변이체 목록에서 시뮬레이션된 AFS /100,000 * 트리뉴클레오티드의 발생이다. 192개의 트리뉴클레오티드를 합산하면 유전자에 대해 예상된 AFS(304)가 생성된다. 이러한 접근법에 따른 특정 유전자의 유전병 유병률 메트릭(즉, 예상된 CAF(480))은 해당 유전자에 대한 예상된 AFS(304)에서 시뮬레이션된 희귀 대립유전자 빈도(즉, AF ≤ 0.001)의 합계(단계 506)로 정의된다.

[0155] 제2 방법론에 따라 도출된 바와 같은 유전병 유병률 메트릭은 제1 방법론을 사용하여 도출된 것과 유사하지만 유해 미스센스 변이체 목록의 정의에 있어서 다르다. 제2 방법론에 따라, 도 25를 참조하면, 병원성 변이체(220)는 본원에서 논의된 바와 같이 그 추정된 고갈이 해당 유전자에서 단백질 절단형 변이체(PTV)의 고갈의 ≥ 75%인 경우 유전자당 예측된 유해 변이체로 정의된다. 예를 들어, 도 25에 도시된 바와 같이, 이러한 컨텍스트에서, 병원성 점수(206)는 관심 변이체(들)(200)에 대해 측정될 수 있다(단계 202). 병원성 점수(들)(206)는 본원에서 논의된 바와 같이 소정의 백분위수 병원성-고갈 관계(436)를 사용하여 고갈(522)을 추정(단계 520)하는 데 사용될 수 있다. 그런 다음, 추정된 고갈(522)은 고갈 임계치 또는 컷오프(524)와 비교되어(결정 블록 526) 병원성 변이체(220)로 간주되는 것과 비병원성 변이체(476)를 분리할 수 있다. 병원성 변이체(220)가 결정되면, 처리는 예상된 CAF(480)를 도출하기 위해 단계 478에서 위에서 논의된 바와 같이 진행될 수 있다.

[0156] 이러한 제2 방법론을 사용하여 도출된 바와 같은 유전병 유병률 메트릭과 관련하여, 도 26은 Clinvar 병원성 변이체의 DiscovEHR 누적 AF와 제2 방법론에 따라 계산된 유전병 유병률의 스피어만 상관 관계가 0.5208임을 보여

준다. 유사하게, 도 27은 제2 방법론에 따라 계산된 유전병 유병률 메트릭과 질병 유병률의 스피어만 상관 관계가 0.4102임을 보여주며, 이는 양호한 상관 관계를 암시한다. 따라서, 제2 방법론을 사용하여 계산된 바와 같은 메트릭은 유전병 유병률의 예측자로 역할을 할 수도 있다.

[0157] **VII.병원성 점수의 재보정**

[0158] 본원에 설명된 바와 같이, 본 교시에 따라 생성된 병원성 점수는 주로 변이체 주변의 DNA 측면 서열, 종 간의 보존, 및 단백질 2차 구조에 기초하여 훈련된 신경망을 사용하여 도출된다. 그러나, 병원성 점수(예를 들어, PrimateAI 점수)와 연관된 변이는 클 수 있다(예를 들어, 약 0.15). 또한, 병원성 점수를 계산하기 위해 본원에서 논의된 일반화된 모델의 특정 구현예는 훈련 동안 인간 모집단에서 관측된 대립유전자 빈도의 정보를 활용하지 않는다. 특정 상황에서, 병원성 점수가 높은 일부 변이체는 대립유전자 카운트 > 1을 갖는 것으로 나타날 수 있으며, 이는 대립유전자 카운트에 기초하여 이러한 병원성 점수에 페널티를 부과할 필요가 있음을 암시한다. 이를 염두에 두고, 이러한 상황을 해결하기 위해 병원성 점수를 재보정하는 것이 유용할 수 있다. 본원에서 논의된 하나의 예시적인 실시예에서, 재보정 접근법은 변이체의 병원성 점수의 백분위수에 초점을 맞출 수 있는데, 이는 더 강력하고 전체 유전자에 가해지는 선택 압력에 의해 덜 영향을 받을 수 있기 때문이다.

[0159] 이를 염두에 두고, 도 28을 참조하면, 재보정 접근법의 일례에서, 실제 병원성 백분위수는 관측된 병원성 점수 백분위수(550)에서 노이즈를 평가하고 설명할 수 있도록 모델링된다. 이러한 모델링 프로세스에서, 실제 병원성 백분위수는 (0,1]에 대해 이산 균일하게 분포된다고 가정할 수 있다(예를 들어, [0.01, 0.02, …, 0.99, 1.00]인 100개의 값을 취함). 관측된 병원성 점수 백분위수(550)는 표준 편차가 0.15인 정규 분포를 따르는 일부 노이즈 항을 갖는 실제 병원성 점수 백분위수에 중심을 두는 것으로 가정할 수 있다:

[0160] (7) $obsAI \sim trueAI + e,\ e \sim N(0,\ sd = 0.15)$

[0161] 이러한 컨텍스트에서 관측된 병원성 점수 백분위수(550)의 분포(554)는 도 29에 도시된 바와 같이 가우시안 노이즈로 오버레이된 실제 병원성 점수 백분위수의 이산 균일 분포이며, 각 선은 실제 병원성 점수 백분위수의 각 값을 중심으로 하는 정규 분포를 나타낸다. 가우시안 노이즈로 오버레이된 관측된 병원성 점수 백분위수의 이러한 이산 균일 분포(556)에 대한 밀도 플롯이 도 30에 도시되어 있고, 단계 562에서 결정된 누적 분포 함수(CDF)(558)는 도 31에 도시되어 있다. 이러한 CDF(558)로부터, 누적 확률은 100개의 간격으로 분할되고 관측된 병원성 점수 백분위수(550)에 대한 분위수(568)가 생성된다(단계 566).

[0162] 실제 병원성 점수 백분위수(도 32의 x축)를 갖는 변이체가 관측된 병원성 점수 백분위수 간격(y축)에 속할 확률을 시각화하기 위해, 이러한 100x100 확률 매트릭스의 각 행은 합계가 1이 되도록 정규화될 수 있고 그 결과는 히트맵(572)(도 32)으로서 플로팅될 수 있다(단계 570). 히트맵(572) 상의 각 점은 관측된 병원성 점수 백분위수(550) 간격 내의 변이체가 실제 병원성 점수 백분위수에서 실제로 나올 확률(즉, 실제 병원성 점수 백분위수(x축)를 갖는 변이체가 관측된 병원성 점수 백분위수 간격(y축)에 속할 확률)을 측정한다.

[0163] 도 33을 참조하면, 미스센스 변이체에 대해, 각 유전자에서 10개의 빈 각각에 대한 고갈 메트릭(522)을 본원에 설명된 방법론을 사용하여 결정하였다. 이 예에서, 본원의 다른 곳에서 논의된 바와 같이, 병원성 점수(206)는 비닝 프로세스의 일부로서 관심 변이체(200)에 대해 계산될 수 있다(단계 202). 결과적으로, 각각의 병원성 점수(206)는 소정의 백분위수 병원성 점수-고갈 관계(436)에 기초하여 고갈(522)을 추정(단계 520)하는 데 사용될 수 있다.

[0164] 이러한 고갈 메트릭(522)은 각 빈 내에 속하는 변이체가 정제 선택에 의해 제거될 수 있는 확률을 측정한다. 이에 대한 예가 유전자 SCN2A에 대하여 도 34 및 35에 도시되어 있다. 특히, 도 34는 SCN2A 유전자의 미스센스 변이체에 대한 10개의 빈의 백분위수에 걸친 고갈 확률을 도시하고 있다. 변이체가 선택에서 생존할 확률은 (1 - 고갈)로 정의될 수 있으며, 생존 확률(580)로 표시되고 단계 582에서 결정된다. 이 확률이 0.05 미만이면, 0.05로 설정될 수 있다. 도 35는 SCN2A 유전자의 미스센스 변이체에 대한 10개의 빈의 백분위수에 걸친 생존 확률(580)을 도시하고 있다. 두 도면에서, x축 상의 1.0에 표시된 다이아몬드는 PTV를 나타낸다.

[0165] 일 구현예에 따르면, 매끄러운 스플라인이 빈(예를 들어, 10개의 빈)에 걸쳐 각 빈의 생존 확률 대 중간 병원성 점수 백분위수에 피팅되었고(단계 584) 병원성 점수의 각 백분위수에 대한 생존 확률을 생성했다. 이러한 접근법에 따르면, 이는 생존 확률 보정 계수(590)를 구성하며, 이는 병원성 점수(206)의 백분위수가 높을수록 변이체가 정제 선택에서 생존할 가능성이 적음을 암시한다. 다른 구현예에서, 매끄러운 스플라인을 피팅하는 대신 보간과 같은 다른 기술이 이용될 수 있다. 그런 다음, 해당 관측된 변이체의 높은 병원성 점수(206)는 이러한

보정 계수(590)에 따라 처벌되거나 교정될 수 있다.

[0166] 이전 내용을 염두에 두고, 도 36을 참조하면, 생존 확률 보정 계수(590)는 재보정을 수행하기 위해 이용될 수 있다. 예를 들어, 이전에 예시된 바와 같이 히트맵(572)으로 시각화될 수 있는 확률 매트릭스의 컨텍스트에서, 특정 유전자의 경우, 히트맵(572)(예를 들어, 차원 50 x 50, 100 x 100 등인 확률 매트릭스)의 각 행에 각각의 생존 확률 보정 계수(590)(예를 들어, 100개의 값의 벡터)를 곱하여(단계 600) 해당 유전자의 예상된 고갈에 의해 히트맵(572)의 값을 감소시킨다. 그런 다음, 히트맵의 각 행은 합계가 1이 되도록 재보정된다. 그런 다음, 재보정된 히트맵(596)은 도 37에 도시된 바와 같이 플로팅되고 표시될 수 있다. 이 예에서 재보정된 히트맵(596)은 x축에 실제 병원성 점수 백분위수를 표시하고 재보정된 관측된 병원성 점수 백분위수는 y축에 있다.

[0167] 실제 병원성 점수 백분위수는 빈(즉, 1%-10%(재보정된 히트맵(596)의 처음 10개 열)를 제1 빈으로 병합, 11%-20%(재보정된 히트맵(596)의 다음 10개 열)를 제2 빈으로 병합, 등)으로 분할되었고, 이는 변이체가 실제 병원성 점수 백분위수 빈 각각에서 나올 수 있는 확률을 나타낸다. 관측된 병원성 점수 백분위수(예를 들어, 재보정된 히트맵(596)의 x번째 행에 대응하는 x%)를 갖는 해당 유전자의 변이체에 대해, 이러한 변이체가 실제 병원성 점수 백분위수 빈(예를 들어, 10개의 빈) 각각 내에 속할 수 있는 확률을 얻을 수 있다(단계 608). 이는 각 빈에 대한 변이체 기여도(612)로 표시될 수 있다.

[0168] 이러한 예에서, 빈(예를 들어, 10개의 빈) 각각 내의 예상된 미스센스 변이체(620) 수(단계 624에서 도출됨)는 각각의 유전자에서 관측된 모든 미스센스 변이체에 걸쳐 해당 빈에 대한 변이체 기여도의 합계이다. 유전자에 대한 각각의 빈 내에 속하는 미스센스 변이체(620)의 이러한 예상 수에 기초하여, 본 명세서에서 논의된 미스센스 변이체에 대한 고갈 공식은 각각의 미스센스 빈에 대한 수정된 고갈 메트릭(634)을 계산하는 데 사용될 수 있다(단계 630). 이는 각 백분위수 빈에 대한 교정된 고갈 메트릭의 예가 플로팅된 도 38에 도시되어 있다. 특히, 도 38은 유전자 SCN2A에서 재보정된 고갈 메트릭 대 원래의 고갈 메트릭의 비교를 도시하고 있다. x축 상의 1.0에 플로팅된 다이아몬드는 PTV의 고갈을 나타낸다.

[0169] 병원성 점수(206)를 재보정하는 이러한 방식으로, 병원성 점수(206)의 백분위수로 노이즈 분포가 모델링되고 예측된 고갈 메트릭(522)에서 노이즈가 감소된다. 이는 본원에서 논의된 바와 같이 미스센스 변이체에서 선택 계수(320)의 추정에 대한 노이즈의 영향을 완화하는 데 도움이 될 수 있다.

[0170] **VIII.신경망 지침서**

[0171] **신경망**

[0172] 이전 논의에서, 신경망 아키텍처 및 사용의 다양한 양태가 병원성 분류 또는 채점 네트워크의 컨텍스트에서 참조된다. 신경망 설계 및 사용의 이러한 다양한 양태에 대한 광범위한 지식은 본원에서 논의된 바와 같이 병원성 분류 네트워크를 이해하고 이용하고자 하는 사람들에게 필요한 것으로 여겨지지 않지만, 추가 세부사항을 원하는 사람들을 위해 다음과 같은 신경망 지침서가 추가 참조를 통해 제공된다.

[0173] 이를 염두에 두고, 일반적인 의미에서 "신경망"이란 용어는 각각의 출력을 수신하도록 훈련되고 그 훈련에 따라 입력이 수정, 분류 또는 이와 달리 처리되는 병원성 점수와 같은 출력을 생성하도록 훈련된 계산 구성으로 이해될 수 있다. 이러한 구조는 생물학적 뇌를 모델로 하여 신경망이라고 지칭될 수 있으며, 구조의 상이한 노드는 "뉴런"과 동일시되며, 이는 노드 간의 복잡한 잠재적 상호 연결이 가능하도록 광범위한 다른 노드와 상호 연결될 수 있다. 일반적으로, 신경망은 경로 및 관련 노드가 통상적으로 (예를 들어, 입력 및 출력이 알려져 있거나 비용 함수가 최적화될 수 있는 샘플 데이터를 사용하여) 예를 들어 훈련되므로 기계 학습의 한 형태로 간주될 수 있고, 신경망이 사용되고 그 성능이나 출력이 수정되거나 재훈련됨에 따라 시간이 지나면서 학습하거나 진화할 수 있다.

[0174] 이를 염두에 두고, 추가 예시를 통해, 도 39는 신경망(700)의, 여기서는 다수 계층(702)을 갖는 완전 연결된 신경망(700)의, 일례의 단순화된 도면이다. 본원에 언급되고 도 39에 도시된 바와 같이, 신경망(700)은 서로 간에 메시지를 교환하는 상호 연결된 인공 뉴런(704)(예를 들어, a_1, a_2, a_3)의 시스템이다. 예시된 신경망(700)은 3개의 입력을 가지며, 은닉 계층에 2개의 뉴런이 있고 출력 계층에 2개의 뉴런이 있다. 은닉 계층은 활성화 함수 $f(●)$를 갖고, 출력 계층은 활성화 함수 $g(●)$를 갖는다. 연결에는 적절하게 훈련된 네트워크가 처리하도록 훈련된 입력을 공급받을 때 정확하게 응답하도록 훈련 프로세스 중에 조정되는 관련된 숫자 가중치(예를 들어, w_{11}, w_{21}, w_{12}, w_{31}, w_{22}, w_{32}, v_{11}, v_{22})가 구비된다. 입력 계층은 원시 입력을 처리하고, 은닉 계층은 입력 계층과 은닉 계층 간의 연결 가중치에 기초하여 입력 계층으로부터의 출력을 처리한다. 출력 계층은 은닉 계층

으로부터 출력을 가져와 은닉 계층과 출력 계층 간의 연결 가중치에 기초하여 이를 처리한다. 하나의 컨텍스트에서, 네트워크(700)는 특징 검출 뉴런의 다수 계층을 포함한다. 각 계층에는 이전 계층으로부터 다양한 입력 조합에 응답하는 많은 뉴런이 있다. 이러한 계층은 제1 계층이 입력 이미지 데이터에서 프리미티브 패턴 세트를 검출하고, 제2 계층이 패턴의 패턴을 검출하고, 제3 계층이 해당 패턴의 패턴을 검출하고, 기타 등등이 이루어지도록 구성될 수 있다.

[0175] **컨벌루션 신경망**

[0176] 신경망(700)은 작동 모드에 기초하여 다양한 유형으로 분류될 수 있다. 예를 들어, 컨벌루션 신경망은 조밀하거나 조밀하게 연결된 계층과 달리 하나 이상의 컨벌루션 계층을 이용하거나 통합하는 신경망의 한 유형이다. 특히, 조밀하게 연결된 계층은 입력 특징 공간에서 전역 패턴을 학습한다. 반대로, 컨벌루션 계층은 로컬 패턴을 학습한다. 예를 들어, 이미지의 경우, 컨벌루션 계층은 작은 윈도우나 입력의 서브세트에서 발견되는 패턴을 학습할 수 있다. 로컬 패턴 또는 특징에 대한 이러한 초점은 컨벌루션 신경망에 다음의 두 가지 유용한 속성을 제공한다: (1) 이들이 학습하는 패턴은 변환 불변이고 (2) 이들은 패턴의 공간 계층 구조를 학습할 수 있다.

[0177] 이러한 첫 번째 속성과 관련하여, 데이터세트의 일 부분 또는 서브세트에서 특정 패턴을 학습한 후, 컨벌루션 계층은 동일하거나 상이한 데이터세트의 다른 부분에서 패턴을 인식할 수 있다. 반대로, 조밀하게 연결된 네트워크는 다른 위치(예를 들어, 새 위치)에 있는 경우 패턴을 새로 학습해야 한다. 이러한 속성은 다른 컨텍스트 및 위치에서 식별되도록 일반화될 수 있는 표현을 학습하는 데 더 적은 훈련 샘플이 필요하기 때문에 컨벌루션 신경망 데이터를 효율적으로 만든다.

[0178] 두 번째 속성과 관련하여, 제1 컨벌루션 계층은 작은 로컬 패턴을 학습할 수 있고, 제2 컨벌루션 계층은 제1 계층의 특징으로 이루어진 더 큰 패턴을 학습할 수 있고, 기타 등등이다. 이를 통해 컨벌루션 신경망은 점점 더 복잡해지고 추상적인 시각적 개념을 효율적으로 학습할 수 있다.

[0179] 이를 염두에 두고, 컨벌루션 신경망은 많은 상이한 계층(702)에 배열된 인공 뉴런(704)의 계층을 종속적으로 만드는 활성화 함수와 상호 연결함으로써 고도의 비선형 매핑을 학습할 수 있다. 하나 이상의 하위 샘플링 계층과 비선형 계층이 산재된 하나 이상의 컨벌루션 계층을 포함하고 통상적으로 하나 이상의 완전 연결된 계층이 뒤따른다. 컨벌루션 신경망의 각 요소는 이전 계층에서 특징 세트로부터 입력을 수신한다. 컨벌루션 신경망은 동일한 특징 맵 내의 뉴런이 동일한 가중치를 갖기 때문에 동시에 학습한다. 이러한 로컬 공유 가중치는, 다차원 입력 데이터가 네트워크에 입력될 때 컨벌루션 신경망이 특징 추출 및 회귀 또는 분류 프로세스에서 데이터 재구성의 복잡성을 방지하도록, 네트워크의 복잡성을 감소시킨다.

[0180] 컨벌루션은 2개의 공간 축(높이 및 폭) 및 깊이 축(채널 축이라고도 함)이 있는, 특징 맵이라고 불리우는, 3D 텐서에 걸쳐 작동한다. 컨벌루션 작업은 입력 특징 맵으로부터 패치를 추출하고 이러한 모든 패치에 동일한 변환을 적용하여, 출력 특징 맵을 생성한다. 이러한 출력 특징 맵은 여전히 3D 텐서이고, 폭 및 높이를 갖는다. 출력 깊이는 계층의 파라미터이기 때문에 깊이는 임의적일 수 있고, 해당 깊이 축에서 서로 다른 채널은 필터를 나타낸다. 필터는 입력 데이터의 특정 양태를 인코딩한다.

[0181] 예를 들어, 제1 컨벌루션 계층이 주어진 크기(28, 28, 1)의 특징 맵을 가져와 크기(26, 26, 32)의 특징 맵을 출력하는 예에서, 입력에 대해 32개의 필터를 계산한다. 이러한 32개의 출력 채널 각각은 입력에 대한 필터의 응답 맵인 26 x 26 그리드 값을 포함하며, 입력의 다른 위치에서 해당 필터 패턴의 응답을 나타낸다. 이것이 이 컨텍스트에서 특징 맵이라는 용어가 의미하는 바이고: 깊이 축에서 모든 차원은 특징(또는 필터)이고, 2D 텐서 출력 [:, :, n]은 입력에 대한 이러한 필터 응답의 2D 공간 맵이다.

[0182] 이전 내용을 염두에 두고, 컨벌루션은 다음의 두 가지 주요 파라미터로 정의된다: (1) 입력으로부터 추출된 패치의 크기 및 (2) 출력 특징 맵의 깊이(즉, 컨벌루션에 의해 계산된 필터 수). 통상적인 구현에에서, 이들은 32의 깊이에서 시작하여 64의 깊이까지 계속되고 128 또는 256의 깊이에서 종료되지만, 특정 구현예는 이 진행과 다를 수 있다.

[0183] 도 40을 참조하면, 컨벌루션 프로세스의 시각적 개요가 도시되어 있다. 이 예에 도시된 바와 같이, 컨벌루션은 3D 입력 특징 맵(720)에서 이러한 윈도우(예를 들어, 크기 3 x 3 또는 5 x 5의 윈도우)를 슬라이딩(예를 들어, 점진적으로 이동)시키고, 모든 위치에서 멈추고, 주변 특징(형상(윈도우_높이, 윈도우_폭, 입력_깊이))의 3D 패치(722)를 추출하여 작동한다. 그런 다음, 각 3D 패치(722)는 (컨벌루션 커널이라고 하는 동일한 학습된 가중치 매트릭스를 갖는 텐서 곱을 통해) 형상(출력_깊이)(즉, 변환된 패치)의 1D 벡터(724)로 변환된다. 그런 다

음, 이러한 벡터(724)는 형상(높이, 폭, 출력_깊이)의 3D 출력 특징 맵(726)으로 공간적으로 재조립된다. 출력 특징 맵(726)의 모든 공간 위치는 입력 특징 맵(720)의 동일한 위치에 대응한다.예를 들어, 3 x 3 윈도우에서, 벡터 출력 [i, j, :]은 3D 패치 입력 [i-1: i+1, j-1:J+1, :]에서 나온다.

[0184] 이전 내용을 염두에 두고, 컨벌루션 신경망은 훈련 프로세스 동안 여러 구배 업데이트 반복을 통해 학습되는 컨벌루션 필터(가중치의 매트릭스)와 입력 값 사이에서 컨벌루션 작업을 수행하는 컨벌루션 계층을 포함한다. (m, n)이 필터 크기이고 W가 가중치의 매트릭스인 경우, 컨벌루션 계층은 내적 $W \cdot x + b$를 계산하여 입력 X와 W의 컨벌루션을 수행하며, 여기서 x는 X의 인스턴스이고 b는 바이어스이다. 컨벌루션 필터가 입력에서 슬라이딩하는 단계 크기를 스트라이드라고 하며, 필터 영역($m \times n$)을 수용 필드라고 한다. 동일한 컨벌루션 필터는 입력의 상이한 위치에 걸쳐 적용되어 학습된 가중치 수를 감소시킨다. 위치 불변 학습도 가능하다, 즉 중요한 패턴이 입력에 존재하는 경우 컨벌루션 필터는 서열의 어디에 있든 이를 학습한다.

[0185] **컨벌루션 신경망 훈련**

[0186] 이전 논의로부터 알 수 있듯이, 컨벌루션 신경망의 훈련은 주어진 관심 작업을 수행하는 네트워크의 중요한 양태이다. 컨벌루션 신경망은 입력 데이터가 특정 출력 추정치로 이어지도록 조정되거나 훈련된다. 컨벌루션 신경망은 출력 추정치가 지상 실측과 점진적으로 부합하거나 이에 접근할 때까지 출력 추정치와 지상 실측의 비교를 기반으로 역전파를 사용하여 조정된다.

[0187] 컨벌루션 신경망은 지상 실측과 실제 출력 간의 차이(즉, 오류, δ)에 기초하여 뉴런 사이의 가중치를 조정하여 훈련된다. 훈련 프로세스에서 중간 단계는 본원에 설명된 바와 같이 컨벌루션 계층을 사용하여 입력 데이터로부터 특징 벡터를 생성하는 단계를 포함한다. 출력에서 시작하여, 각 계층에서 가중치에 대한 구배가 계산된다. 이를 역방향 패스 또는 후진으로 지칭된다. 네트워크에서 가중치는 음의 구배와 이전 가중치의 조합을 사용하여 업데이트된다.

[0188] 일 구현예에서, 컨벌루션 신경망(150)은 경사 하강법에 의해 오류의 역전파를 수행하는 확률적 경사 업데이트 알고리즘(예를 들어, ADAM)을 사용한다. 알고리즘은 네트워크에서 모든 뉴런의 활성화 계산을 포함하여, 순방향 패스에 대한 출력을 산출한다. 그런 다음, 오류 및 정확한 가중치가 계층당 계산된다. 일 구현예에서, 컨벌루션 신경망은 경사 하강법 최적화를 사용하여 모든 계층에 걸쳐 오류를 계산한다.

[0189] 일 구현예에서, 컨벌루션 신경망은 확률적 경사 하강법(SGD)을 사용하여 비용 함수를 계산한다. SGD는 하나의 무작위 데이터 쌍으로부터만 계산하여 손실 함수에서 가중치에 대한 구배를 근사화한다. 다른 구현예에서, 컨벌루션 신경망은 유클리드 손실 및 소프트맥스 손실과 같은 다양한 손실 함수를 사용한다. 추가 구현예에서, 아담 확률적 최적화기가 컨벌루션 신경망에 의해 사용된다.

[0190] **컨벌루션 계층**

[0191] 컨벌루션 신경망의 컨벌루션 계층은 특징 추출기 역할을 한다. 특히, 컨벌루션 계층은 입력 데이터를 학습하고 계층적 특징으로 분해할 수 있는 적응식 특징 추출기 역할을 한다. 컨벌루션 작업은 통상적으로 입력 데이터에 필터로 적용되는 "커널"을 포함하며, 출력 데이터를 생성한다.

[0192] 컨벌루션 작업은 입력 데이터에 대해 커널을 슬라이딩(예를 들어, 증분적으로 이동)시키는 것을 포함한다. 커널의 각 위치에 대해, 커널과 입력 데이터의 중복 값을 곱하고 그 결과를 추가한다. 곱의 합계는 커널이 중앙에 있는 입력 데이터의 지점에서 출력 데이터의 값이다. 많은 커널로부터 생성된 상이한 출력을 특징 맵이라고 한다.

[0193] 컨벌루션 계층이 훈련되면, 새로운 추론 데이터에 대한 인식 작업을 수행하기 위해 적용된다. 컨벌루션 계층이 훈련 데이터로부터 학습하기 때문에, 이들은 명시적인 특징 추출을 방지하고 훈련 데이터로부터 암묵적으로 학습한다. 컨벌루션 계층은 훈련 프로세스의 일부로 결정되고 업데이트되는 컨벌루션 필터 커널 가중치를 사용한다. 컨벌루션 계층은 더 높은 계층에서 조합되는 입력의 다양한 특징을 추출한다. 컨벌루션 신경망은 다양한 개수의 컨벌루션 계층을 사용하고, 그 각각은 커널 크기, 스트라이드, 패딩, 특징 맵 및 가중치 수와 같은 상이한 컨벌루션 파라미터가 있다.

[0194] **하위 샘플링 계층**

[0195] 컨벌루션 신경망 구현의 추가 양태는 계층의 하위 샘플링을 포함할 수 있다. 이러한 컨텍스트에서, 하위 샘플링 계층은 컨벌루션 계층에 의해 추출된 특징의 해상도를 줄여 추출된 특징 또는 특징 맵을 노이즈 및 왜곡에 대해 강건하게 만든다. 일 구현예에서, 하위 샘플링 계층은 두 가지 유형의 풀링 작업인 평균 풀링 및 최대 풀

링을 사용한다. 풀링 작업은 입력을 비중첩 공간 또는 영역으로 분할한다. 평균 풀링의 경우, 지역에서 값의 평균이 계산된다. 최대 풀링의 경우, 값의 최대 값이 선택된다.

[0196] 일 구현예에서, 하위 샘플링 계층은, 출력을 최대 풀링에서 입력 중 하나에만 매핑하고 평균 풀링에서 입력 평균에 출력을 매핑하여, 이전 계층에서 뉴런 세트에 대한 풀링 작업을 포함한다. 최대 풀링에서, 풀링 뉴런의 출력은 입력 내에 있는 최대 값이다. 평균 풀링에서, 풀링 뉴런의 출력은 입력 뉴런 세트 내에 있는 입력 값의 평균 값이다.

[0197] **비선형 계층**

[0198] 본 개념과 관련된 신경망 구현예의 추가 양태는 비선형 계층을 사용하는 것이다. 비선형 계층은 서로 다른 비선형 트리거 함수를 사용하여 각 은닉 계층에서 가능성이 있는 특징의 고유한 식별 신호를 전송한다. 비선형 계층은 다양한 특정 함수를 사용하여, 정류 선형 유닛(ReLU), 하이퍼볼릭 탄젠트, 하이퍼볼릭 탄젠트 절대값, 시그모이드 및 연속 트리거(비선형) 함수를 포함하나 이에 제한되지 않는, 비선형 트리거링을 구현한다. 일 구현예에서, ReLU 활성화는 함수 $y = max(x, 0)$를 구현하고, 계층의 입력 및 출력 크기를 동일하게 유지한다. ReLU를 사용하는 잠재적인 하나의 이점은 컨벌루션 신경망이 몇 배 더 빠르게 훈련될 수 있는 점이다. ReLU는 입력 값이 0보다 크고 그렇지 않으면 0인 경우에 입력에 대해 선형인 비연속 비포화 활성화 함수이다.

[0199] 다른 구현예에서, 컨벌루션 신경망은 연속적인 비포화 함수인 파워 유닛 활성화 함수를 사용할 수 있다. 파워 활성화 함수는 c가 홀수이면 x 및 y-반대칭 활성화를 산출할 수 있고, c가 짝수이면 y-대칭 활성화를 산출할 수 있다. 일부 구현예에서, 유닛은 비정류 선형 활성화를 산출한다.

[0200] 또 다른 구현예에서, 컨벌루션 신경망은 연속적인 포화 함수인 시그모이드 유닛 활성화 함수를 사용할 수 있다. 시그모이드 유닛 활성화 함수는 음의 활성화를 산출하지 않으며 y축에 대한 반대칭일 뿐이다.

[0201] **잔차 연결**

[0202] 컨벌루션 신경망의 추가 특징은 도 41에 도시된 바와 같이 특징 맵 추가를 통해 하류에 사전 정보를 재주입하는 잔차 연결을 사용하는 것이다. 이 예에 나타낸 바와 같이, 잔차 연결(730)은 과거 출력 텐서를 이후 출력 텐서에 추가함으로써 데이터의 하류 흐름에 이전 표현을 재주입하는 것을 포함하며, 이는 데이터 처리 흐름을 따라 정보 손실을 방지하는 데 도움이 된다. 잔차 연결(730)은 이전 계층의 출력을 이후 계층에 대한 입력으로서 이용 가능하게 만들어, 순차적 네트워크에서 숏컷을 효과적으로 생성하는 것을 포함한다. 이후 활성화에 연결되는 대신, 이전의 출력은 이후 활성화와 합산되며, 이는 두 활성화가 동일한 크기임을 가정한다. 크기가 서로 다른 경우, 이전 활성화를 목표 형상으로 재형상화하기 위한 선형 변환이 사용될 수 있다. 잔차 연결은 임의의 대규모 딥 러닝 모델에 존재할 수 있는 두 문제인 (1) 구배 소실 및 (2) 표상적 병목 현상을 해결한다. 일반적으로, 10개보다 많은 계층을 갖는 임의의 모델에 잔차 연결(730)을 추가하는 것이 유리할 수 있다.

[0203] **잔차 학습 및 스킵 연결**

[0204] 본 기술 및 접근법과 관련된 컨벌루션 신경망에 존재하는 다른 개념은 스킵 연결을 사용하는 것이다. 잔차 학습의 기본 원리는 잔차 매핑이 원래 매핑보다 학습되기 쉽다는 점이다. 잔차 네트워크는 훈련 정확도의 저하를 완화하기 위해 다수의 잔차 유닛을 적층한다. 잔차 블록은 심층 신경망에서 구배가 소실되는 것을 방지하기 위해 특수 추가 스킵 연결을 사용한다. 잔차 블록의 초반에서, 데이터 흐름은 다음과 같은 2개의 스트림으로 분리된다: (1) 제1 스트림은 블록의 미변경 입력을 전달하고, (2) 제2 스트림은 가중치 및 비선형성을 적용한다. 블록의 끝에서, 2개의 스트림은 요소별 합계를 사용하여 병합된다. 이러한 구성의 한 이점은 구배가 네트워크를 통해 더 용이하게 흐를 수 있는 점이다.

[0205] 이러한 잔차 네트워크의 이점을 통해, 심층 컨벌루션 신경망(CNN)을 쉽게 훈련할 수 있으며 데이터 분류, 객체 검출 등의 정확도를 높일 수 있다. 컨벌루션 피드-포워드 네트워크는 I번째 계층을 (I+1)번째 계층에 입력으로서 연결한다. 잔차 블록은 식별 함수를 갖는 비선형 변환을 우회시키는 스킵 연결을 추가한다. 잔차 블록의 이점은 구배가 아이덴티티 함수를 통해 이후 계층으로부터 이전 계층으로 직접 흐를 수 있는 점이다.

[0206] **배치 정규화**

[0207] 현재의 병원성 분류 접근법에 적용될 수 있는 컨벌루션 신경망의 구현과 관련된 추가 양태는, 데이터 표준화를 네트워크 아키텍처의 필수적인 부분으로 만들어 심층 네트워크 훈련을 가속화하는 방법인 배치 정규화이다. 배치 정규화는 훈련 중에 시간이 지남에 따라 평균 및 분산이 변경되더라도 데이터를 적응적으로 정규화할 수 있으며 훈련 중에 표시되는 데이터의 배치별 평균 및 분산의 지수 이동 평균을 내부적으로 유지함으로써

작동한다. 배치 정규화의 한 가지 효과는, 잔차 연결과 마찬가지로, 구배 전파에 도움이 되므로 심층 네트워크의 사용을 용이하게 한다는 점이다.

[0208] 그러므로, 배치 정규화는, 완전 연결된 또는 컨벌루션 계층과 마찬가지로, 모델 아키텍처에 삽입될 수 있는 또 다른 계층으로 볼 수 있다. 배치 정규화 계층은 통상적으로 컨벌루션 또는 조밀하게 연결된 계층 후에 사용될 수 있지만, 컨벌루션 또는 조밀하게 연결된 계층 이전에도 사용될 수 있다.

[0209] 배치 정규화는 입력을 피드-포워드하고 역방향 패스를 통해 파라미터 및 자체 입력에 대해 구배를 계산하기 위한 정의를 제공한다. 실제로, 배치 정규화 계층은 통상적으로 컨벌루션 또는 완전 연결된 계층 후, 그러나 출력이 활성화 함수에 공급되기 전에 삽입된다. 컨벌루션 계층의 경우, 서로 다른 위치에 있는 동일한 특징 맵의 상이한 요소(즉, 활성화)는 컨벌루션 속성을 따르기 위해 동일한 방식으로 정규화된다. 따라서, 미니 배치에서 모든 활성화는 활성화마다가 아닌 모든 위치에 걸쳐 정규화된다.

[0210] **1D 컨벌루션**

[0211] 본 접근법에 적용될 수 있는 컨벌루션 신경망의 구현에 사용되는 추가 기술은 서열로부터 로컬 1D 패치 또는 하위 서열을 추출하기 위해 1D 컨벌루션을 사용하는 것과 관련된다. 1D 컨벌루션 접근법은 입력 서열에서 윈도우 또는 패치로부터 각 출력 단계를 얻는다. 1D 컨벌루션 계층은 서열에서 로컬 패턴을 인식한다. 모든 패치에 대해 동일한 입력 변환이 수행되기 때문에, 입력 서열의 특정 위치에서 학습된 패턴은 나중에 다른 위치에서 인식될 수 있어, 1D 컨벌루션 계층 변환이 변환에 대해 불변하게 된다. 예를 들어, 크기 5의 컨벌루션 윈도우를 사용하여 염기의 서열을 처리하는 1D 컨벌루션 계층은 길이 5 이하의 염기 또는 염기 서열을 학습할 수 있어야 하며, 입력 서열의 모든 컨텍스트에서 기본 모티프를 인식할 수 있어야 한다. 따라서, 기본 수준의 1D 컨벌루션은 기본 형태에 대해 학습할 수 있다.

[0212] **전역 평균 풀링**

[0213] 현재 컨텍스트에서 유용하거나 활용될 수 있는 컨벌루션 신경망의 다른 양태는 전역 평균 풀링과 관련된다. 특히, 전역 평균 풀링은 채점을 위해 마지막 계층에 있는 특징의 공간 평균을 취함으로써 분류를 위해 완전 연결된(FC) 계층을 대체하는 데 사용될 수 있다. 이는 훈련 부하를 줄이고 오버피팅 문제를 우회한다. 전역 평균 풀링은 모델 이전에 구조를 적용하며 기정된 가중치를 사용하는 선형 변환과 동일하다. 전역 평균 풀링은 파라미터 수를 줄이고 완전 연결된 계층을 제거한다. 완전 연결된 계층은 통상적으로 가장 파라미터 및 연결 집약적인 계층으로서, 전역 평균 풀링은 유사한 결과를 달성하기 위해 훨씬 저렴한 접근법을 제공한다. 전역 평균 풀링의 주요 아이디어는 각 마지막 계층 특징 맵으로부터 평균 값을 채점을 위한 신뢰 요인으로 생성하여, 소프트맥스 계층에 직접 공급하는 것이다.

[0214] 전역 평균 풀링은 다음을 포함하나 이에 제한되지 않는 특정 이점을 제공할 수 있다: (1) 전역 평균 풀링 계층에 추가 파라미터가 없으므로 전역 평균 풀링 계층에서 오버피팅이 방지되고; (2) 전역 평균 풀링의 출력은 전체 특징 맵의 평균이므로, 전역 평균 풀링은 공간 변환에 강하고; (3) 일반적으로 전체 네트워크의 모든 파라미터에서 50% 넘게 차지하는 완전 연결된 계층에서의 엄청난 수의 파라미터로 인해, 이들을 전역 평균 풀링 계층으로 대체하면 모델의 크기를 상당히 줄일 수 있으며 이는 모델 압축에서 전역 평균 풀링을 매우 유용하게 만든다.

[0215] 전역 평균 풀링은 마지막 계층에서의 더 강력한 특징이 더 높은 평균 값을 가질 것으로 예상되므로 의미가 있다. 일부 구현예에서, 전역 평균 풀링은 분류 점수에 대한 프록시로 사용될 수 있다. 전역 평균 풀링 하에서 특징 맵은 신뢰 맵으로 해석될 수 있으며 특징 맵과 범주 간의 대응을 강제할 수 있다. 전역 평균 풀링은 마지막 계층 특징이 직접 분류를 위해 충분히 추상화된 경우 특히 효과적일 수 있다. 그러나, 다단계 특징을 부품 모델과 같은 그룹으로 조합되어야 하는 경우 전역 평균 풀링만으로는 충분하지 않거나 적합하지 않을 수 있으며, 이는 전역 평균 풀링 후에 간단한 완전 연결된 계층 또는 다른 분류기를 추가하여 더 적합하게 처리될 수 있다.

[0216] **IX.컴퓨터 시스템**

[0217] 이해할 수 있는 바와 같이, 설명된 신경망에 의해 출력된 병원성 분류기에 대해 수행되는 분석 및 처리뿐만 아니라 본 논의의 신경망 양태는 컴퓨터 시스템 또는 시스템에서 구현될 수 있다. 이를 염두에 두고, 추가 컨텍스트를 통해, 도 42는 현재 개시된 기술이 작동될 수 있는 예시적인 컴퓨팅 환경(800)을 도시하고 있다. 병원성 분류기(160), 2차 구조 서브네트워크(130), 및 용매 접근성 서브네트워크(132)를 갖는 심층 컨벌루션 신경망(102)은 하나 이상의 훈련 서버(802)(그 수는 처리될 데이터의 양 또는 계산 부하에 따라 조정될 수 있음)에서

훈련된다. 훈련 서버에 의해 접근, 생성, 및/또는 활용될 수 있는 이러한 접근법의 다른 양태는 훈련 프로세스에서 사용되는 훈련 데이터세트(810), 본원에서 논의된 바와 같은 양성 데이터세트 생성기(812). 및 본원에서 논의된 바와 같은 준지도 학습기(814) 양태를 포함하지만 이에 제한되지 않는다. 관리 인터페이스(816)는 훈련 서버 작동과의 상호 작용 및/또는 제어가 가능하도록 제공될 수 있다. 훈련된 모델의 출력은, 도 42에 도시된 바와 같이, 프로덕션 환경의 운영 및/또는 테스트에 사용하기 위해 프로덕션 서버(804)에 제공될 수 있는 테스트 데이터(820) 세트를 포함할 수 있지만 이에 제한되지 않는다.

[0218] 프로덕션 환경과 관련하여, 도 42에 도시된 바와 같이, 병원성 분류기(160), 2차 구조 서브네트워크(130), 및 용매 접근성 서브네트워크(132)를 갖는 훈련된 심층 컨벌루션 신경망(102)은 클라이언트 인터페이스(826)를 통해 요청 클라이언트로부터 입력 서열(예를 들어, 프로덕션 데이터(824))를 수신하는 하나 이상의 프로덕션 서버(804)에 배치된다. 프로덕션 서버(804)의 수는 사용자 수, 처리될 데이터의 양, 또는 더 일반적으로 계산 부하에 기초하여 조정될 수 있다. 프로덕션 서버(804)는 병원성 분류기(160), 2차 구조 서브네트워크(130), 및 용매 접근성 서브네트워크(132) 중 적어도 하나를 통해 입력 서열을 처리하여, 클라이언트 인터페이스(826)를 통해 클라이언트로 전송되는 출력(즉, 병원성 점수 또는 클래스를 포함할 수 있는 추론 데이터(828))을 생성한다. 추론 데이터(828)는 본원에서 논의된 바와 같이 병원성 점수 또는 분류기, 선택 계수, 고갈 메트릭, 보정 계수 또는 재보정된 메트릭, 히트맵, 대립유전자 빈도 및 누적 대립유전자 빈도 등을 포함할 수 있지만 이에 제한되지 않는다.

[0219] 훈련 서버(802), 프로덕션 서버(804), 관리 인터페이스(816), 및/또는 클라이언트 인터페이스(들)(826)를 실행하거나 지원하기 위해 활용될 수 있는 실제 하드웨어 아키텍처와 관련하여, 이러한 하드웨어는 물리적으로 하나 이상의 컴퓨터 시스템(예를 들어, 서버, 워크스테이션 등)으로 구현될 수 있다. 이러한 컴퓨터 시스템(850)에서 찾을 수 있는 구성요소의 예는 도 43에 도시되어 있지만, 본 예시는 이러한 시스템의 모든 실시예에서 찾을 수 없는 구성요소를 포함할 수 있거나 이러한 시스템에서 찾을 수 있는 모든 구성요소를 예시하지 않을 수 있음을 이해해야 한다. 더 나아가, 실제로 본 접근법의 양태는 가상 서버 환경에서 또는 클라우드 플랫폼의 일부로서 부분적으로 또는 전체적으로 구현될 수 있다. 그러나, 이러한 컨텍스트에서, 다양한 가상 서버 인스턴스화는 여전히 도 43과 관련하여 설명된 바와 같이 하드웨어 플랫폼에서 구현될 것이지만, 설명된 특정 기능적 양태는 가상 서버 인스턴스의 수준에서 구현될 수 있다.

[0220] 이를 염두에 두고, 도 43은 개시된 기술을 구현하는 데 사용될 수 있는 컴퓨터 시스템(850)의 단순화된 블록도이다. 컴퓨터 시스템(850)은 통상적으로, 버스 서브시스템(858)을 통해 다수의 주변 장치와 통신하는 적어도 하나의 프로세서(예를 들어, CPU)(854)를 포함한다. 이러한 주변 장치는, 예를 들어 메모리 장치(866)(예를 들어, RAM(874) 및 ROM(878)) 및 파일 저장 서브시스템(870)을 포함하는 저장 서브시스템(862), 사용자 인터페이스 입력 장치(882), 사용자 인터페이스 출력 장치(886), 및 네트워크 인터페이스 서브시스템(890)을 포함할 수 있다. 입력 및 출력 장치는 컴퓨터 시스템(850)과의 사용자 상호 작용을 가능하게 한다. 네트워크 인터페이스 서브시스템(890)은 다른 컴퓨터 시스템에서 대응하는 인터페이스 장치에 대한 인터페이스를 포함하여 외부 네트워크에 인터페이스를 제공한다.

[0221] 컴퓨터 시스템(850)이 병원성 분류기를 구현하거나 훈련하는 데 사용되는 일 구현예에서, 양성 데이터세트 생성기(812), 변이체 병원성 분류기(160), 2차 구조 분류기(130), 용매 접근성 분류기(132), 및 준지도 학습기(814)와 같은 신경망(102)은 저장 서브시스템(862) 및 사용자 인터페이스 입력 장치(882)에 통신 가능하게 연결된다.

[0222] 도시된 예에서, 컴퓨터 시스템(850)이 본원에서 논의된 바와 같이 신경망을 구현하거나 훈련하는 데 사용되는 컨텍스트에서, 하나 이상의 딥 러닝 프로세서(894)는 컴퓨터 시스템(850)의 일부로서 또는 이와 달리 컴퓨터 시스템(850)과 통신하여 존재할 수 있다. 이러한 실시예에서, 딥 러닝 프로세서는 GPU 또는 FPGA일 수 있으며 Google Cloud Platform, Xilinx 및 Cirrascale과 같은 딥 러닝 클라우드 플랫폼에 의해 호스팅될 수 있다. 딥 러닝 프로세서의 예로는 Google의 텐서 처리 유닛(TPU), GX4 랙마운트 시리즈와 같은 랙마운트 솔루션, GX8 랙마운트 시리즈, NVIDIA DGX-1, Microsoft의 Stratix V FPGA, Graphcore의 지능형 프로세서 유닛(IPU), Snapdragon 프로세서가 탑재된 Qualcomm의 Zeroth 플랫폼, NVIDIA의 Volta, NVIDIA의 DRIVE PX, NVIDIA의 JETSON TX1/TX2 MODULE, Intel의 Nirvana, Movidius VPU, Fujitsu DPI, ARM의 DynamicIQ, IBM TrueNorth 등을 포함한다.

[0223] 컴퓨터 시스템(850)의 컨텍스트에서, 사용자 인터페이스 입력 장치(882)는 키보드; 마우스, 트랙볼, 터치패드, 또는 그래픽 태블릿과 같은 포인팅 장치; 스캐너; 디스플레이 내에 통합된 터치 스크린; 음성 인식 시스템 및

마이크로폰과 같은 오디오 입력 장치; 및 다른 유형의 입력 장치를 포함할 수 있다. 일반적으로, "입력 장치"란 용어의 사용은 컴퓨터 시스템(850)에 정보를 입력하기 위한 모든 가능한 유형의 장치 및 방식을 포함하는 것으로 해석될 수 있다.

[0224] 사용자 인터페이스 출력 장치(886)는 디스플레이 서브시스템, 프린터, 팩스기, 또는 오디오 출력 장치와 같은 비시각적 디스플레이를 포함할 수 있다. 디스플레이 서브시스템은 음극선관(CRT), 액정 디스플레이(LCD)와 같은 평면 패널 장치, 프로젝션 장치, 또는 가시 이미지를 생성하기 위한 일부 다른 메커니즘을 포함할 수 있다. 디스플레이 서브시스템은 오디오 출력 장치와 같은 비시각적 디스플레이를 제공할 수도 있다. 일반적으로, "출력 장치"란 용어의 사용은 컴퓨터 시스템(850)으로부터 사용자로 또는 다른 기계나 컴퓨터 시스템으로 정보를 출력하기 위한 모든 가능한 유형의 장치 및 방식을 포함하는 것으로 해석될 수 있다.

[0225] 저장 서브시스템(862)은 본원에 설명된 모듈 및 방법의 일부 또는 전부의 기능을 제공하는 프로그래밍 및 데이터 구성을 저장한다. 이러한 소프트웨어 모듈은 일반적으로 프로세서(854) 단독으로 또는 다른 프로세서(854)와 조합하여 실행된다.

[0226] 저장 서브시스템(862)에 사용되는 메모리(866)는 프로그램 실행 동안 명령어 및 데이터의 저장을 위한 메인 랜덤 액세스 메모리(RAM)(878) 및 고정된 명령어가 저장되는 리드 온리 메모리(ROM)(874)를 포함하는 다수의 메모리를 포함할 수 있다. 파일 저장 서브시스템(870)은 프로그램 및 데이터 파일을 위한 영구 스토리지를 제공할 수 있고, 하드 디스크 드라이브, 관련된 착탈식 매체와 함께 플로피 디스크 드라이브, CD-ROM 드라이브, 광학 드라이브, 또는 착탈식 매체 카트리지를 포함할 수 있다. 특정 구현예의 기능을 구현하는 모듈은 저장 서브시스템(862) 내의 파일 저장 서브시스템(870)에 의해, 또는 프로세서(854)에 의해 액세스 가능한 다른 기계에 저장될 수 있다.

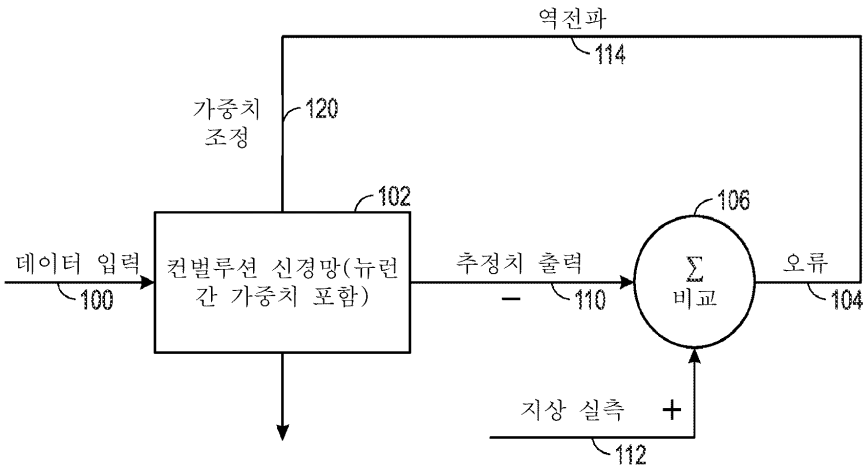
[0227] 버스 서브시스템(858)은 컴퓨터 시스템(850)의 다양한 구성요소 및 서브시스템이 의도된 대로 서로 통신하게 하기 위한 메커니즘을 제공한다. 버스 서브시스템(858)이 개략적으로 단일 버스로서 도시되어 있지만, 버스 서브시스템(858)의 대안적인 구현예는 다수의 버스를 사용할 수 있다.

[0228] 컴퓨터 시스템(850) 자체는 개인용 컴퓨터, 휴대용 컴퓨터, 워크스테이션, 컴퓨터 단말기, 네트워크 컴퓨터, 텔레비전, 메인프레임, 독립형 서버, 서버 팜, 광범위하게 분산된 느슨하게 네트워킹된 컴퓨터 세트, 또는 기타 데이터 처리 시스템이나 사용자 장치를 포함하는 다양한 유형일 수 있다. 끊임없이 변화하는 컴퓨터와 네트워크의 특성으로 인해, 도 43에 도시된 컴퓨터 시스템(850)의 설명은 개시된 기술을 예시하기 위한 특정 예로서만 의도된다. 도 43에 도시된 컴퓨터 시스템(850)보다 더 많거나 적은 구성요소를 갖는 컴퓨터 시스템(850)의 여러 다른 구성이 가능하다.

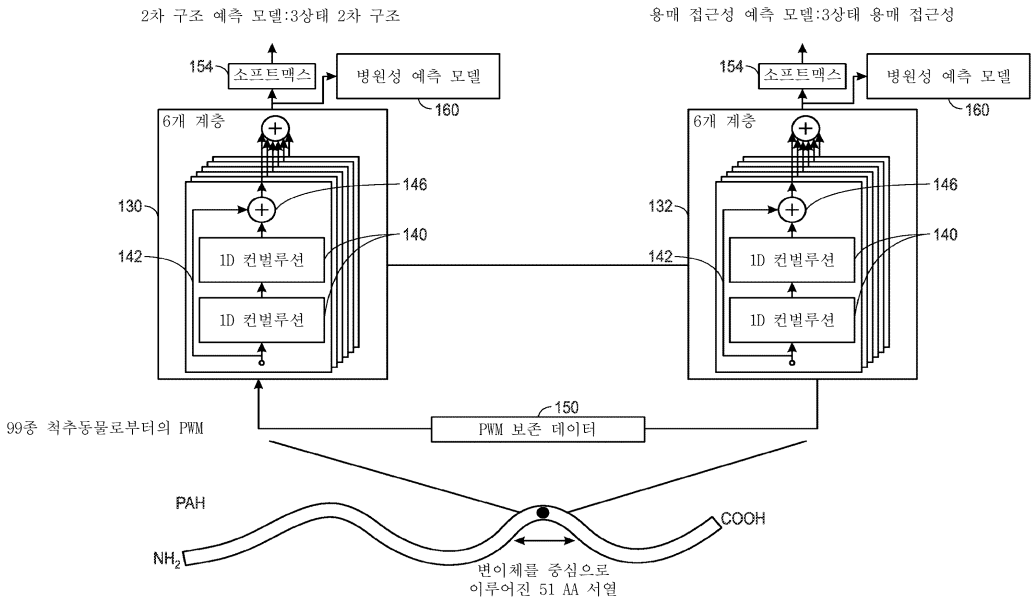
[0229] 이러한 서면 설명은 베스트 모드를 포함하여 본 발명을 개시하고 또한 임의의 장치 또는 시스템을 제조 및 사용하고 임의의 통합된 방법을 수행하는 것을 포함하여 당업자가 본 발명을 실시할 수 있도록 예시를 사용한다. 본 발명의 특허 가능한 범주는 청구범위에 의해 한정되며, 당업자에게 상기되는 다른 예를 포함할 수 있다. 이러한 다른 예는, 이들이 청구범위의 문자적 언어와 다르지 않은 구조적 요소를 갖는 경우 또는 이들이 청구범위의 문자적 언어와 실질적으로 다르지 않은 등가의 구조적 요소를 갖는 경우, 청구범위의 범주 내에 있는 것으로 의도된다.

도면

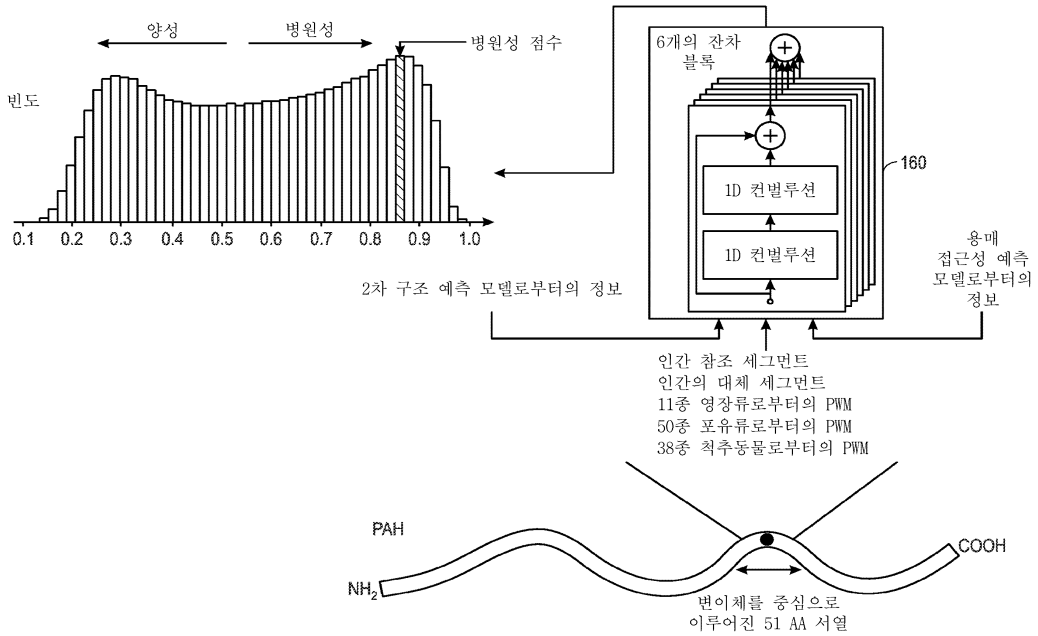
도면1


역전파
114
가중치
조정
120
102
106
데이터 입력
100
컨벌루션 신경망(뉴런
간 가중치 포함)
추정치 출력
110
Σ
비교
오류
104
지상 실측
112

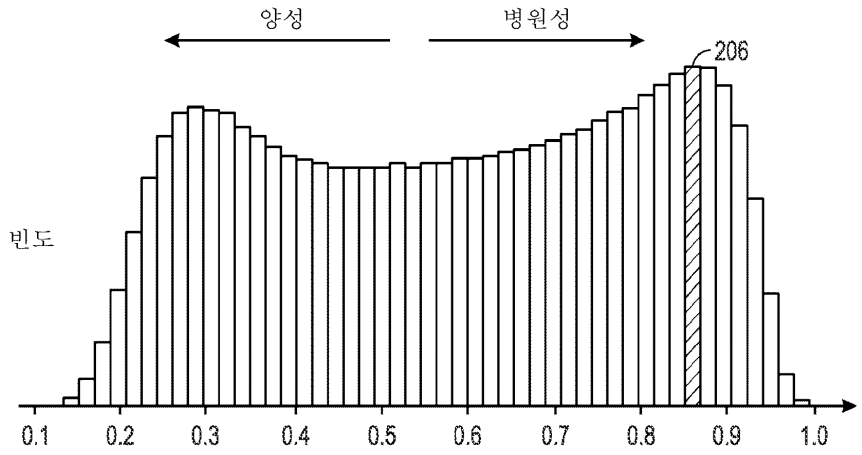
도면2


2차 구조 예측 모델:3상태 2차 구조
용매 접근성 예측 모델:3상태 용매 접근성
154
소프트맥스
병원성 예측 모델
160
6개 계층
130
132
142
146
140
1D 컨벌루션
150
PWM 보존 데이터
99종 척추동물로부터의 PWM
PAH
COOH
NH_2
변이체를 중심으로
이루어진 51 AA 서열

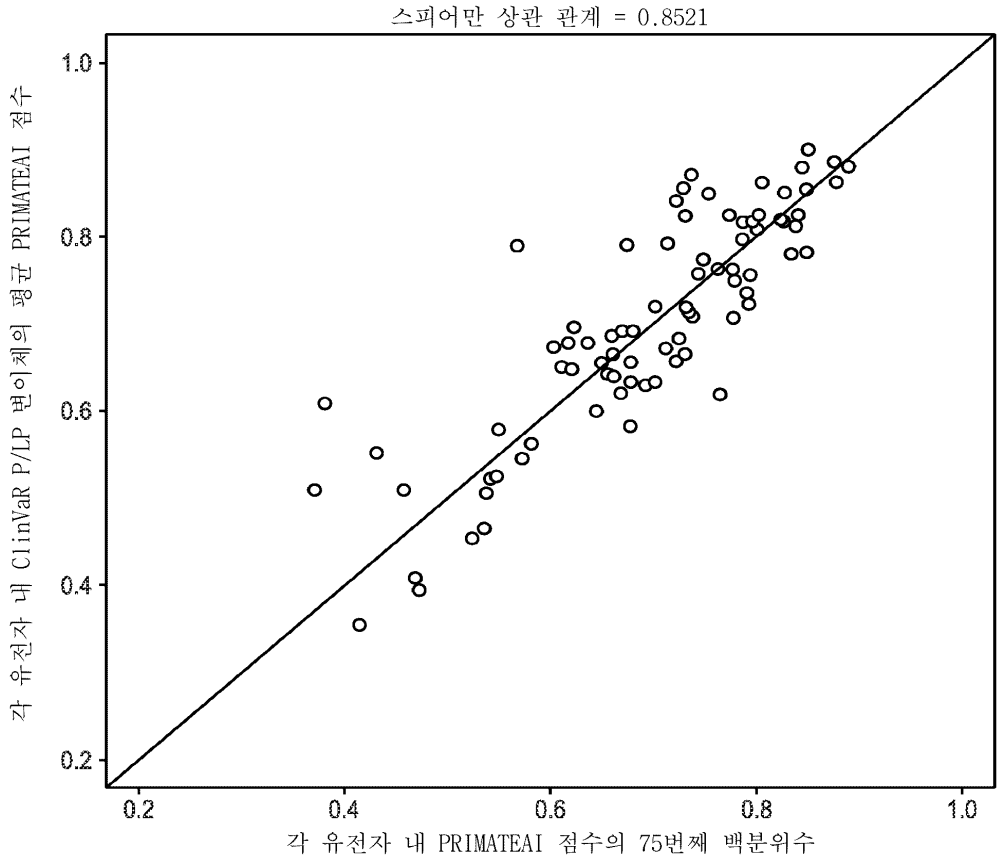
도면3


양성
병원성
병원성 점수
빈도
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
6개의 잔차 블록
1D 컨벌루션
1D 컨벌루션
160
2차 구조 예측 모델로부터의 정보
용매 접근성 예측 모델로부터의 정보
인간 참조 세그먼트
인간의 대체 세그먼트
11종 영장류로부터의 PWM
50종 포유류로부터의 PWM
38종 척추동물로부터의 PWM
PAH
NH2
COOH
변이체를 중심으로 이루어진 51 AA 서열

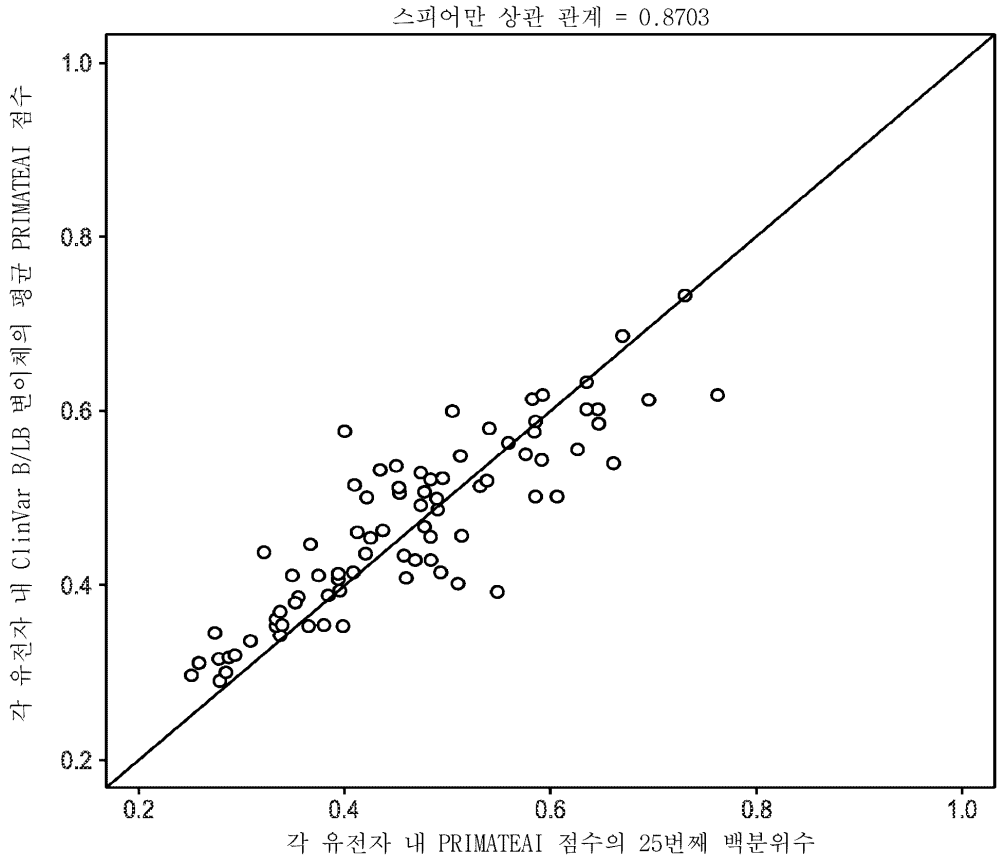
도면4


양성
병원성
206
빈도
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

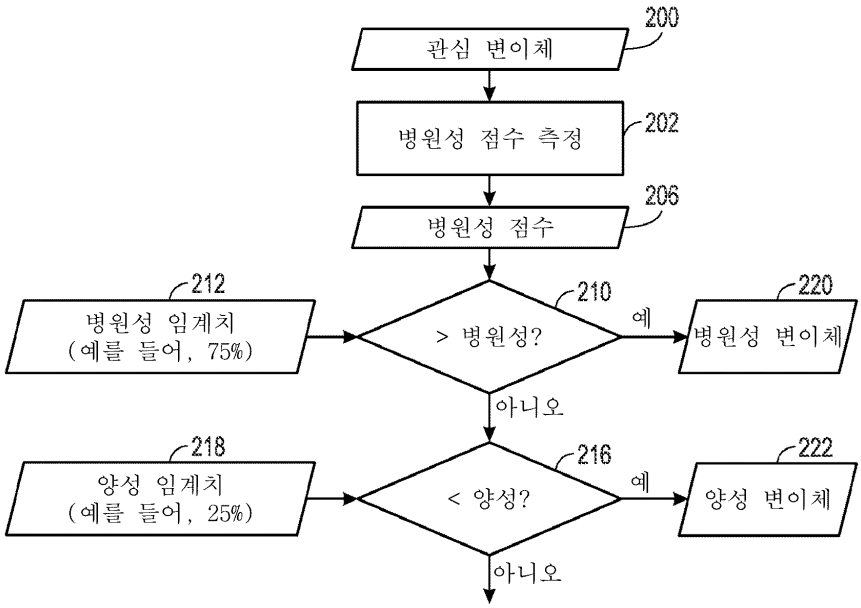
도면5


스피어만 상관 관계 = 0.8521
각 유전자 내 ClinVaR P/LP 변이체의 평균 PRIMATEAI 점수
각 유전자 내 PRIMATEAI 점수의 75번째 백분위수
1.0
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1.0

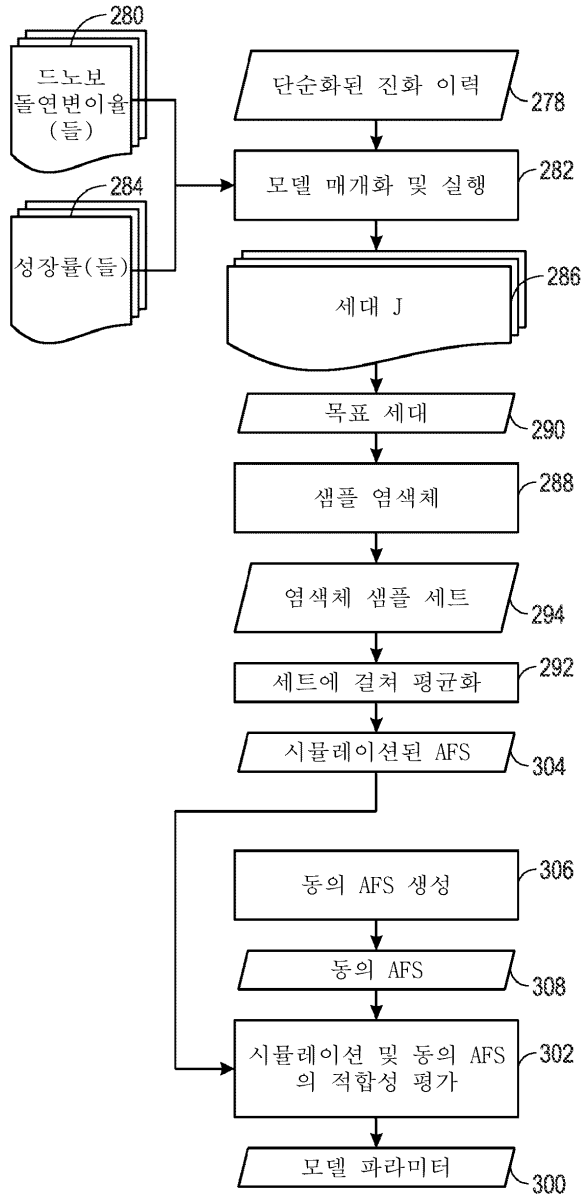
도면6


스피어만 상관 관계 = 0.8703
각 유전자 내 ClinVar B/LB 변이체의 평균 PRIMATEAI 점수
각 유전자 내 PRIMATEAI 점수의 25번째 백분위수
1.0
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1.0

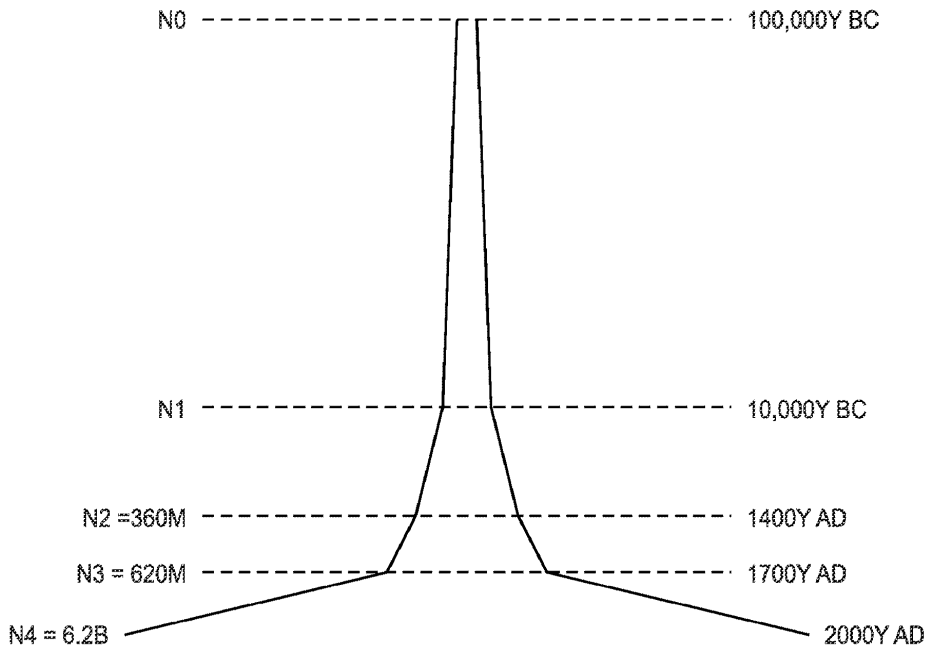
도면7


200
관심 변이체
202
병원성 점수 측정
206
병원성 점수
212
병원성 임계치
(예를 들어, 75%)
210
> 병원성?
예
220
병원성 변이체
아니오
218
양성 임계치
(예를 들어, 25%)
216
< 양성?
예
222
양성 변이체
아니오

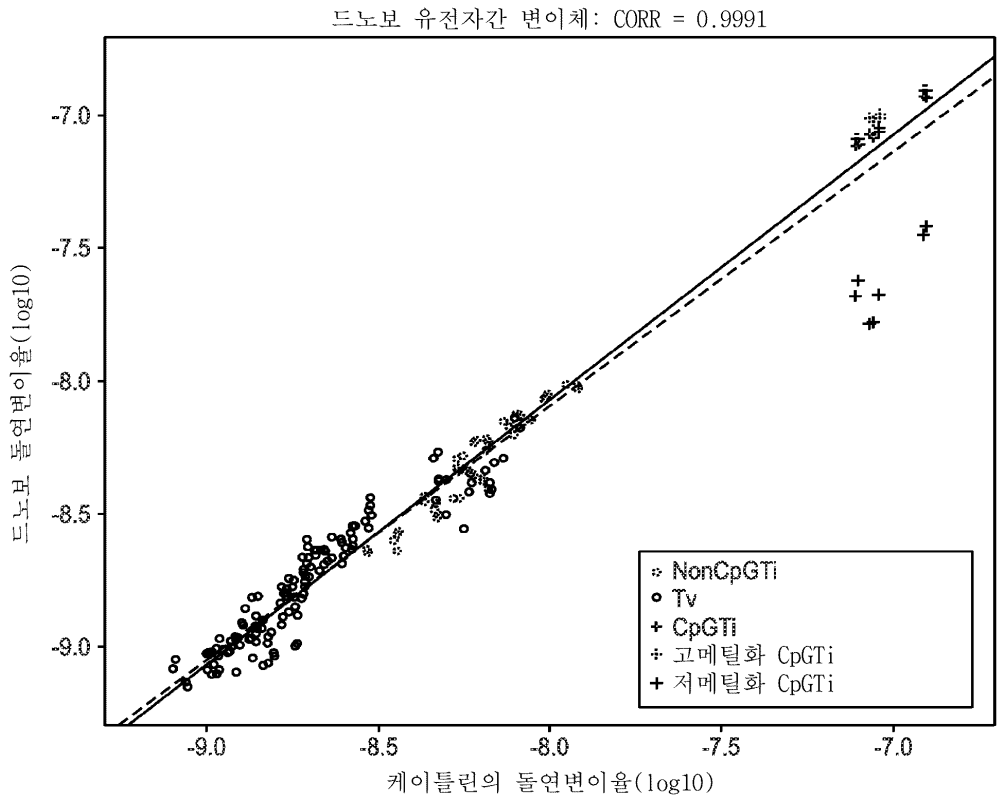
도면8


280
드노보 돌연변이율(들)
284
성장률(들)
단순화된 진화 이력
278
모델 매개화 및 실행
282
세대 J
286
목표 세대
290
샘플 염색체
288
염색체 샘플 세트
294
세트에 걸쳐 평균화
292
시뮬레이션된 AFS
304
동의 AFS 생성
306
동의 AFS
308
시뮬레이션 및 동의 AFS의 적합성 평가
302
모델 파라미터
300

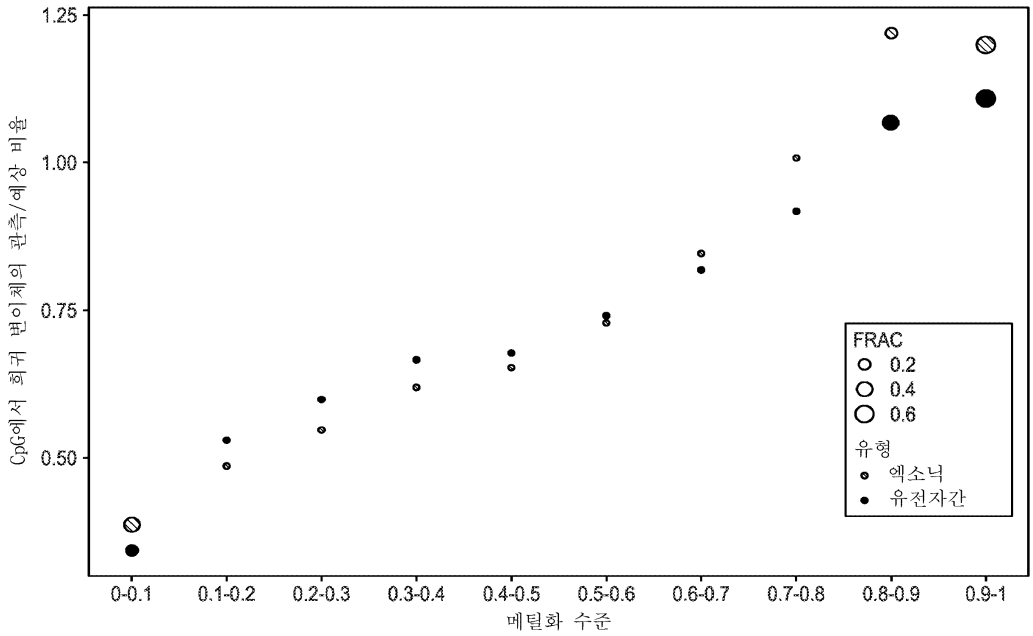
도면9


N0
100,000Y BC
N1
10,000Y BC
N2 =360M
1400Y AD
N3 = 620M
1700Y AD
N4 = 6.2B
2000Y AD

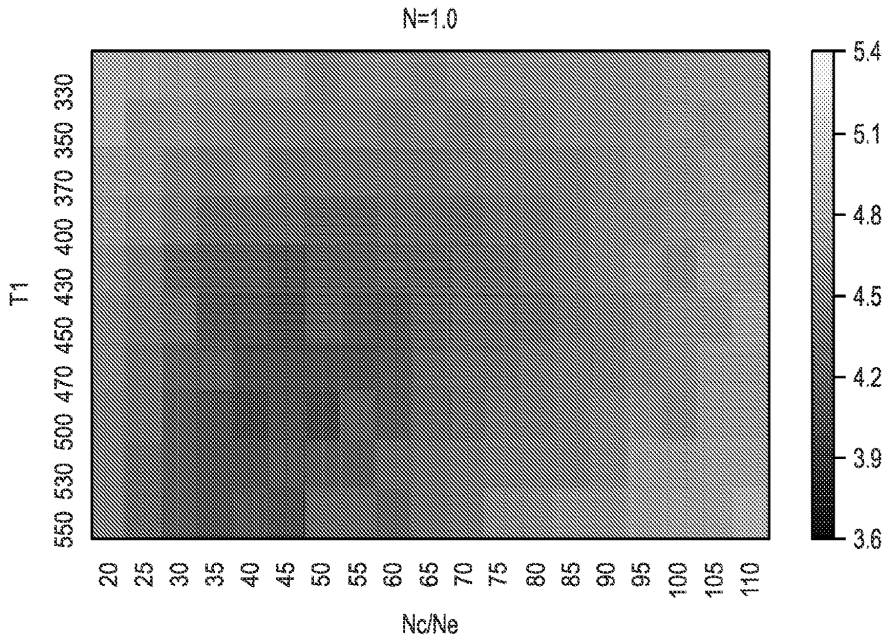
도면10


드노보 유전자간 변이체: CORR = 0.9991
-7.0
-7.5
-8.0
-8.5
-9.0
드노보 돌연변이율(log10)
-9.0
-8.5
-8.0
-7.5
-7.0
케이틀린의 돌연변이율(log10)
NonCpGTi
Tv
CpGTi
고메틸화 CpGTi
저메틸화 CpGTi

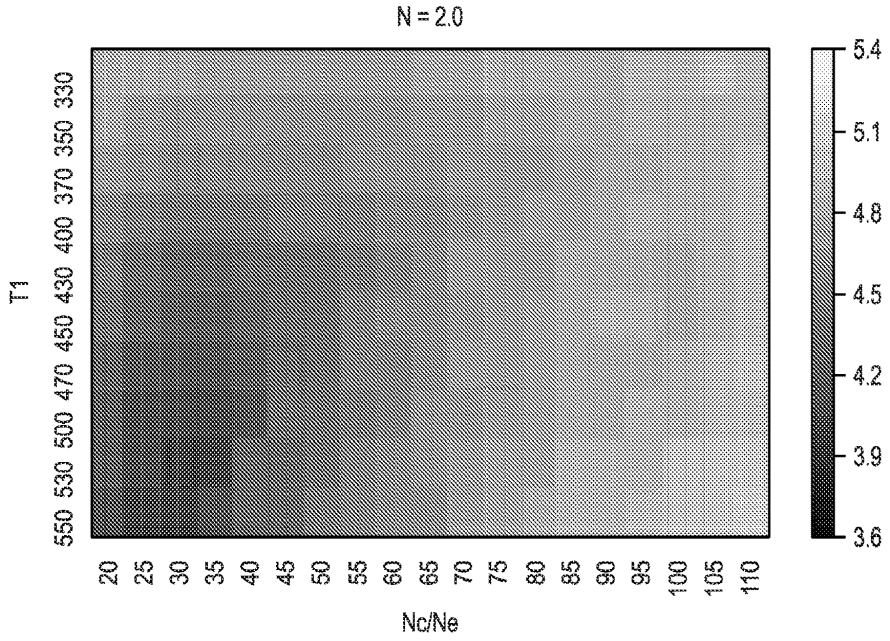
도면11


CpG에서 회귀 변이체의 관측/예상 비율
1.25
1.00
0.75
0.50
0-0.1
0.1-0.2
0.2-0.3
0.3-0.4
0.4-0.5
0.5-0.6
0.6-0.7
0.7-0.8
0.8-0.9
0.9-1
메틸화 수준
FRAC
0.2
0.4
0.6
유형
엑소닉
유전자간

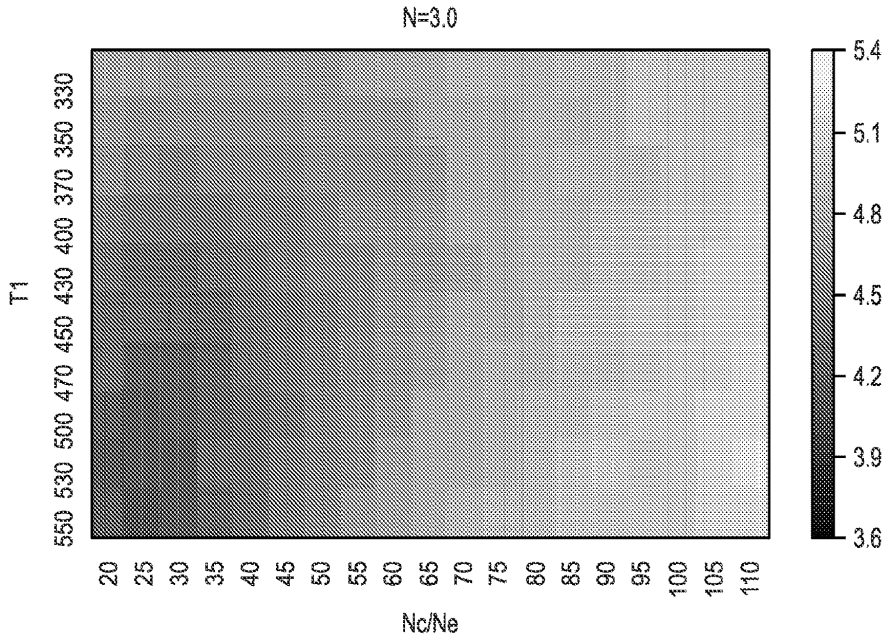
도면12a


N=1.0
T1
550 530 500 470 450 430 400 370 350 330
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110
Nc/Ne
5.4
5.1
4.8
4.5
4.2
3.9
3.6

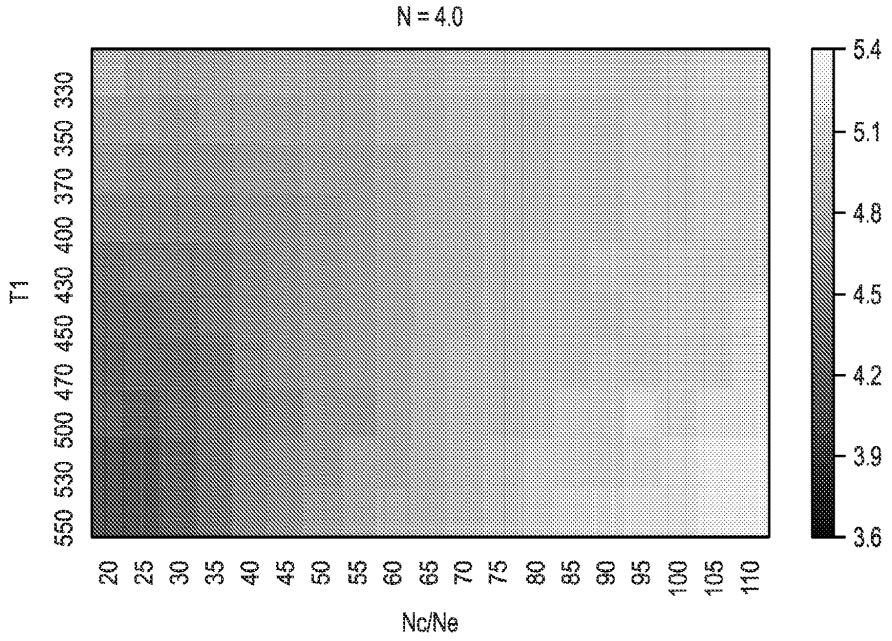
도면12b


N = 2.0
T1
330
350
370
400
430
450
470
500
530
550
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
110
Nc/Ne
5.4
5.1
4.8
4.5
4.2
3.9
3.6

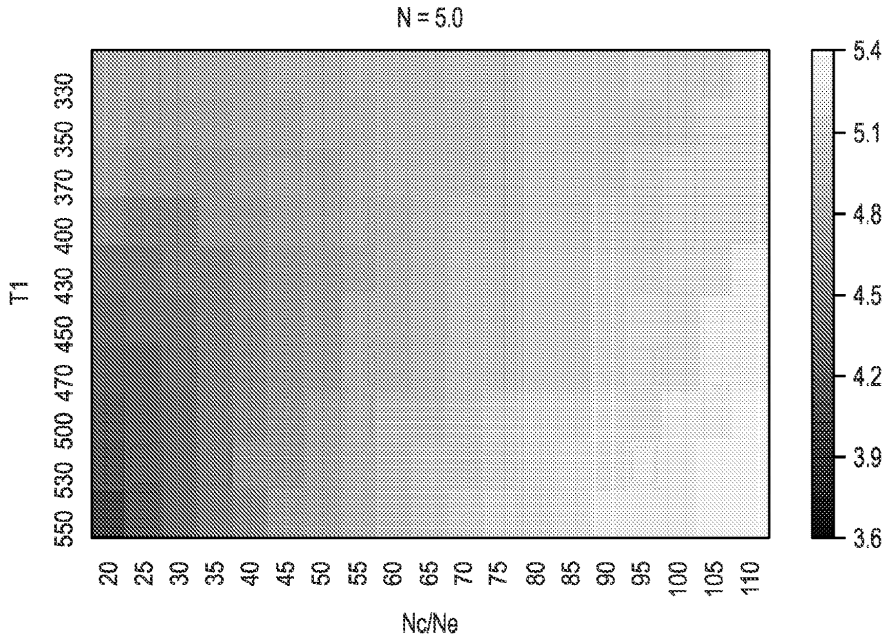
도면12c


N=3.0
T1
330
350
370
400
430
450
470
500
530
550
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
110
Nc/Ne
5.4
5.1
4.8
4.5
4.2
3.9
3.6

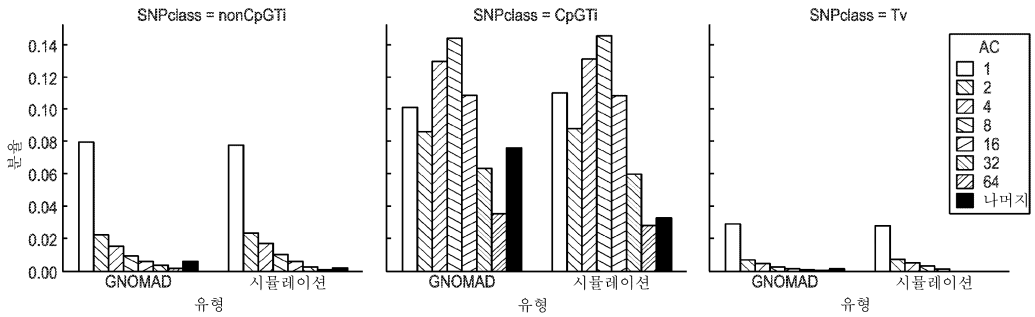
도면12d



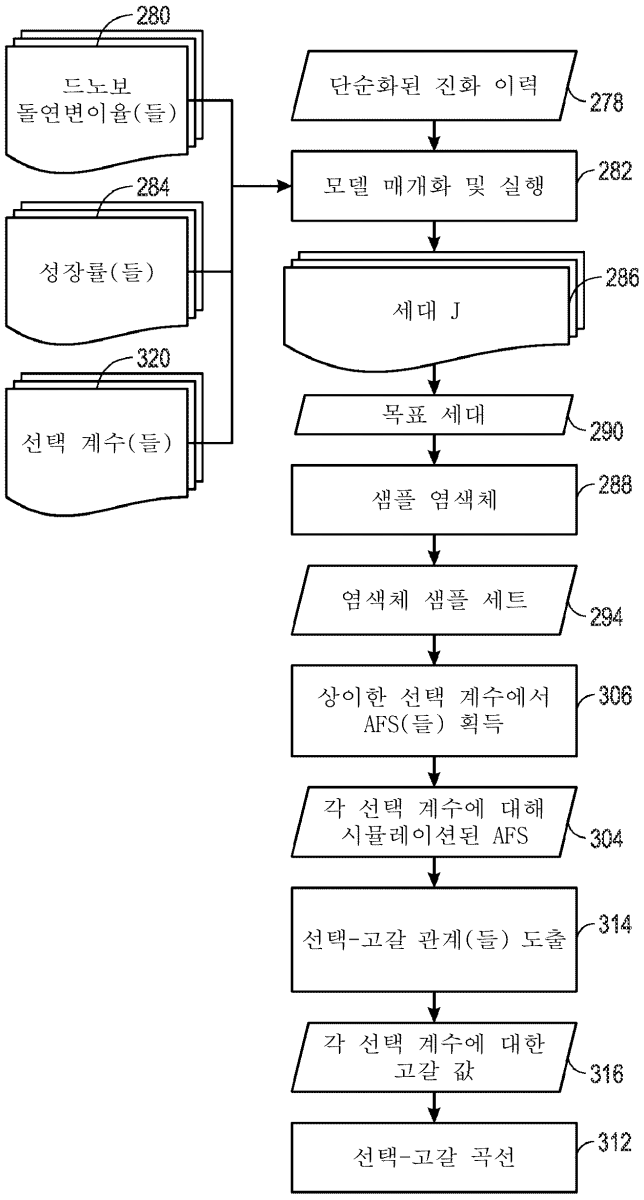
도면12e



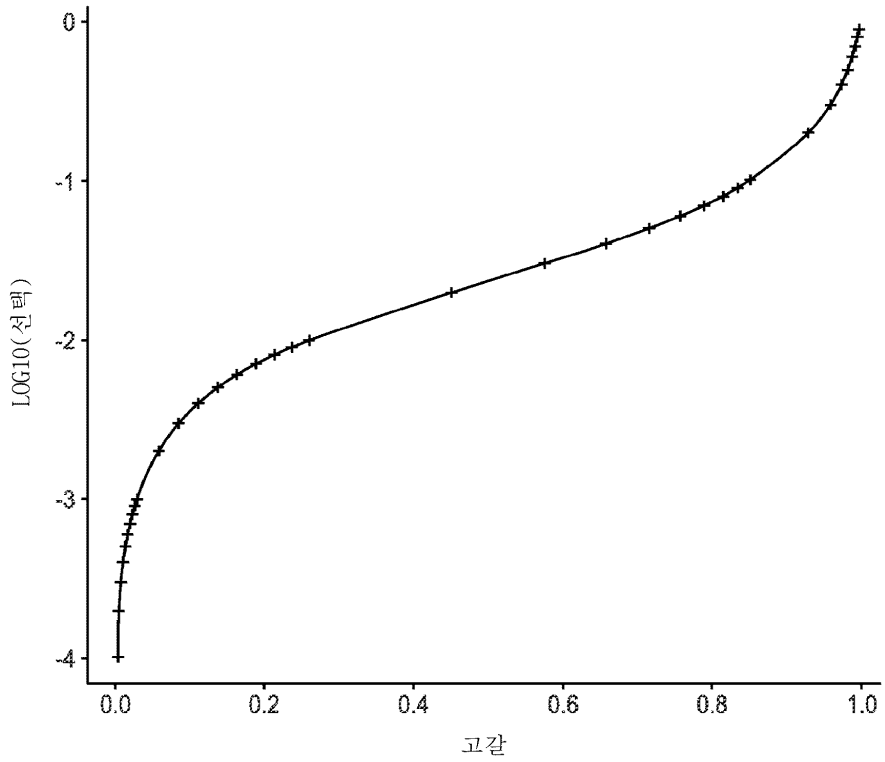
도면13



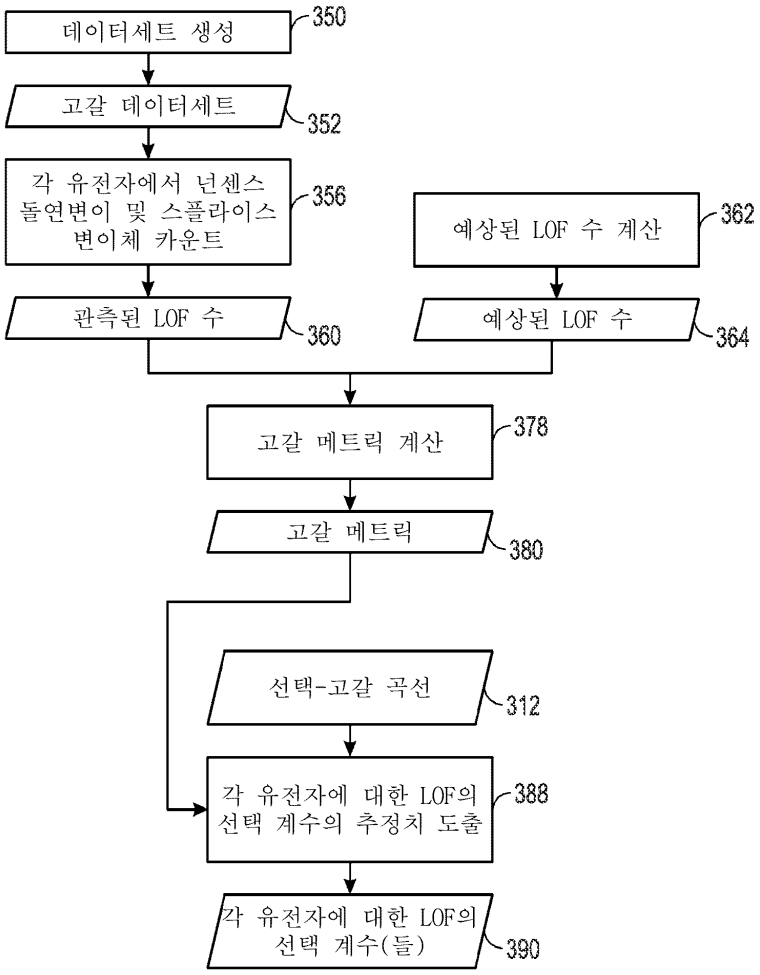
도면14


280
드노보 돌연변이율(들)
284
성장률(들)
320
선택 계수(들)
단순화된 진화 이력
278
모델 매개화 및 실행
282
세대 J
286
목표 세대
290
샘플 염색체
288
염색체 샘플 세트
294
상이한 선택 계수에서 AFS(들) 획득
306
각 선택 계수에 대해 시뮬레이션된 AFS
304
선택-고갈 관계(들) 도출
314
각 선택 계수에 대한 고갈 값
316
선택-고갈 곡선
312

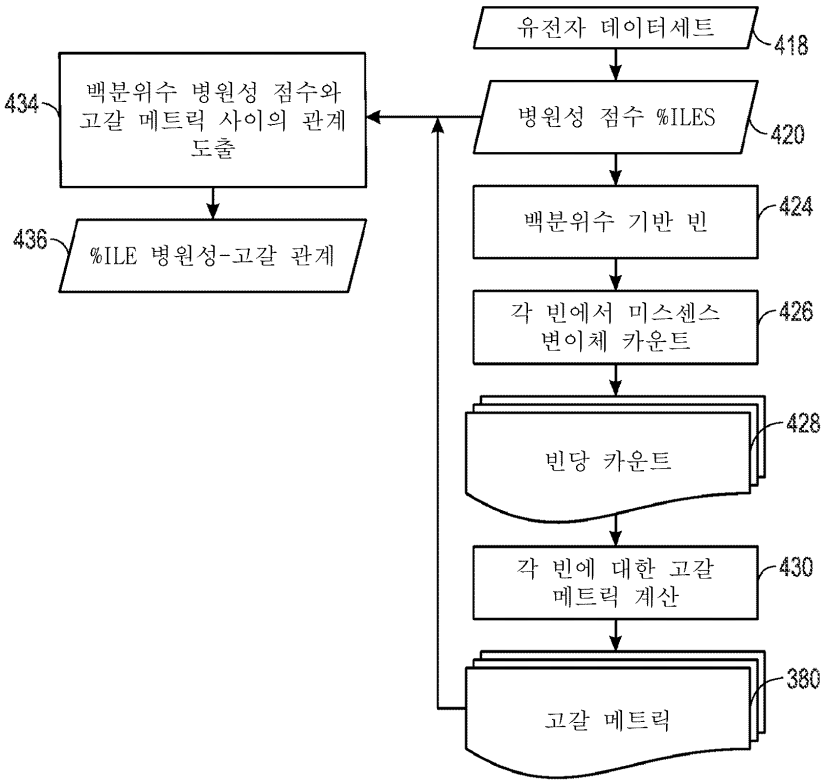
도면15



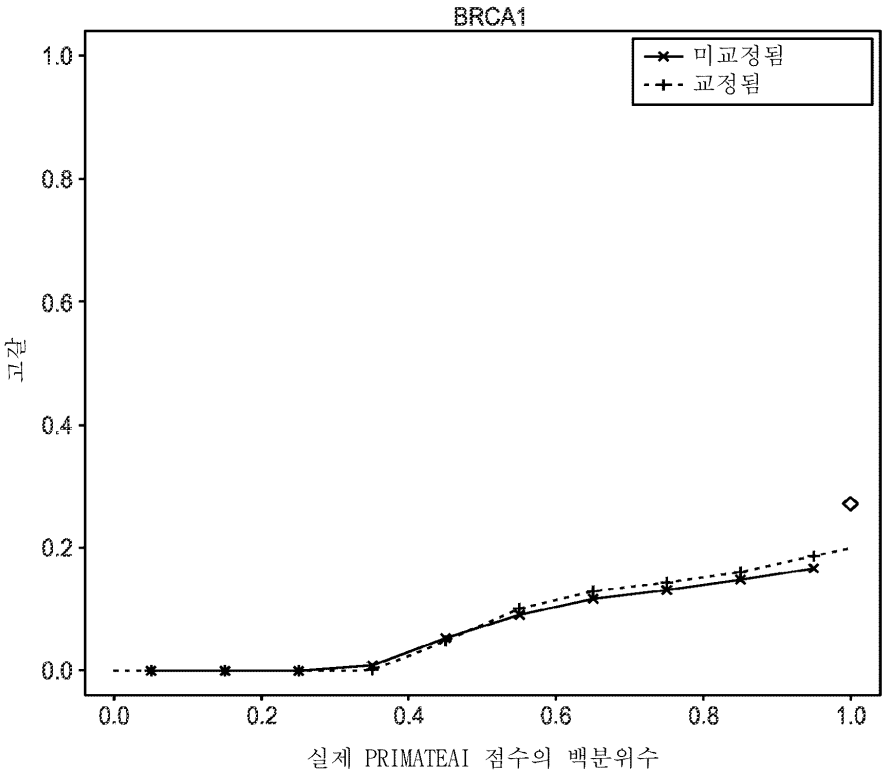
도면16


데이터세트 생성
350
고갈 데이터세트
352
각 유전자에서 넌센스 돌연변이 및 스플라이스 변이체 카운트
356
관측된 LOF 수
360
예상된 LOF 수 계산
362
예상된 LOF 수
364
고갈 메트릭 계산
378
고갈 메트릭
380
선택-고갈 곡선
312
각 유전자에 대한 LOF의 선택 계수의 추정치 도출
388
각 유전자에 대한 LOF의 선택 계수(들)
390

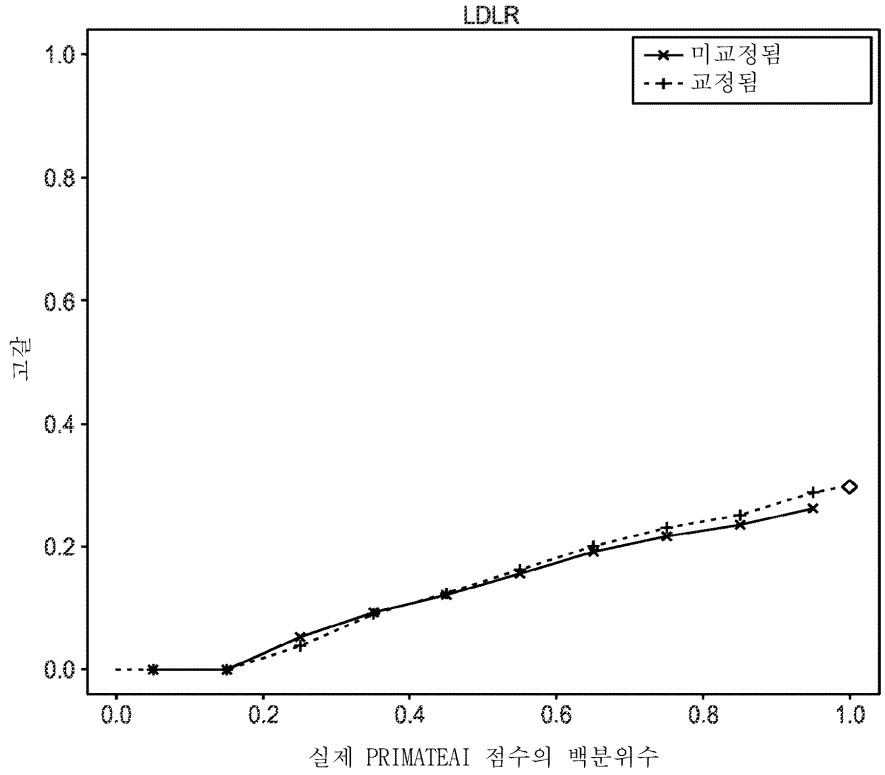
도면17


유전자 데이터세트
418
병원성 점수 %ILES
420
백분위수 기반 빈
424
각 빈에서 미스센스 변이체 카운트
426
빈당 카운트
428
각 빈에 대한 고갈 메트릭 계산
430
고갈 메트릭
380
백분위수 병원성 점수와 고갈 메트릭 사이의 관계 도출
434
%ILE 병원성-고갈 관계
436

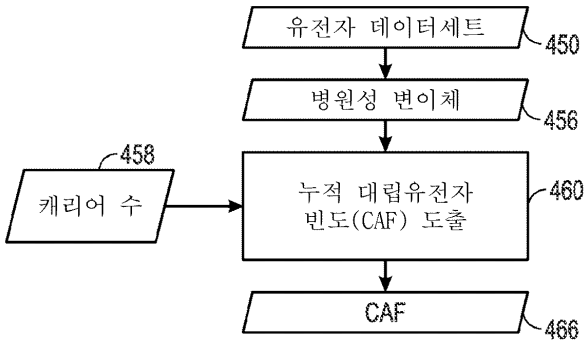
도면18


BRCA1
미교정됨
교정됨
고갈
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
실제 PRIMATEAI 점수의 백분위수

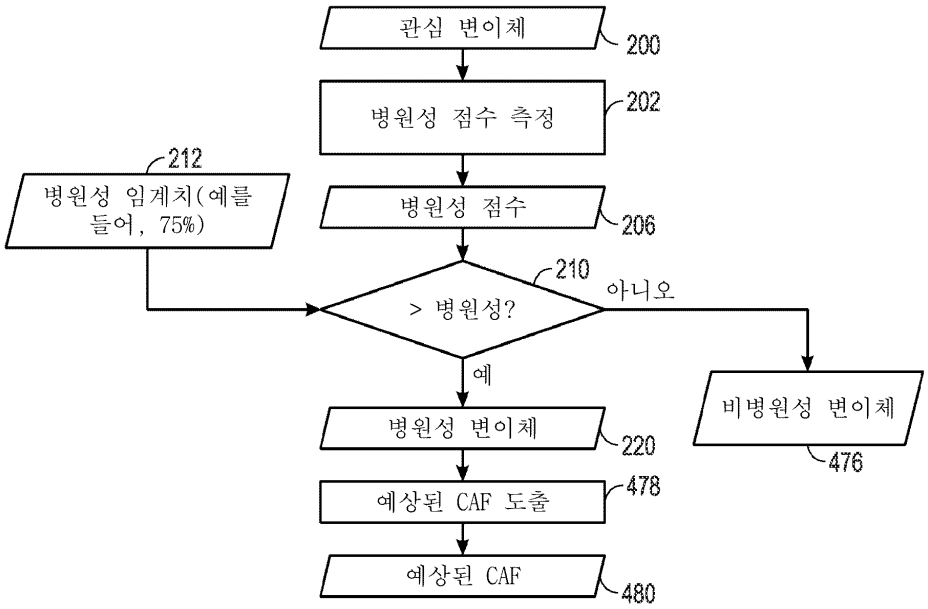
도면19


LDLR
미교정됨
교정됨
1.0
0.8
0.6
0.4
0.2
0.0
고갈
0.0
0.2
0.4
0.6
0.8
1.0
실제 PRIMATEAI 점수의 백분위수

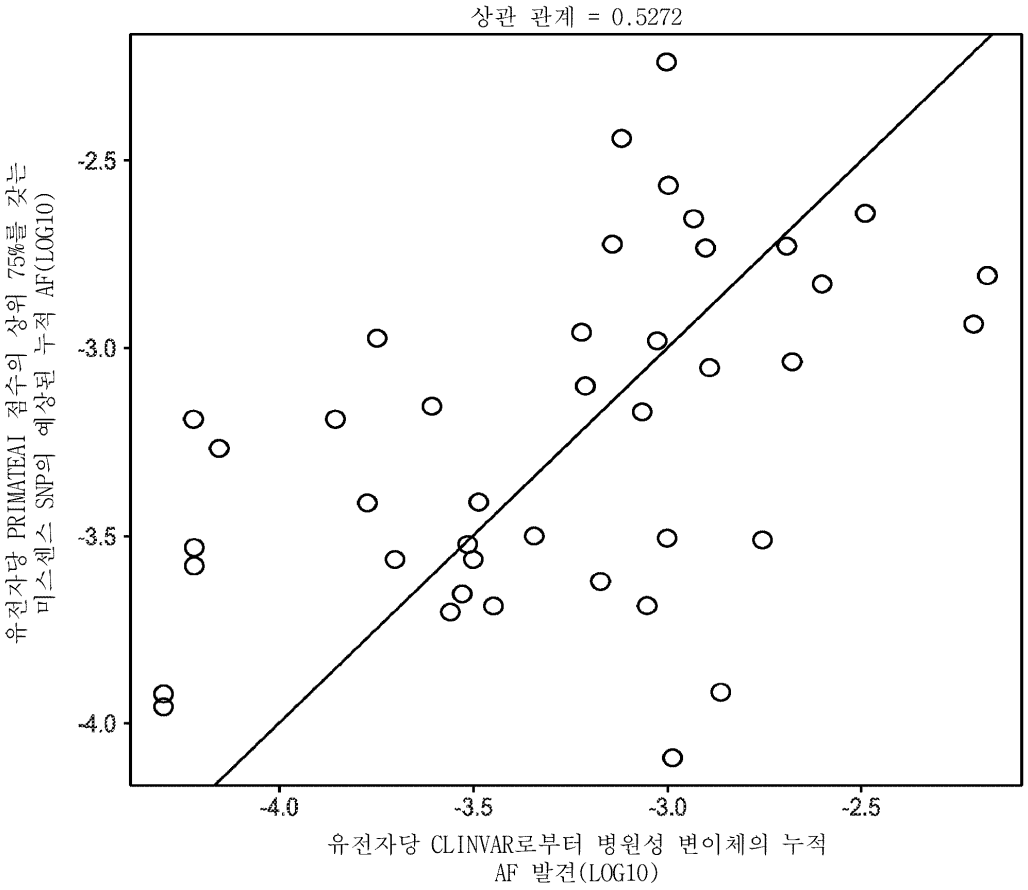
도면20


유전자 데이터세트
450
병원성 변이체
456
458
캐리어 수
누적 대립유전자
빈도(CAF) 도출
460
CAF
466

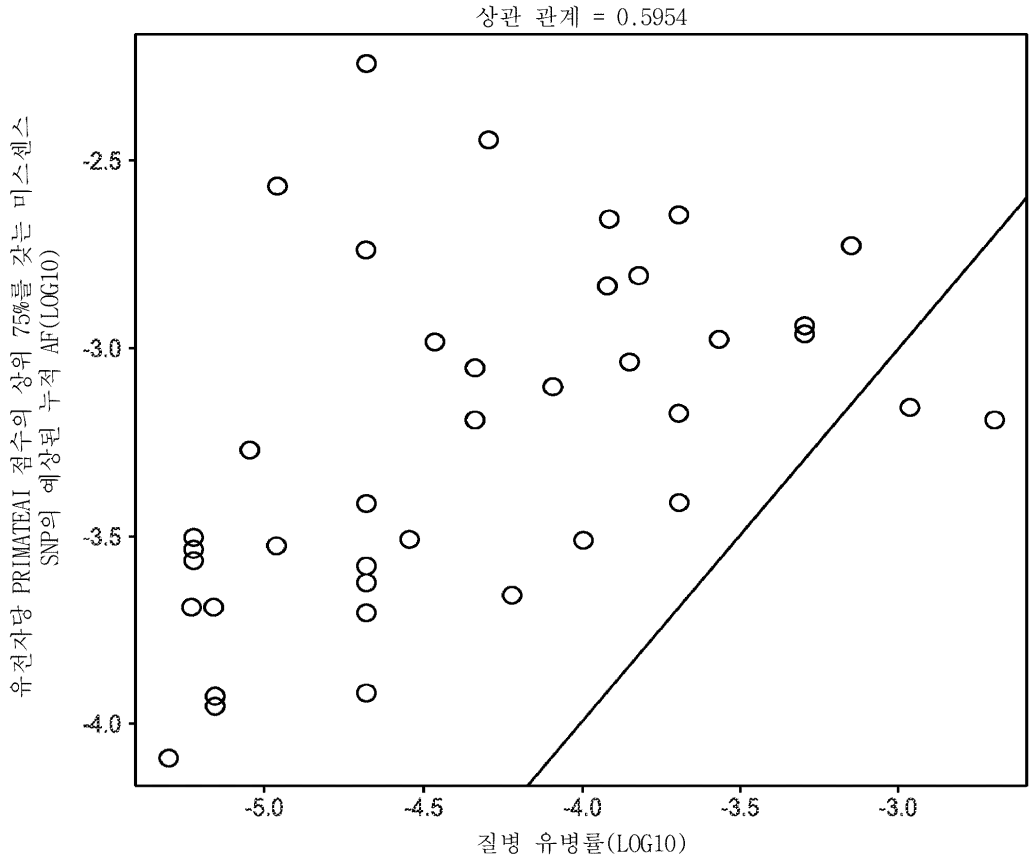
도면21


관심 변이체
200
병원성 점수 측정
202
212
병원성 임계치(예를 들어, 75%)
병원성 점수
206
210
> 병원성?
아니오
예
비병원성 변이체
476
병원성 변이체
220
478
예상된 CAF 도출
예상된 CAF
480

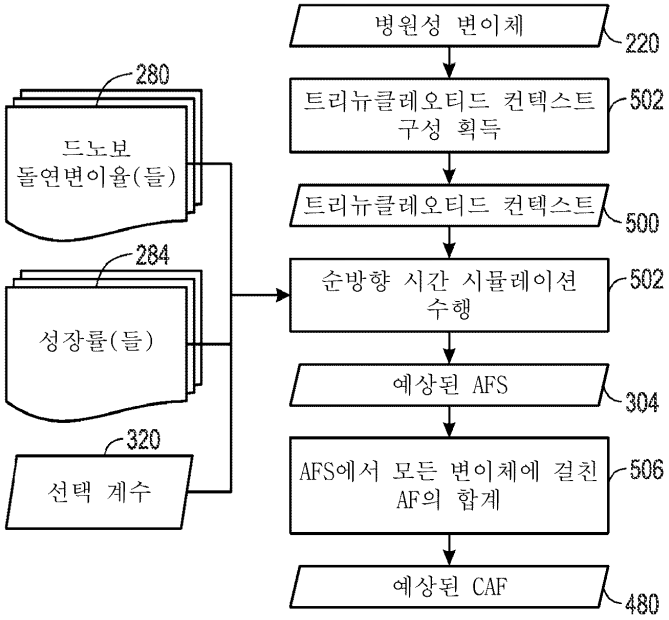
도면22


상관 관계 = 0.5272
유전자당 PRIMATEAI 점수의 상위 75%를 갖는 미스센스 SNP의 예상된 누적 AF(LOG10)
-2.5
-3.0
-3.5
-4.0
-4.0
-3.5
-3.0
-2.5
유전자당 CLINVAR로부터 병원성 변이체의 누적 AF 발견(LOG10)

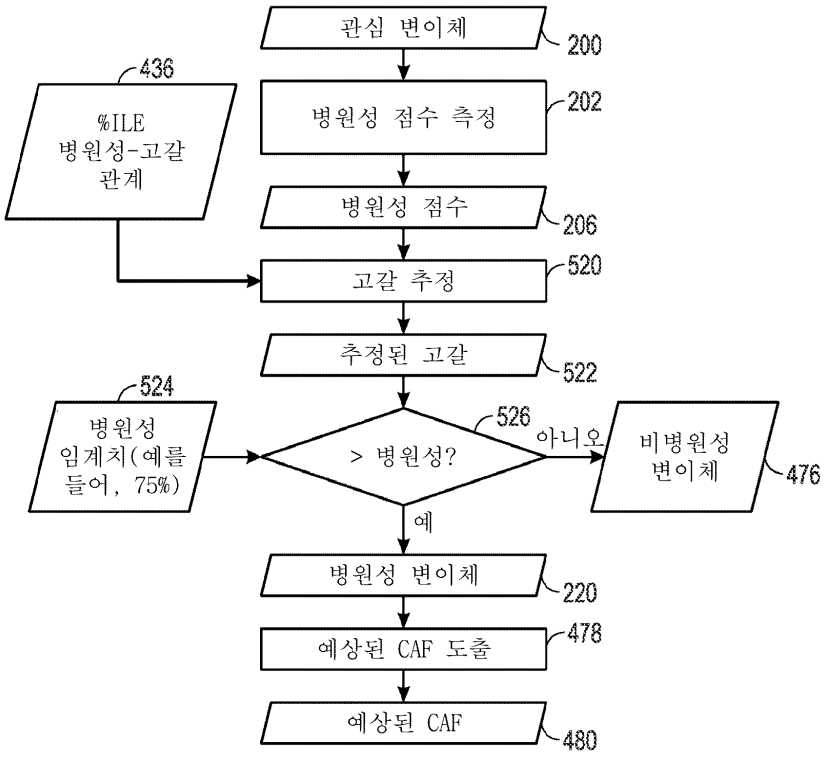
도면23


상관 관계 = 0.5954
유전자당 PRIMATEAI 점수의 상위 75%를 갖는 미스센스 SNP의 예상된 누적 AF(LOG10)
-2.5
-3.0
-3.5
-4.0
-5.0
-4.5
-4.0
-3.5
-3.0
질병 유병률(LOG10)

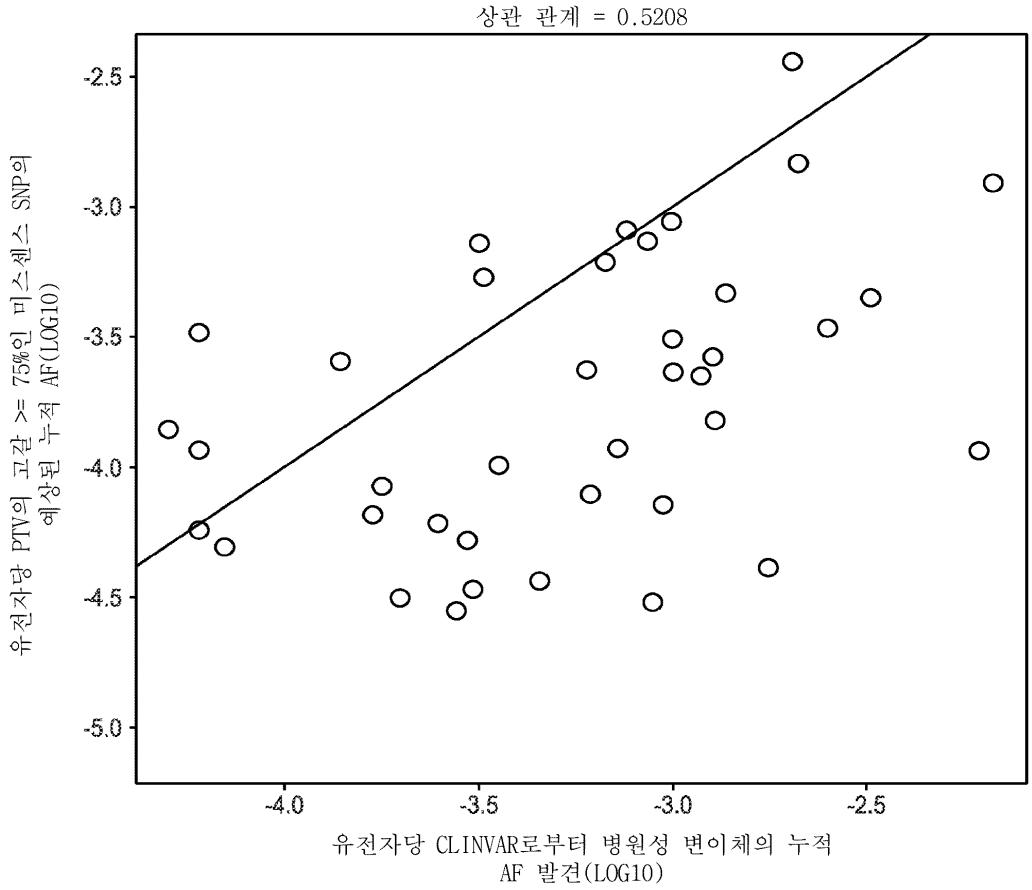
도면24


병원성 변이체
220
트리뉴클레오티드 컨텍스트 구성 획득
502
트리뉴클레오티드 컨텍스트
500
순방향 시간 시뮬레이션 수행
502
예상된 AFS
304
AFS에서 모든 변이체에 걸친 AF의 합계
506
예상된 CAF
480
280
드노보 돌연변이율(들)
284
성장률(들)
320
선택 계수

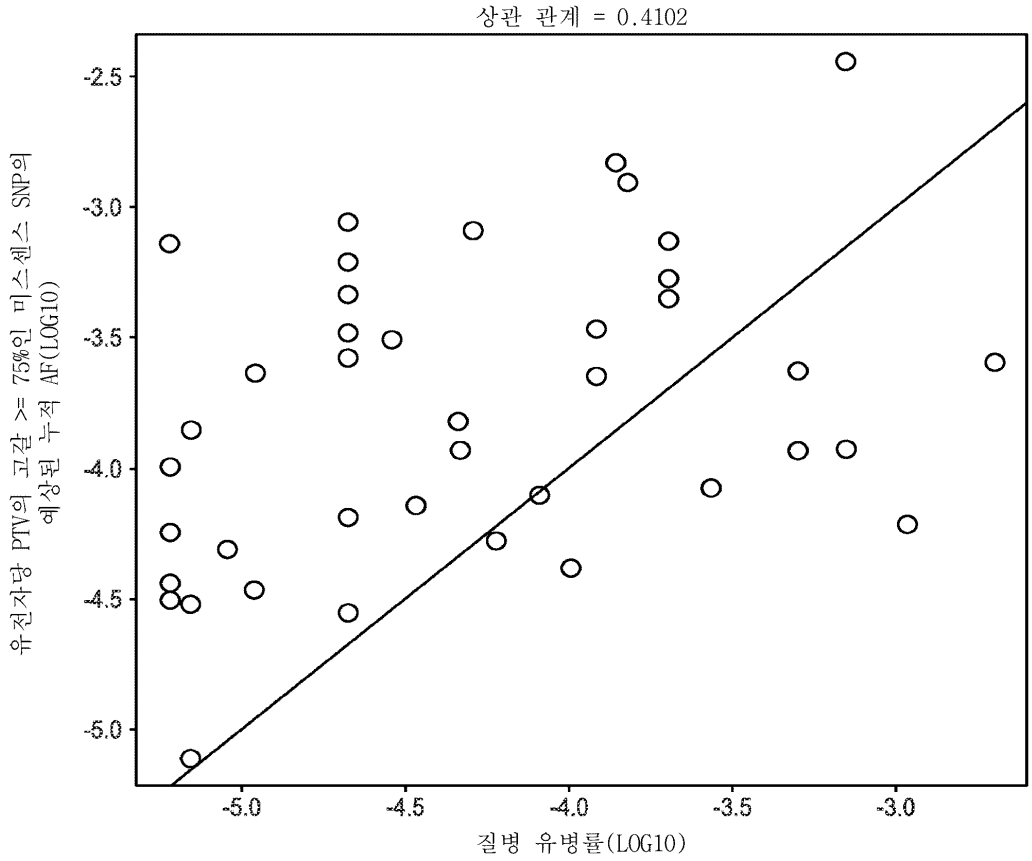
도면25


관심 변이체
200
436
%ILE
병원성-고갈
관계
병원성 점수 측정
202
병원성 점수
206
고갈 추정
520
추정된 고갈
522
524
병원성
임계치(예를
들어, 75%)
526
> 병원성?
아니오
비병원성
변이체
476
예
병원성 변이체
220
예상된 CAF 도출
478
예상된 CAF
480

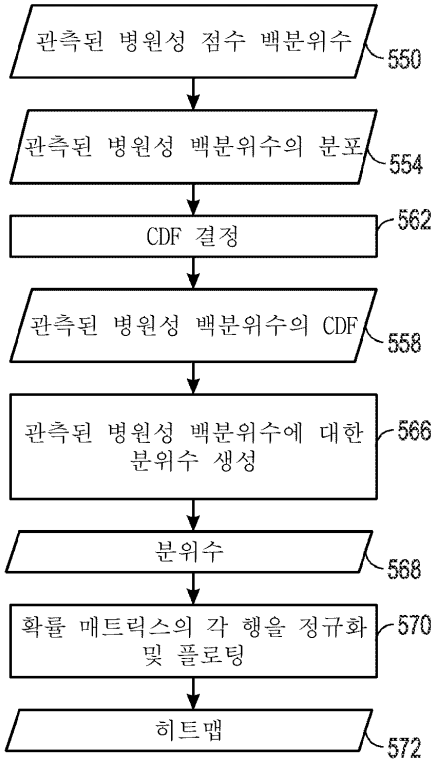
도면26


상관 관계 = 0.5208
유전자당 PTV의 고갈 >= 75%인 미스센스 SNP의
예상된 누적 AF(LOG10)
-2.5
-3.0
-3.5
-4.0
-4.5
-5.0
-4.0
-3.5
-3.0
-2.5
유전자당 CLINVAR로부터 병원성 변이체의 누적
AF 발견(LOG10)

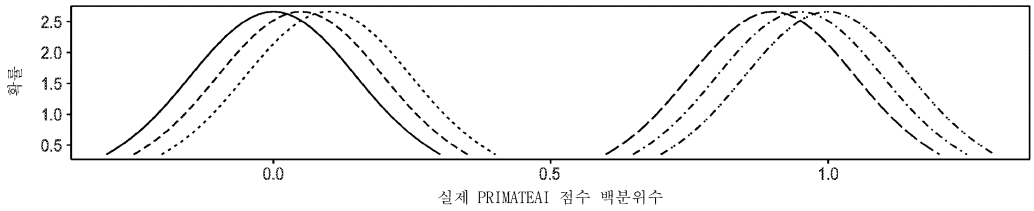
도면27


상관 관계 = 0.4102
유전자당 PTV의 고갈 >= 75%인 미스센스 SNP의 예상된 누적 AF(LOG10)
-2.5
-3.0
-3.5
-4.0
-4.5
-5.0
-5.0
-4.5
-4.0
-3.5
-3.0
질병 유병률(LOG10)

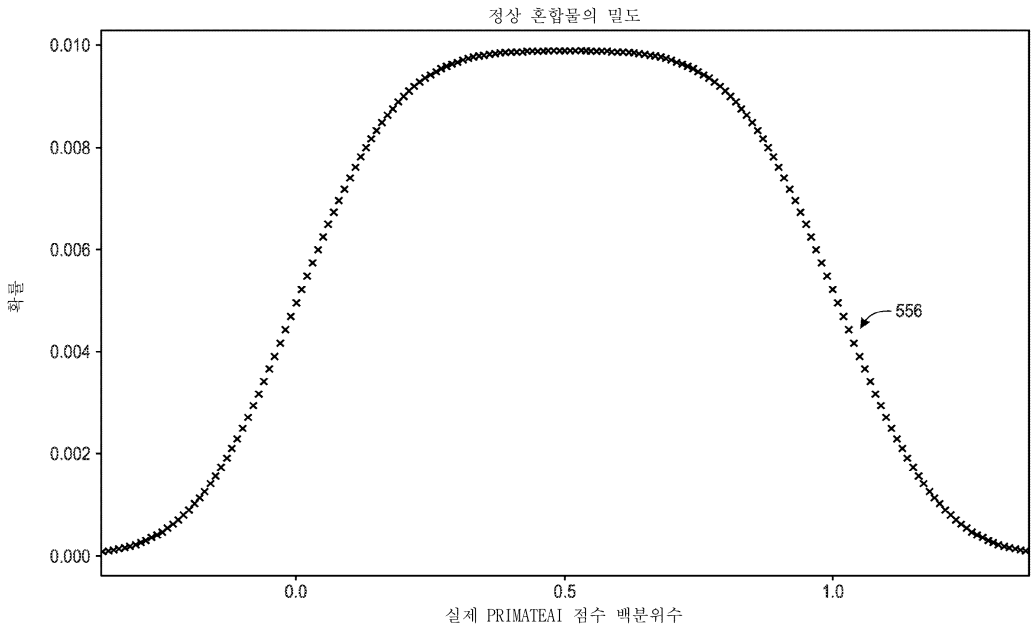
도면28


관측된 병원성 점수 백분위수
550
관측된 병원성 백분위수의 분포
554
CDF 결정
562
관측된 병원성 백분위수의 CDF
558
관측된 병원성 백분위수에 대한 분위수 생성
566
분위수
568
확률 매트릭스의 각 행을 정규화 및 플로팅
570
히트맵
572

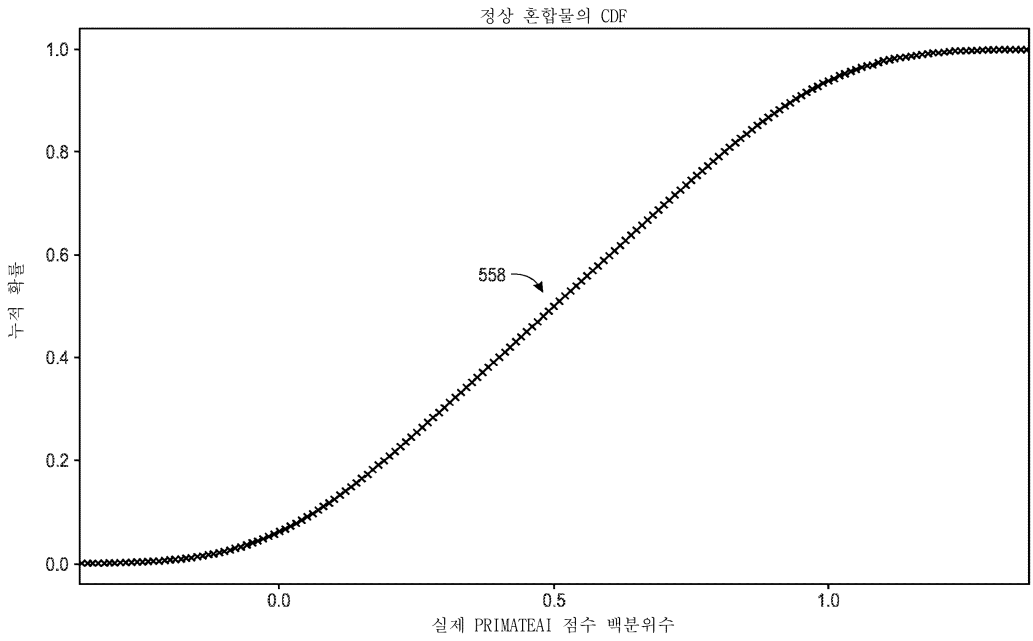
도면29



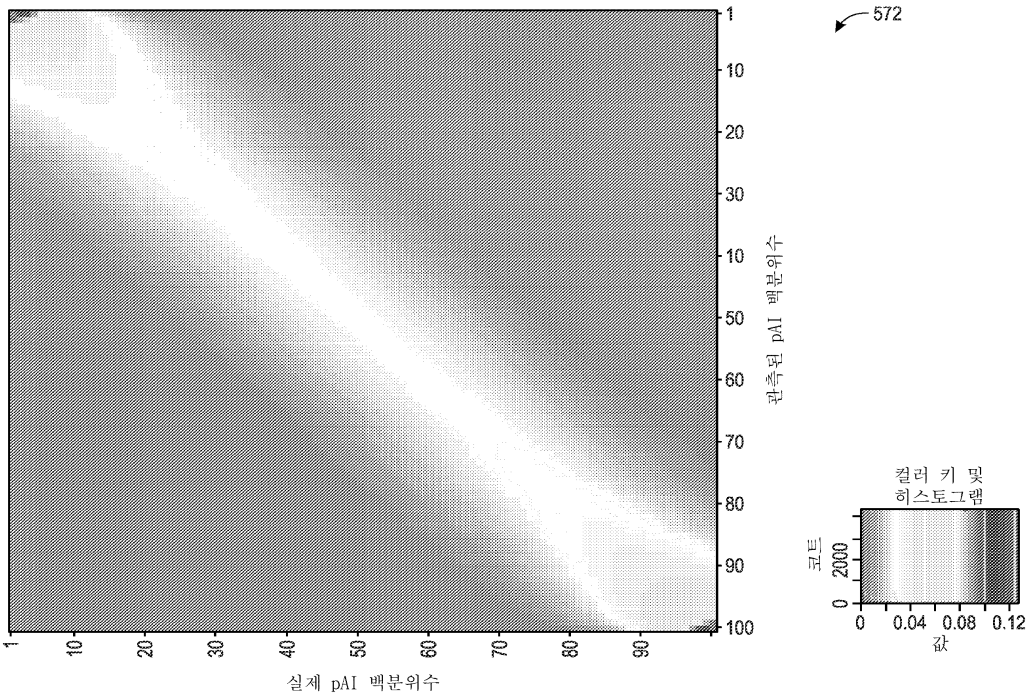
도면30



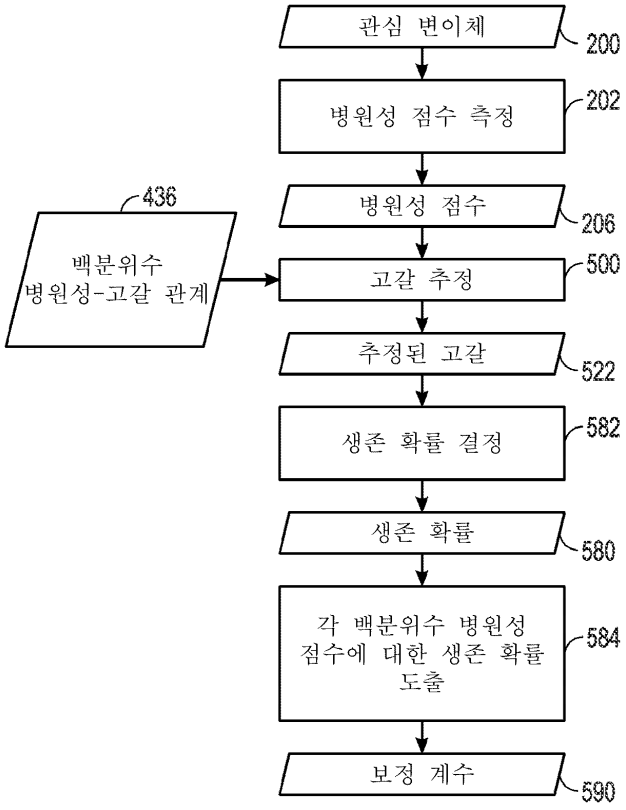
도면31



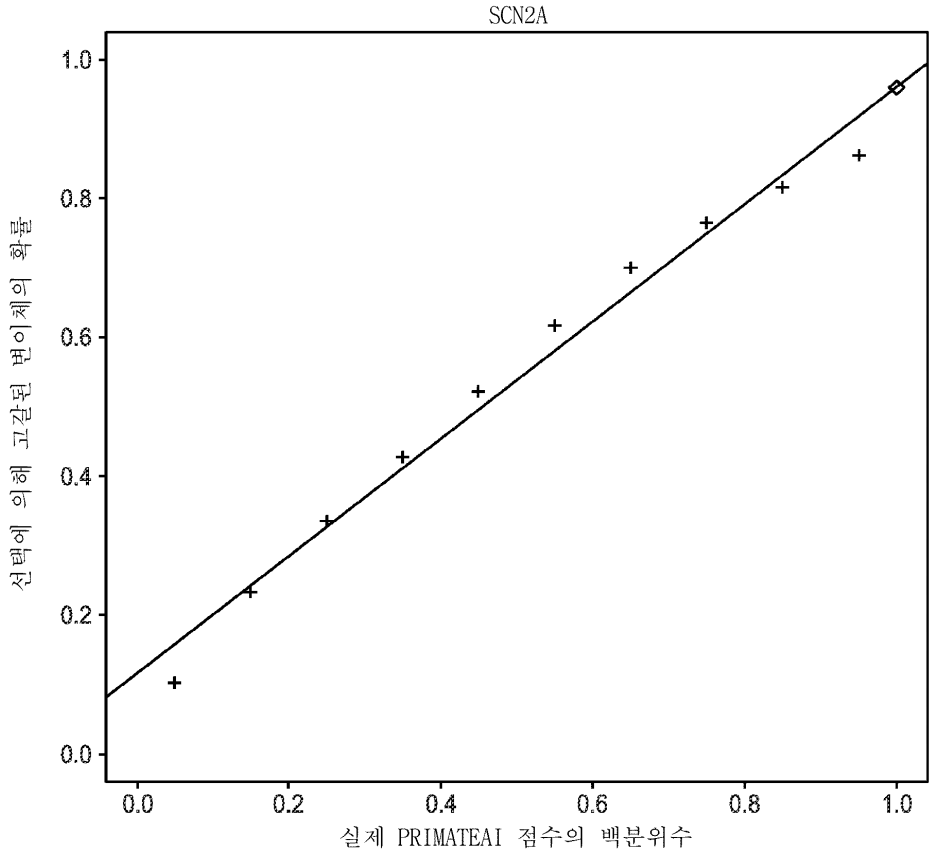
도면32


572
관측된 pAI 백분위수
실제 pAI 백분위수
컬러 키 및
히스토그램
카운트
값

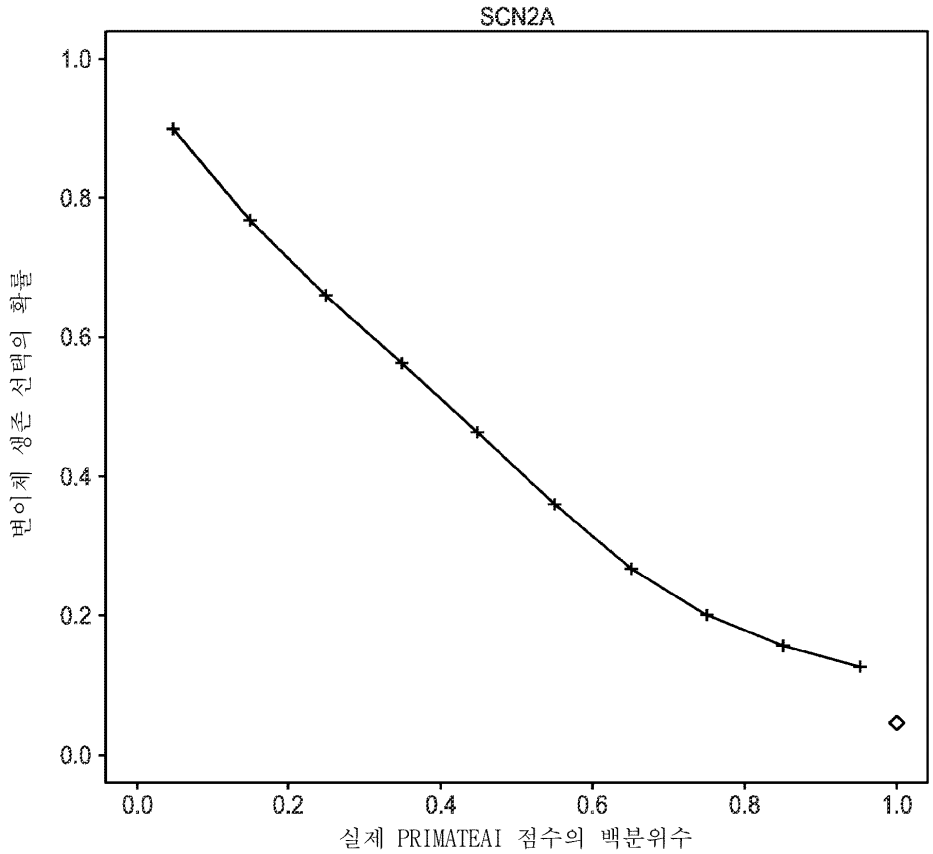
도면33


관심 변이체
200
병원성 점수 측정
202
병원성 점수
206
436
백분위수
병원성-고갈 관계
고갈 추정
500
추정된 고갈
522
생존 확률 결정
582
생존 확률
580
각 백분위수 병원성
점수에 대한 생존 확률
도출
584
보정 계수
590

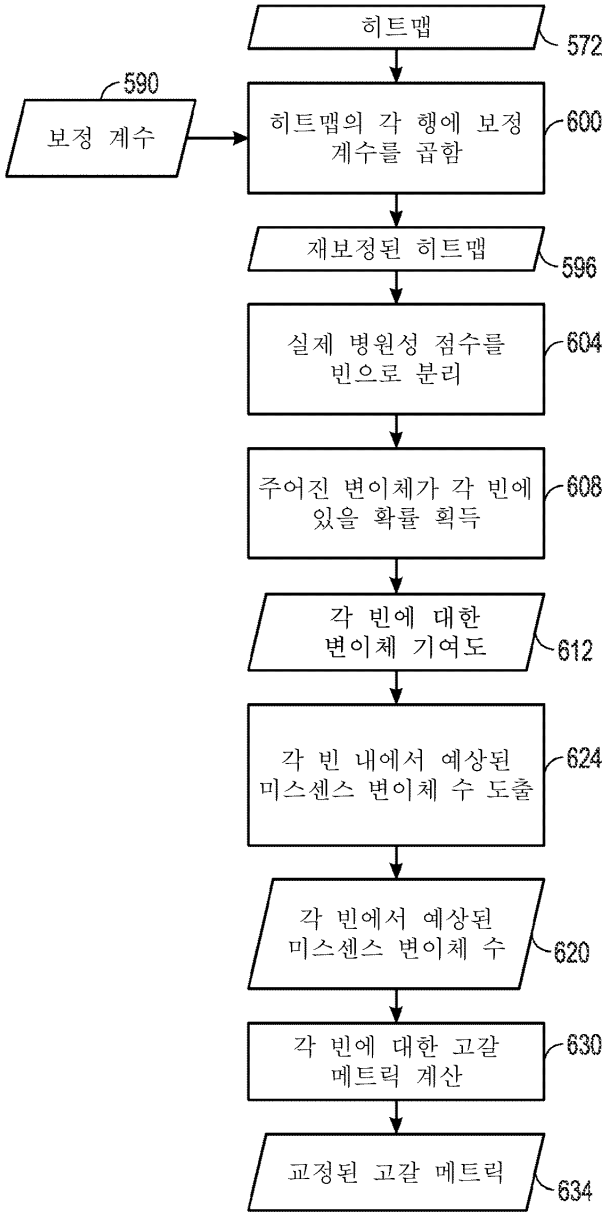
도면34


SCN2A
선택에 의해 고갈된 변이체의 확률
실제 PRIMATEAI 점수의 백분위수
0.0
0.2
0.4
0.6
0.8
1.0

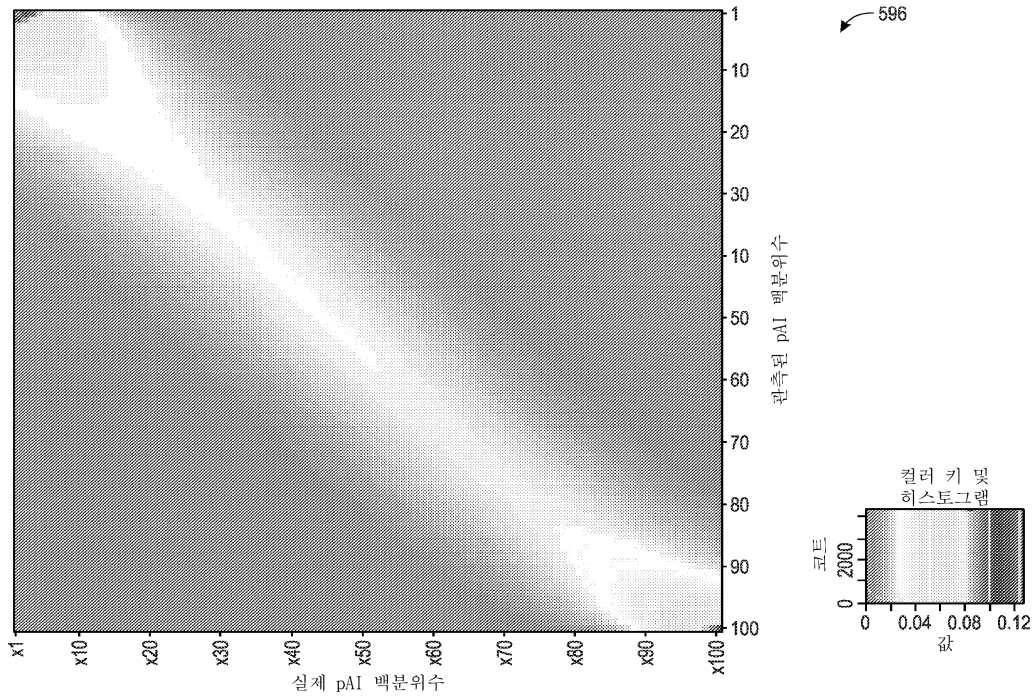
도면35


SCN2A
변이체 생존 선택의 확률
실제 PRIMATEAI 점수의 백분위수
0.0
0.2
0.4
0.6
0.8
1.0

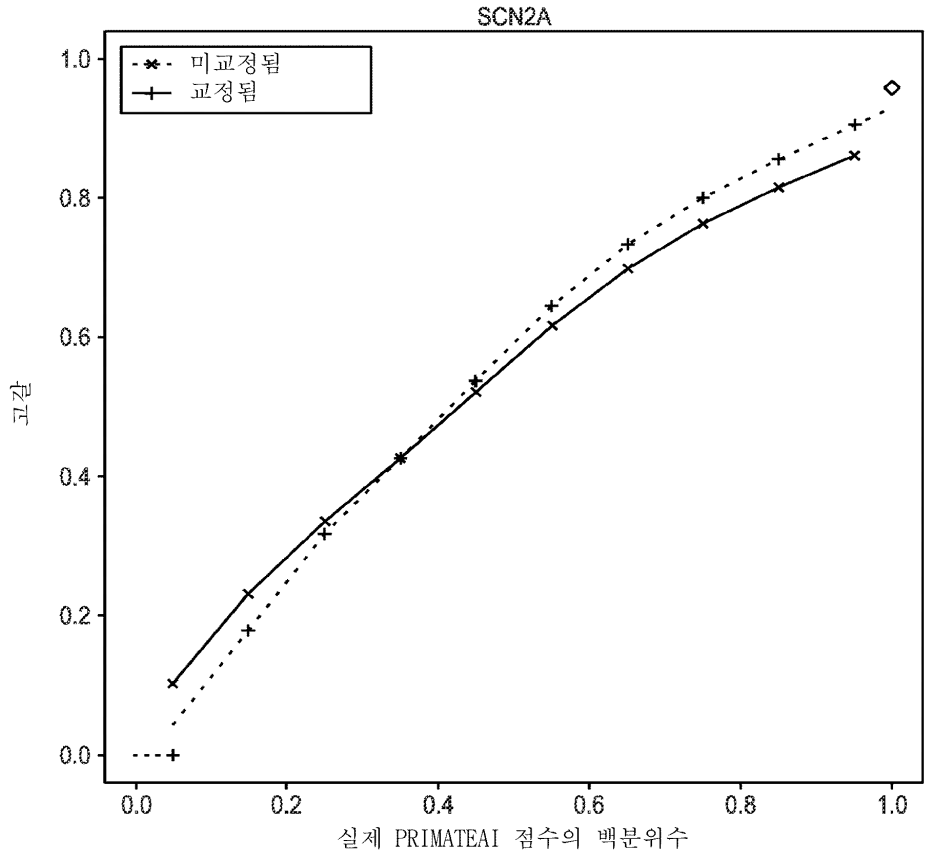
도면36


히트맵
572
590
보정 계수
히트맵의 각 행에 보정 계수를 곱함
600
재보정된 히트맵
596
실제 병원성 점수를 빈으로 분리
604
주어진 변이체가 각 빈에 있을 확률 획득
608
각 빈에 대한 변이체 기여도
612
각 빈 내에서 예상된 미스센스 변이체 수 도출
624
각 빈에서 예상된 미스센스 변이체 수
620
각 빈에 대한 고갈 메트릭 계산
630
교정된 고갈 메트릭
634

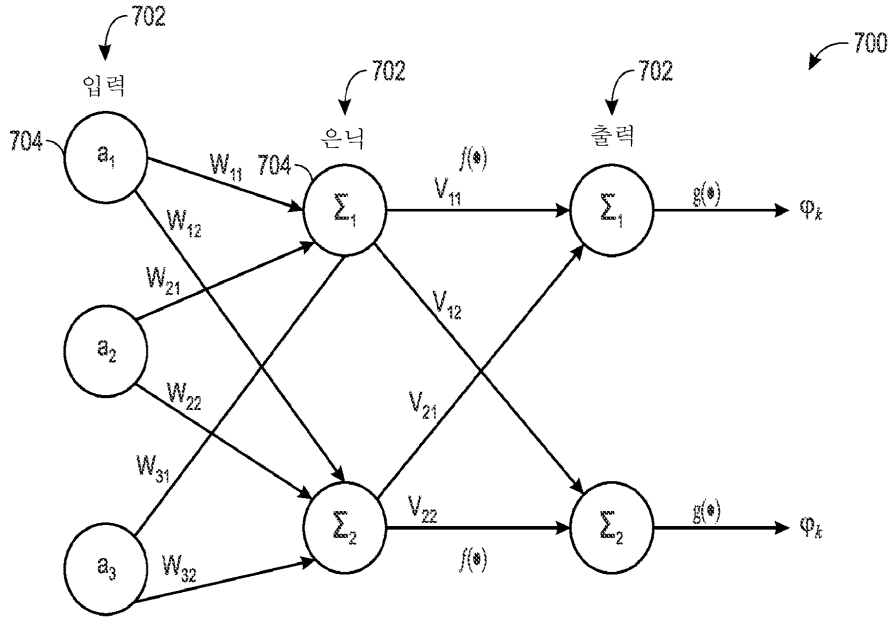
도면37


596
관측된 pAI 백분위수
실제 pAI 백분위수
컬러 키 및
히스토그램
카운트
값

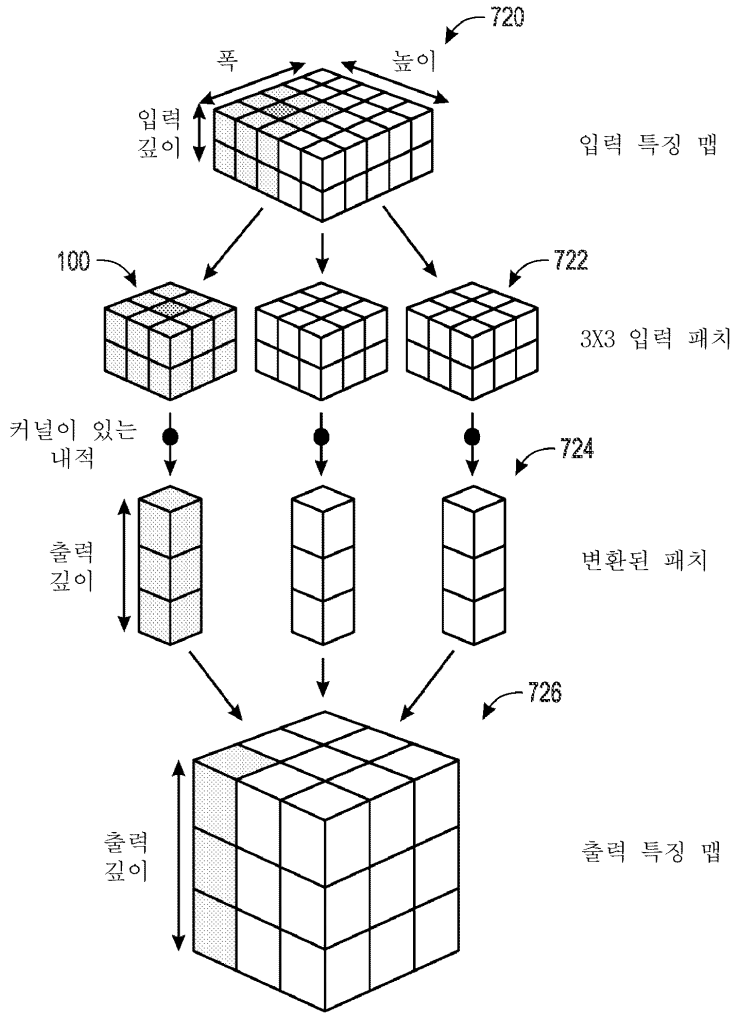
도면38


SCN2A
미교정됨
교정됨
고갈
실제 PRIMATEAI 점수의 백분위수

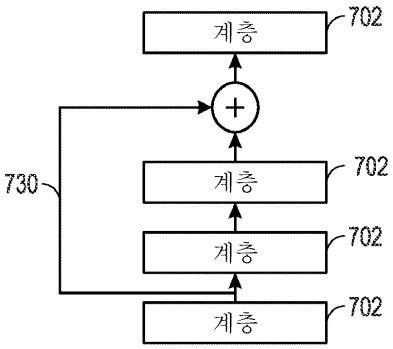
도면39


702
입력
702
은닉
702
출력
700
704
a_1
a_2
a_3
W_{11}
W_{12}
W_{21}
W_{22}
W_{31}
W_{32}
704
Σ_1
Σ_2
$f(\cdot)$
V_{11}
V_{12}
V_{21}
V_{22}
$f(\cdot)$
Σ_1
Σ_2
$g(\cdot)$
$g(\cdot)$
φ_k
φ_k

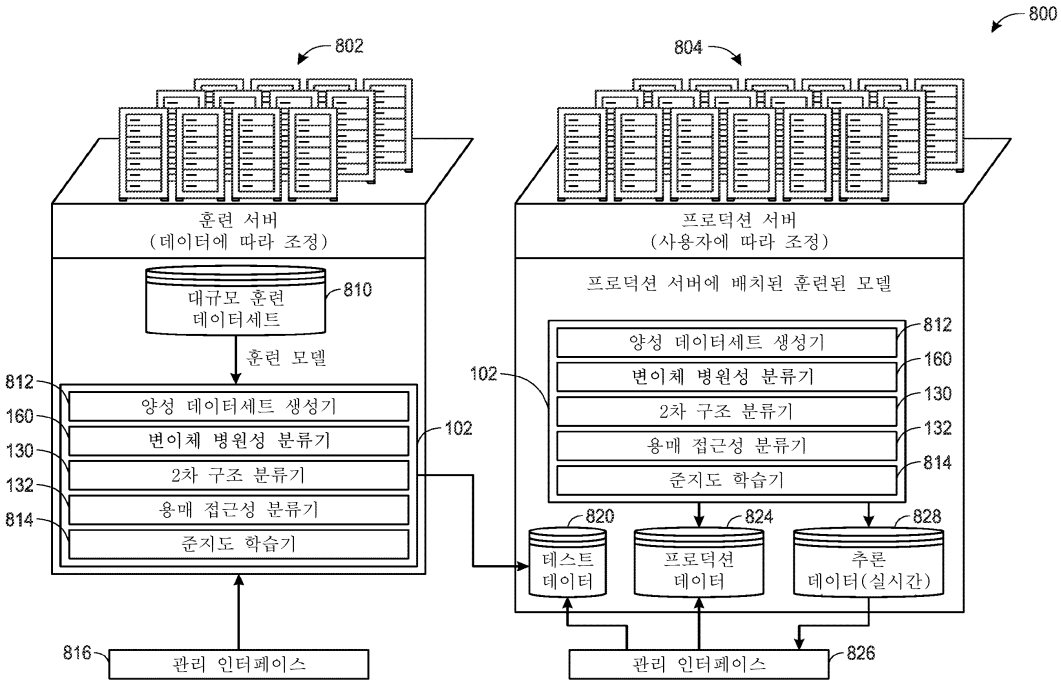
도면40


720
폭
높이
입력
깊이
입력 특징 맵
100
722
3X3 입력 패치
커널이 있는
내적
724
출력
깊이
변환된 패치
726
출력
깊이
출력 특징 맵

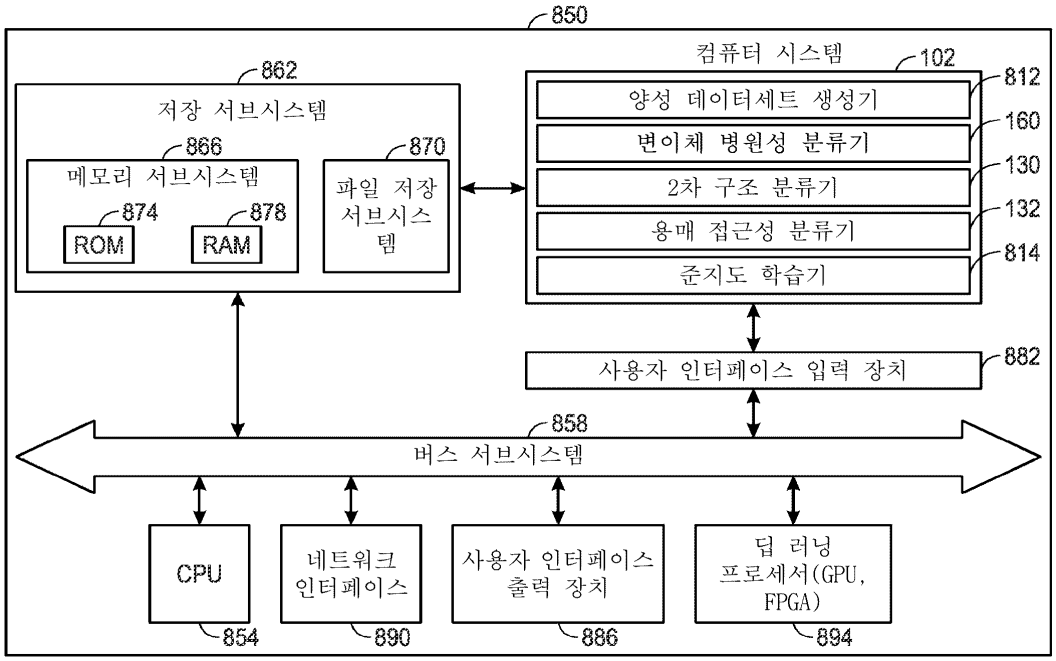
도면41


계층
702
730

도면42


800
802
804
훈련 서버
(데이터에 따라 조정)
대규모 훈련
데이터세트
810
훈련 모델
812
양성 데이터세트 생성기
160
변이체 병원성 분류기
130
2차 구조 분류기
132
용매 접근성 분류기
814
준지도 학습기
102
816
관리 인터페이스
프로덕션 서버
(사용자에 따라 조정)
프로덕션 서버에 배치된 훈련된 모델
820
테스트
데이터
824
프로덕션
데이터
828
추론
데이터(실시간)
826

도면43


850
컴퓨터 시스템
102
862
저장 서브시스템
866
메모리 서브시스템
874
ROM
878
RAM
870
파일 저장 서브시스템
양성 데이터세트 생성기
812
변이체 병원성 분류기
160
2차 구조 분류기
130
용매 접근성 분류기
132
준지도 학습기
814
사용자 인터페이스 입력 장치
882
858
버스 서브시스템
CPU
854
네트워크 인터페이스
890
사용자 인터페이스 출력 장치
886
딥 러닝 프로세서(GPU, FPGA)
894