



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0042048
(43) 공개일자 2023년03월27일

(51) 국제특허분류(Int. Cl.)
G16B 40/20 (2019.01) A61K 39/00 (2006.01)
G06N 3/04 (2023.01) G16B 15/30 (2019.01)
G16B 30/00 (2019.01) G16H 50/20 (2018.01)
(52) CPC특허분류
G16B 40/20 (2019.02)
A61K 39/0011 (2021.08)
(21) 출원번호 10-2023-7005289
(22) 출원일자(국제) 2021년07월16일
심사청구일자 없음
(85) 번역문제출일자 2023년02월14일
(86) 국제출원번호 PCT/US2021/042105
(87) 국제공개번호 WO 2022/016125
국제공개일자 2022년01월20일
(30) 우선권주장
63/053,307 2020년07월17일 미국(US)

(71) 출원인
제넨테크, 인크.
미합중국 캘리포니아 (우편번호 94080-4990) 사우
쓰샌프란시스코 디엔에이 웨이 1
(72) 발명자
리우, 카이
미국 94080-4990 캘리포니아 사우쓰샌프란시스코
디엔에이 웨이 1
로운스베리, 니콜라스, 윈스턴
미국 94080-4990 캘리포니아 사우쓰샌프란시스코
디엔에이 웨이 1
(뒷면에 계속)
(74) 대리인
장덕순, 이귀동

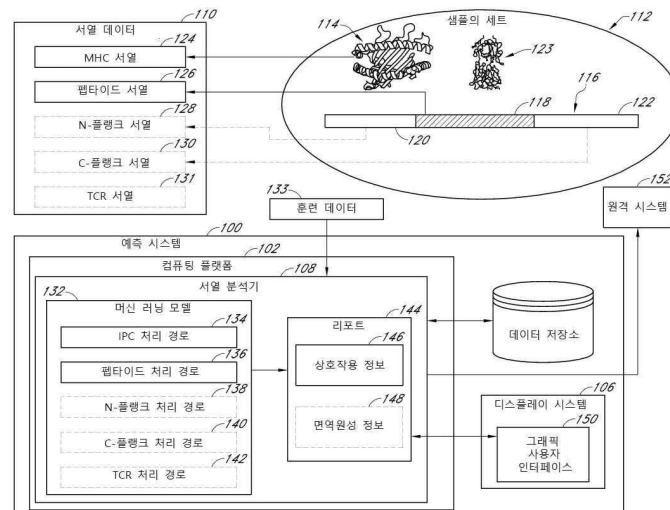
전체 청구항 수 : 총 113 항

(54) 발명의 명칭 펩타이드 결합, 제시, 및 면역원성을 예측하기 위한 어텐션-기반 신경망

(57) 요약

본 명세서에 개시된 실시예는 일반적으로 어텐션-기반 머신 러닝 모델을 이용해 대응하는 펩타이드-면역단백질 복합체(IPC) 결합체에 대한 표적 상호작용에 대한 상호작용 예측, 상호작용 친화도 예측, 또는 표적 상호작용과 관련된 면역원성 예측 중 적어도 하나를 포함하는 출력을 생성하는 것과 관련된다. 표적 상호작용은 펩타이드와 면역원성 복합체(IPC), 가령, 주 조직적합성 복합체(MHC), T 세포 수용체(TCR), 또는 둘 모두 간에 있을 수 있다. 하나 이상의 표적 상호작용이 있을 가능성이 높은 것으로 예측되는 하나 이상의 펩타이드를 포함하는 약학 조성물이 식별, 제조, 및/또는 사용될 수 있다. 이러한 약학 조성물의 투여를 포함하는 치료 방법이 정의 및/또는 사용될 수 있다.

대표도 - 도1



(52) CPC특허분류

G06N 3/04 (2023.01)
G16B 15/30 (2019.02)
G16B 30/00 (2019.02)
G16H 50/20 (2018.01)

(72) 발명자

트리프트, 윌리엄, 존

미국 94080-4990 캘리포니아 사우쓰샌프란시스코
디엔에이 웨이 1

브로드웰, 앤드릭 쿠에이드

미국 94080-4990 캘리포니아 사우쓰샌프란시스코
디엔에이 웨이 1

델라마르, 렐리아, 이본느

미국 94080-4990 캘리포니아 사우쓰샌프란시스코
디엔에이 웨이 1

준준왈라, 수치트, 수실

미국 94080-4990 캘리포니아 사우쓰샌프란시스코
디엔에이 웨이 1

명세서

청구범위

청구항 1

펩타이드 세트를 특징짓는 펩타이드 서열 세트를 액세스하는 단계이며, 펩타이드 서열 세트의 각각의 펩타이드 서열은 대상체로부터의 질병 샘플을 처리함으로써 식별된 것인 단계,

대상체의 면역단백질 복합체(IPC: immunoprotein complex)에 대해 식별된 면역단백질 복합체(IPC) 서열을 액세스하는 단계,

어텐션-기반 머신 러닝 모델의 초기 어텐션 서브시스템 내 제1 어텐션 블록을 이용해 펩타이드 서열 세트를 나타내는 펩타이드 표현 세트 및 초기 어텐션 서브시스템 내 제2 어텐션 블록을 이용해 IPC 서열을 나타내는 면역단백질 복합체(IPC) 표현을 처리하여 출력을 생성하는 단계이며, 상기 출력은 대응하는 펩타이드-IPC 결합체에 대한 상호작용 예측, 상호작용 친화도 예측, 또는 면역원성 예측 중 적어도 하나를 포함하는 것인 단계, 및

상기 출력에 기초하여 리포트를 생성하는 단계

를 포함하는 방법.

청구항 2

제1항에 있어서, 펩타이드 서열 세트의 적어도 하나의 펩타이드 서열은 대응하는 참조 서열에 대한 변이를 포함하는 변이-코딩 서열을 포함하는 것인 방법.

청구항 3

제1항 또는 제2항에 있어서,

상기 처리는

펩타이드 서열 세트의 대응하는 펩타이드 서열에 대한 펩타이드 표현 세트의 펩타이드 표현을 수신하고,

제1 어텐션 블록을 통해 펩타이드 표현을 변환 펩타이드 표현으로 변환하는 것

을 포함하고,

상기 제1 어텐션 블록은 어텐션 서브-블록 세트의 각각의 어텐션 서브-블록이 셀프-어텐션 층을 포함하는 어텐션 서브-블록 세트를 포함하는 것인 방법.

청구항 4

제1항 내지 제3항 중 어느 한 항에 있어서,

상기 처리는

IPC 표현을 수신하고,

제2 어텐션 블록을 통해 IPC 표현을 변환 IPC 표현으로 변환하는 것

을 포함하고,

상기 제2 어텐션 블록은 어텐션 서브-블록 세트의 각각의 어텐션 서브-블록이 셀프-어텐션 층을 포함하는 어텐션 서브-블록 세트를 포함하는 것인 방법.

청구항 5

제1항 내지 제4항 중 어느 한 항에 있어서,

펩타이드 표현의 적어도 일부가 펩타이드 서열 내 단량체에 대응하고 IPC 표현의 적어도 일부가 IPC 서열 내 단량체에 대응하며,

상기 처리는

제1 어텐션 블록 및 제1 가중치 세트를 이용해 펩타이드 표현에 기초하여 변환 펩타이드 표현을 생성하고,
제2 어텐션 블록 및 제2 가중치 세트를 이용해 IPC 표현에 기초하여 변환 IPC 표현을 생성하고,
변환 펩타이드 표현 및 변환 MHC 표현을 이용해 복합 표현을 생성하는 것
을 포함하는 것인 방법.

청구항 6

제1항 내지 제5항 중 어느 한 항에 있어서,

상기 펩타이드 서열 세트의 펩타이드 서열을 임베딩하여 상기 펩타이드 서열에 대한 임베딩된 펩타이드 표현을 생성하는 단계, 및

상기 펩타이드 서열에 대한 임베딩된 펩타이드 서열을 위치적으로 암호화하여 펩타이드 서열을 나타내는 펩타이드 표현 세트의 펩타이드 표현을 생성하는 단계

를 더 포함하는 방법.

청구항 7

제1항 내지 제6항 중 어느 한 항에 있어서,

제1 어텐션 블록은 어텐션 서브-블록 세트를 포함하고,

어텐션 서브-블록 세트의 각각의 어텐션 서브-블록은 적어도 하나의 셀프-어텐션 층을 포함하는 신경망을 포함하는 것인 방법.

청구항 8

제1항 내지 제7항 중 어느 한 항에 있어서,

제2 어텐션 블록은 어텐션 서브-블록 세트를 포함하고,

어텐션 서브-블록 세트의 각각의 어텐션 서브-블록은 적어도 하나의 셀프-어텐션 층을 포함하는 신경망을 포함하는 것인 방법.

청구항 9

제1항 내지 제8항 중 어느 한 항에 있어서,

제1 어텐션 블록은 제1 복수의 어텐션 서브-블록을 포함하고,

제2 어텐션 블록은 제1 복수의 어텐션 서브-블록을 포함하며,

어텐션 서브-블록의 제1 세트 및 어텐션 서브-블록의 제2 세트의 각각의 어텐션 서브-블록은 적어도 하나의 셀프-어텐션 층을 포함하는 신경망을 포함하는 것인 방법.

청구항 10

제1항 내지 제9항 중 어느 한 항에 있어서,

상기 펩타이드 표현 세트의 펩타이드 표현이 제1 어텐션 블록을 이용해 처리되는 통합 표현의 제1 부분을 형성하고,

통합 표현의 제2 부분이 N-플랭크 서열 또는 C-플랭크 서열 중 적어도 하나를 나타내는 것인 방법.

청구항 11

제1항 내지 제10항 중 어느 한 항에 있어서,

상기 펩타이드 서열 세트의 펩타이드 서열은 통합 서열의 제1 부분을 형성하고,

통합 서열의 제2 부분이 N-플랭크 서열 또는 C-플랭크 서열 중 적어도 하나를 포함하며,

어텐션-기반 머신 러닝 모델은 통합 서열을 수신 및 처리하여 상기 펩타이드 표현 세트 중 펩타이드 서열에 대응하는 펩타이드 표현을 포함하는 통합 표현을 형성하는 표현 블록을 포함하고, 상기 통합 표현은 제1 어텐션 블록에 의해 처리되는 것인 방법.

청구항 12

제1항 내지 제11항 중 어느 한 항에 있어서,

IPC 서열을 임베딩하여 IPC 서열의 임베딩된 IPC 표현을 생성하는 단계, 및

IPC 서열의 임베딩된 IPC 표현을 위치적으로 암호화하여 IPC 표현을 생성하는 단계

를 더 포함하는 방법.

청구항 13

제1항 내지 제12항 중 어느 한 항에 있어서, 어텐션-기반 머신 러닝 모델은 복수의 셀프-어텐션 층을 포함하고 복수의 셀프-어텐션 층의 각각에 대해 대응하는 다운스트림 피드포워드 신경망을 포함하는 것인 방법.

청구항 14

제1항 내지 제13항 중 어느 한 항에 있어서,

제1 어텐션 블록은 펩타이드 표현 세트의 펩타이드 표현을 수신 및 처리하여 변환 펩타이드 표현을 생성하도록 구성된 제1 신경망을 포함하고,

제2 어텐션 블록은 IPC 표현을 수신 및 처리하여 변환 IPC 표현을 생성하도록 구성된 제2 신경망을 포함하며,

제1 신경망 및 제2 신경망 각각은 적어도 하나의 셀프-어텐션 층을 포함하고,

어텐션-기반 머신-러닝 모델은 변환 펩타이드 표현 및 변환 IPC 표현을 이용해 복합 표현을 생성하도록 구성되는 것인 방법.

청구항 15

제1항 내지 제14항 중 어느 한 항에 있어서,

어텐션-기반 머신 러닝 모델은

복합 표현을 수신 및 처리하도록 구성된 신경망을 포함하는 복합 어텐션 블록

을 더 포함하고,

신경망은 셀프-어텐션 층을 포함하는 것인 방법.

청구항 16

제1항 내지 제15항 중 어느 한 항에 있어서,

어텐션-기반 머신 러닝 모델은

어텐션 서브-블록 세트를 포함하는 복합 어텐션 블록

을 더 포함하고,

어텐션 서브-블록 세트의 각각의 어텐션 서브-블록은 적어도 하나의 셀프-어텐션 층을 포함하는 신경망을 포함하는 것인 방법.

청구항 17

제1항 내지 제16항 중 어느 한 항에 있어서,

IPC는 주 조직적합성 복합체(MHC: major histocompatibility complex)를 포함하고 대응하는 펩타이드-IPC 결합

체는 상기 펩타이드 세트의 펩타이드 및 MHC를 포함하며,

대응하는 펩타이드-IPC 결합체에 대한 상호작용 친화도 예측은 펩타이드와 MHC 간 결합 친화도를 예측하고,

대응하는 펩타이드-IPC 결합체에 대한 상호작용 예측은 MHC가 세포 표면에서 펩타이드를 제시할 것인지 여부를 예측하는 것인 방법.

청구항 18

제1항 내지 제17항 중 어느 한 항에 있어서, 어텐션-기반 머신 러닝 모델은 복수의 훈련 펩타이드 서열 및 훈련 MHC 서열 세트에 대한 실험적 상호작용 친화도 데이터 또는 실험적 상호작용 데이터 중 적어도 하나를 포함하는 훈련 데이터 세트를 이용해 훈련되는 것인 방법.

청구항 19

제1항 내지 제18항 중 어느 한 항에 있어서,

IPC는 T 세포 수용체(TCR)이고 대응하는 펩타이드-IPC 쌍은 상기 펩타이드 세트의 펩타이드와 TCR 또는 TCR과 주 조직적합성 복합체(MHC)를 포함하며,

대응하는 펩타이드-IPC 결합체에 대한 면역원성 예측은 TCR에 대한 펩타이드의 면역원성을 예측하고,

어텐션-기반 머신 러닝 모델은 복수의 훈련 펩타이드 서열 및 훈련 TCR 서열 세트에 대한 실험적 면역원성 데이터를 포함하는 훈련 데이터를 이용해 훈련되는 것인 방법.

청구항 20

제1항 내지 제19항 중 어느 한 항에 있어서,

훈련 데이터 세트는 복수의 훈련 데이터 요소를 포함하며,

복수의 훈련 데이터 요소의 적어도 하나의 훈련 데이터 요소는

상기 펩타이드 세트에 포함되지 않는 훈련 펩타이드를 특징짓는 훈련 펩타이드 서열,

IPC와 상이한 훈련 IPC를 특징짓는 훈련 IPC 서열, 및

훈련 펩타이드와 훈련 IPC 간 상호작용 친화도 표시자를 식별하는 실험-기반 결과

중 적어도 하나를 포함하고,

상기 상호작용 친화도 표시자는 검정 또는 바이오센서-기반 방법을 이용해 검출된 것인 방법.

청구항 21

제1항 내지 제20항 중 어느 한 항에 있어서,

훈련 데이터 세트는 복수의 훈련 데이터 요소를 포함하며,

복수의 훈련 데이터 요소의 적어도 하나의 훈련 데이터 요소는

상기 펩타이드 세트에 포함되지 않는 훈련 펩타이드를 특징짓는 훈련 펩타이드 서열,

IPC와 상이한 훈련 MHC를 특징짓는 훈련 MHC 서열, 및

훈련 펩타이드가 훈련 MHC에 의해 세포 표면에 제시되었는지 여부를 식별하는 상호작용 표시자를 포함하는 실험-기반 결과

중 적어도 하나를 포함하고,

상기 상호작용 표시자를 결정하기 위해 면역침강법 또는 질량 분석법 중 적어도 하나가 사용된 것인 방법.

청구항 22

제1항 내지 제21항 중 어느 한 항에 있어서,

처리 단계 전에, 복수의 펩타이드-IPC 결합체에 대한 결합 친화도, 상호작용 표시자, 또는 면역원성 표시자 중

적어도 하나를 포함하는 훈련 데이터 세트를 이용해 어텐션-기반 머신 러닝 모델을 훈련하는 단계
를 더 포함하며,

상기 훈련 데이터 세트는 복수의 훈련 펩타이드 서열 및 복수의 훈련 주 조직적합성 복합체(MHC) 서열 또는 복
수의 훈련 T 세포 수용체(TCR) 서열 중 적어도 하나를 포함하는 것인 방법.

청구항 23

제1항 내지 제22항 중 어느 한 항에 있어서,

상기 처리는

제1 어텐션 블록을 이용해 펩타이드 표현 세트를 처리하고 제2 어텐션 블록을 이용해 IPC 표현을 처리하여 펩타
이드-IPC 결합체 세트에 대한 복합 표현 세트를 생성하고,

상기 복합 표현 세트를 처리하여 결과 세트를 생성하고,

펩타이드-IPC 결합체 세트 중 서브세트를 선택하는 것

을 포함하고,

선택된 상호작용 세트가 펩타이드-IPC 결합체 세트의 나머지 서브세트에 비교해서 상기 서브세트의 각각의 펩타
이드-IPC 결합체에서 발생할 가능성이 높으며,

상기 리포트는 상기 서브세트 내 각각의 펩타이드를 식별하는 것인 방법.

청구항 24

제1항 내지 제23항 중 어느 한 항에 있어서,

펩타이드 세트의 각각의 펩타이드가 사용되어 펩타이드-IPC 결합체 세트를 형성하며,

어텐션-기반 머신 러닝 모델은 펩타이드-IPC 결합체 세트의 각각의 펩타이드-IPC 결합체에 대한 번역원성 예측
을 생성하도록 구성되며,

펩타이드-IPC 결합체 세트의 펩타이드-IPC 결합체에 대한 번역원성 예측은 펩타이드-IPC 결합체 내 펩타이드의
종양-특이적 번역원성의 예측인 방법.

청구항 25

제1항 내지 제24항에 있어서, 상기 리포트는 상기 펩타이드 세트로부터 펩타이드 세트의 나머지 부분에 비해 증
가된 종양-특이적 번역원성을 갖는 펩타이드의 서브세트를 식별하는 것인 방법.

청구항 26

제1항 내지 제25항 중 어느 한 항에 있어서,

IPC는 주 조직적합성 복합체(MHC)이며,

펩타이드 세트의 각각의 펩타이드가 사용되어 펩타이드-MHC 결합체 세트를 형성하고,

어텐션-기반 머신 러닝 모델은 펩타이드-MHC 결합체 세트의 각각의 펩타이드-MHC 결합체에 대한 상호작용 예측
을 생성하도록 구성되며,

펩타이드-MHC 결합체 세트의 펩타이드-MHC 결합체에 대한 상호작용 예측은 펩타이드-MHC 결합체 내 펩타이드가
세포 표면에서 MHC에 의해 제시되는지 여부의 예측인 방법.

청구항 27

제26항에 있어서, 상기 리포트는 상기 펩타이드 세트로부터, 상기 펩타이드 세트의 나머지 부분에 비해 MHC에
의한 증가된 제시 가능성을 갖는 펩타이드의 서브세트를 식별하는 것인 방법.

청구항 28

제1항 내지 제27항 중 어느 한 항에 있어서,
 상기 펩타이드 서열 세트의 펩타이드 서열은 돌연변이 펩타이드를 특징짓는 변이-코딩 서열이고,
 상기 변이-코딩 서열은
 돌연변이 펩타이드의 N-말단에서의 서열을 식별하는 제1 부분, 및
 돌연변이 펩타이드의 항원결정기의 서열을 식별하는 제2 부분
 을 포함하고,
 상기 처리는
 초기 어텐션 서브시스템의 제1 셀프-어텐션 층을 이용해 변이-코딩 서열의 제1 부분의 제1 표현을 처리하고,
 초기 어텐션 서브시스템의 제2 셀프-어텐션 층을 이용해 변이-코딩 서열의 제2 부분의 제2 표현을 처리하는 것
 을 포함하는 것인 방법.

청구항 29

제28항에 있어서, 제1 표현 및 제2 표현은 제1 어텐션 블록 내에서 처리되는 것인 방법.

청구항 30

제1항 내지 제29항 중 어느 한 항에 있어서,
 어텐션-기반 머신 러닝 모델은 하나 이상의 변환기 인코더(transformer encoder)를 포함하고,
 상기 하나 이상의 변환기 인코더의 각각은 셀프-어텐션 층을 포함하는 것인 방법.

청구항 31

제1항 내지 제30항 중 어느 한 항에 있어서, IPC 서열 및 펩타이드 서열 세트의 각각은 아미노산 식별자의 정렬된 세트를 포함하는 것인 방법.

청구항 32

제1항 내지 제31항 중 어느 한 항에 있어서, IPC 서열은 상기 질병 샘플을 이용해 식별되는 것인 방법.

청구항 33

제1항 내지 제32항 중 어느 한 항에 있어서, IPC 서열은 상기 대상체로부터의 생물학적 샘플을 이용해 식별되는 것인 방법.

청구항 34

제1항 내지 제33항 중 어느 한 항에 있어서, 상기 질병 샘플은 암 세포를 포함하는 것인 방법.

청구항 35

제1항 내지 제34항 중 어느 한 항에 있어서,
 상기 대상체의 IPC는 주 조직적합성 복합체(MHC)를 포함하고,
 IPC 서열은 MHC 서열을 포함하며,
 IPC 표현은 MHC 표현을 포함하는 것인 방법.

청구항 36

제35항에 있어서, MHC는 MHC 클래스-I 분자를 포함하는 것인 방법.

청구항 37

제35항에 있어서, MHC는 MHC 클래스-II 분자를 포함하는 것인 방법.

청구항 38

제1항 내지 제35항 중 어느 한 항에 있어서,
대상체의 IPC는 T 세포 수용체(TCR)를 포함하고,
IPC 서열은 TCR 서열을 포함하며,
IPC 표현은 TCR 표현을 포함하는 것인 방법.

청구항 39

제1항 내지 제38항 중 어느 한 항에 있어서, 상기 질병 샘플은 조직을 포함하는 것인 방법.

청구항 40

제1항 내지 제39항 중 어느 한 항에 있어서, 상기 펩타이드 세트의 적어도 하나의 펩타이드는 신생항원인 방법.

청구항 41

제1항 내지 제40항 중 어느 한 항에 있어서, 상기 펩타이드 서열 세트의 적어도 하나의 펩타이드 서열은 상기 질병 샘플로부터 도출된 게놈 서열인 방법.

청구항 42

제1항 내지 제41항 중 어느 한 항에 있어서, 변이-코딩 서열 세트의 적어도 하나의 각각은 상기 질병 샘플의 RNA 서열에 기초하는 것인 방법.

청구항 43

제1항 내지 제42항 중 어느 한 항에 있어서,
대응하는 펩타이드-IPC 결합체는 펩타이드 세트로부터의 펩타이드 및 IPC를 포함하고,
IPC는 주 조직적합성 복합체(MHC)이며,
상호작용 친화도 예측은 펩타이드와 MHC 간 결합에 대한 결합 친화도의 예측이고,
상호작용 예측은 MHC에 의한 세포 표면에서의 펩타이드의 제시의 예측인 방법.

청구항 44

제1항 내지 제43항 중 어느 한 항에 있어서,
사용자에 의해 입력된 입력 데이터를 수신하는 단계이며, 상기 입력 데이터는 상기 대상체에 대응하는 것인 단계
를 더 포함하고,
상기 입력 데이터의 수신에 응답하여, 데이터 저장소로부터의 검색을 통해 펩타이드 서열 세트 및 IPC 서열이 액세스되며,
상기 리포트는 상기 펩타이드 세트로부터 상기 대상체의 병태를 치료하기 위한 개별화된 백신에 포함될 펩타이드의 서브세트를 식별하는 것인 방법.

청구항 45

제44항에 있어서,
개별화된 백신을 포함하는, 대상체에 대한 권장 치료제를 생성하는 단계
를 더 포함하는 방법.

청구항 46

제1항 내지 제45항 중 어느 한 항에 있어서,

사용자에 의해 입력된 입력 데이터를 수신하는 단계이며, 상기 입력 데이터는 상기 대상체에 대응하고, 상기 입력 데이터의 수신에 응답하여, 데이터 저장소로부터의 검색을 통해 펩타이드 서열 세트 및 IPC 서열이 액세스되는 것인 단계, 및

상기 리포트에 기초하여 개별화된 백신에 포함되도록 치료 펩타이드 세트를 결정하는 단계, 및
치료 펩타이드 세트를 포함하는 개별화된 백신의 제조를 용이하게 하는 동작을 개시하는 단계
를 더 포함하는 방법.

청구항 47

제46항에 있어서, 상기 동작을 개시하는 단계는

개별화된 백신의 제조에 포함되는 컴퓨터화된 프로세스를 트리거하는 알림을 생성하는 것
을 포함하는 것인 방법.

청구항 48

제1항 내지 제47항 중 어느 한 항에 있어서, 상기 처리는

어텐션-기반 머신 러닝 모델 내 임베딩 블록으로부터 복수의 요소를 포함하는 표현을 수신하고, 표현은 펩타이드 서열 세트 내 펩타이드 서열을 나타내는 펩타이드 표현 세트의 펩타이드 표현 또는 IPC 서열을 나타내는 IPC 표현이고, 다중-요소 데이터 세트의 각각의 요소가 펩타이드 서열 또는 IPC 서열 내 단량체에 대응하며,

복수의 요소의 각각의 요소에 대해, 어텐션-기반 머신 러닝 모델의 셀프-어텐션 층과 연관된 키 가중치 세트, 값 가중치 세트, 및 쿼리 가중치 세트에 각각 기반하여 키 벡터, 값 벡터, 및 쿼리 벡터를 결정하고,

복수의 요소의 변환(transformation)을 수행하여 복수의 수정 요소를 형성하고, 상기 변환은 복수의 요소에 대해 생성된 어텐션 점수 및 복수의 요소의 각각에 대해 결정된 값 벡터를 이용해 수행되며,

복수의 수정 요소에 기초하여 출력을 생성하는 것

을 포함하는 것인 방법.

청구항 49

제48항에 있어서, 상기 복수의 요소의 선택된 요소에 대해 변환을 수행하는 것은

요소의 키 벡터 및 쿼리 벡터를 이용해 상기 선택된 요소의 어텐션 점수를 결정하고, 상기 복수의 요소 중 상기 선택된 요소가 아닌 나머지 부분이 나머지 요소 세트를 형성하고,

상기 나머지 요소의 키 벡터 및 상기 선택된 요소의 쿼리 벡터를 이용해 상기 나머지 요소 세트의 각각의 나머지 요소에 대해 추가 어텐션 점수를 결정하여 추가 어텐션 점수 세트를 형성하며,

어텐션 점수, 추가 어텐션 점수 세트, 복수의 요소의 각각의 요소의 값 벡터를 이용해 수정 요소를 생성하는 것
을 포함하는 것인 방법.

청구항 50

제1항 내지 제49항 중 어느 한 항에 있어서,

디스플레이 시스템 상의 그래픽 사용자 인터페이스 상에 상기 리포트를 디스플레이하는 단계
를 더 포함하는 방법.

청구항 51

제1항 내지 제50항 중 어느 한 항에 있어서,

처리는 제1 컴퓨팅 플랫폼 상에서 수행되며,

유선 통신 링크 또는 무선 통신 링크 중 적어도 하나를 포함하는 통신 링크 세트를 통해 상기 리포트를 제2 컴퓨팅 플랫폼으로 전송하는 단계

를 더 포함하는 방법.

청구항 52

제1항 내지 제51항 중 어느 한 항에 있어서,

상기 펩타이드 세트의 적어도 하나의 펩타이드를 상기 리포트에 기초하여 면역요법을 위한 표적으로서 포함하도록 결정하는 단계

를 더 포함하는 방법.

청구항 53

제52항에 있어서, 상기 면역요법은 T 세포 요법, 개별화된 암 요법, 항원-특이적 면역요법, 항원-의존적 면역요법, 백신 및 자연 살해(NK) 세포 요법으로 구성된 군 중에서 선택되는 것인 방법.

청구항 54

제1항 내지 제53항 중 어느 한 항에 있어서,

상기 펩타이드 세트의 적어도 하나의 펩타이드를 상기 리포트에 기초하여 면역요법을 위한 표적으로서 배제하도록 결정하는 단계

를 더 포함하는 방법.

청구항 55

제54항에 있어서, 상기 면역요법은 T 세포 요법, 개별화된 암 요법, 항원-특이적 면역요법, 항원-의존적 면역요법, 백신 및 자연 살해(NK) 세포 요법으로 구성된 군 중에서 선택되는 것인 방법.

청구항 56

제1항 내지 제55항 중 어느 한 항에 있어서, IPC는 인간 백혈구 항원(HLA: human leukocyte antigen) 분자인 방법.

청구항 57

제1항 내지 제56항 중 어느 한 항에 있어서,

상기 대상체로부터의 질병 샘플을 시퀀싱하는 단계,

상기 대상체로부터의 질병 샘플의 시퀀싱에 기초하여 펩타이드 서열 세트를 정의하는 단계,

상기 리포트에 기초하여, 펩타이드 서열 세트의 서브세트를 식별하는 단계,

상기 펩타이드 세트의 상기 서브세트에 포함되는 적어도 하나의 펩타이드에 대해 코딩하는 mRNA를 합성하는 단계,

mRNA-리포플렉스 치료제를 생성하기 위해 mRNA를 지질과 복합화하는 단계, 및

mRNA-리포플렉스 치료제를 상기 대상체에게 투여하는 단계

를 더 포함하는 방법.

청구항 58

하나 이상의 펩타이드,

하나 이상의 펩타이드를 암호화하는 복수의 핵산, 또는

하나 이상의 펩타이드를 발현하는 복수의 세포

를 포함하는 백신이며,

상기 하나 이상의 펩타이드는 청구항 제1항 내지 제49항 중 어느 한 항에 따른 방법에 의해 생성된 리포트에 기초하여 펩타이드 세트 중에서 선택되며,

상기 하나 이상의 펩타이드는 상기 펩타이드 세트의 불완전한 서브세트인 백신.

청구항 59

제58항에 있어서, 복수의 핵산을 포함하는 DNA 또는 복수의 핵산을 포함하는 RNA를 포함하는 백신.

청구항 60

제58항 또는 제59항에 있어서, 복수의 핵산을 포함하는 mRNA를 포함하는 백신.

청구항 61

제58항 내지 제60항 중 어느 한 항에 있어서, 종양 백신인 백신.

청구항 62

백신을 생성하는 단계

를 포함하는, 백신을 제조하는 방법이며,

상기 백신은

하나 이상의 펩타이드,

하나 이상의 펩타이드를 암호화하는 복수의 핵산, 또는

하나 이상의 펩타이드를 발현하는 복수의 세포

를 포함하고,

상기 하나 이상의 펩타이드는 청구항 제1항 내지 제49항 중 어느 한 항에 따른 방법에 의해 생성된 리포트에 기초하여 펩타이드 세트 중에서 선택되며,

상기 하나 이상의 펩타이드는 상기 펩타이드 세트의 불완전한 서브세트인 방법.

청구항 63

제62항에 있어서, 상기 백신은 복수의 핵산을 포함하는 DNA, 복수의 핵산을 포함하는 RNA, 또는 복수의 핵산을 포함하는 mRNA를 포함하는 것인 방법.

청구항 64

제62항 또는 제63항에 있어서,

하나 이상의 펩타이드 내 아미노산에 기초하여, 하나 이상의 펩타이드를 암호화하는 복수의 핵산을 식별하는 단계

를 더 포함하고,

상기 백신은 복수의 핵산을 포함하는 것인 방법.

청구항 65

제62항 내지 제64항 중 어느 한 항에 있어서, 상기 백신은 종양 백신인 방법.

청구항 66

제65항에 있어서, 상기 하나 이상의 펩타이드의 각각의 펩타이드에 대해, 상기 종양 백신은,

각각의 펩타이드를 암호화하는 뉴클레오타이드 서열,
 각각의 펩타이드에 대응하는 아미노산 서열,
 각각의 펩타이드에 대응하는 RNA,
 각각의 펩타이드에 대응하는 DNA,
 각각의 펩타이드에 대응하는 세포,
 각각의 펩타이드에 대응하는 플라스미드, 또는
 각각의 펩타이드에 대응하는 벡터
 중 적어도 하나를 포함하는 것인 방법.

청구항 67

제62항 내지 제66항 중 어느 한 항에 있어서, 상기 백신은 부형제 또는 보조제 중 적어도 하나를 더 포함하는 것인 방법.

청구항 68

제62항 내지 제67항 중 어느 한 항에 있어서, 상기 백신은 RNA 분자를 포함하며,
 상기 RNA 분자는 5'→3' 방향으로:
 5' 캡,
 5' 비번역 지역(UTR),
 분비 신호 펩타이드를 암호화하는 폴리뉴클레오타이드 서열,
 하나 이상의 펩타이드를 암호화하는 폴리뉴클레오타이드 서열,
 주 조직적합성 복합체(MHC) 분자의 막횡단 및 세포질 도메인의 적어도 일부분을 암호화하는 폴리뉴클레오타이드 서열,
 3' UTR, 및
 폴리(A) 서열
 을 포함하고,
 3' UTR은
 AES(Amino-Terminal Enhancer of Split) mRNA 또는 이의 단편의 3' 비번역 지역, 및
 미토콘드리아로 암호화된 12S RNA의 비-코딩 RNA 또는 이의 단편
 을 포함하는 것인 방법.

청구항 69

청구항 제1항 내지 제49항 중 어느 한 항에 따른 방법에 의해 생성된 리포트에 기초하여 펩타이드 세트 중에서 선택된 하나 이상의 펩타이드
 를 포함하고,
 상기 하나 이상의 펩타이드는 상기 펩타이드 세트의 불완전한 서브세트인 약학 조성물.

청구항 70

청구항 제1항 내지 제49항 중 어느 한 항에 따른 방법에 의해 생성된 리포트에 기초하여 펩타이드 세트 중에서 선택된 하나 이상의 펩타이드를 암호화하는 핵산 서열
 을 포함하고,

상기 하나 이상의 펩타이드는 상기 펩타이드 세트의 불완전한 서브세트인 약학 조성물.

청구항 71

청구항 제1항 내지 제49항 중 어느 한 항에 따른 방법에 의해 생성된 리포트에 기초하여 식별된 면역원성 펩타이드.

청구항 72

청구항 제1항 내지 제49항 중 어느 한 항에 따른 방법에 의해 생성된 리포트에 기초하여 식별된 핵산 서열.

청구항 73

제72항에 있어서, 상기 핵산 서열은 DNA 서열을 포함하는, 핵산 서열.

청구항 74

제72항 또는 제73항에 있어서, RNA 서열을 포함하는 핵산 서열.

청구항 75

제72항 내지 제74항 중 어느 한 항에 있어서, mRNA 서열을 포함하는 핵산 서열.

청구항 76

청구항 제1항 내지 제49항 중 어느 한 항에 따른 방법에 의해 생성된 리포트에 기초하여 식별된 하나 이상의 펩타이드, 하나 이상의 약학 조성물, 또는 하나 이상의 핵산 서열 중 적어도 하나를 투여하는 단계

를 포함하는, 대상체를 치료하는 방법.

청구항 77

대상체로부터 획득된 생물학적 샘플 세트를 처리하여 펩타이드 세트를 특징짓는 펩타이드 서열 세트를 생성하는 단계,

상기 대상체로부터 획득된 생물학적 샘플 세트를 처리하여 대상체의 면역단백질 복합체(IPC)에 대해 식별된 면역단백질 복합체(IPC) 서열을 생성하는 단계,

어텐션-기반 머신 러닝 모델의 초기 어텐션 서브시스템 내 제1 어텐션 블록을 이용해 펩타이드 서열 세트를 나타내는 펩타이드 표현 세트를 생성하는 단계,

상기 초기 어텐션 서브시스템 내 제2 어텐션 블록을 이용해 IPC 서열을 나타내는 면역단백질 복합체(IPC) 표현을 생성하는 단계,

펩타이드 표현 세트 및 IPC 표현을 처리하여 출력을 생성하는 단계

를 포함하고,

상기 출력은 대응하는 펩타이드-IPC 결합체에 대한 상호작용 예측, 상호작용 친화도 예측, 면역원성 예측 중 적어도 하나를 포함하고,

대응하는 펩타이드-IPC 결합체는 상기 펩타이드 세트의 펩타이드를 포함하는 것인 방법.

청구항 78

제77항에 있어서, 상기 대상체로부터 획득된 생물학적 샘플 세트를 처리하여 펩타이드 서열 세트를 생성하는 것은

상기 대상체로부터 획득된 생물학적 샘플 세트 내 질병 샘플을 처리하여 펩타이드 서열 세트를 생성하는 것

을 포함하는 것인 방법.

청구항 79

제77항 또는 제78항에 있어서,
상기 대상체로부터 생물학적 샘플 세트를 획득하는 단계
를 더 포함하고,
상기 생물학적 샘플 세트는 질병 샘플을 포함하는 것인 방법.

청구항 80

제77항 내지 제79항 중 어느 한 항에 있어서,
상기 출력에 기초하여 리포트를 생성하는 단계
를 더 포함하는 방법.

청구항 81

사용자 장치에서, 대상체에 대한 개별화된 백신을 설계하기 위한 요청을 수신하는 단계,
상기 사용자 장치로부터, 원격 시스템으로 통신을 전송하는 단계이며,
상기 통신은 상기 대상체의 식별자를 포함하고,
상기 원격 시스템은

펩타이드 세트를 특징짓는 펩타이드 서열 세트를 액세스하며, 펩타이드 서열 세트의 각각의 펩타이드 서열은
대상체로부터의 질병 샘플을 처리함으로써 식별된 것이고,

대상체의 면역단백질 복합체(IPC)에 대해 식별된 면역단백질 복합체(IPC) 서열을 액세스하고,

어텐션-기반 머신 러닝 모델의 초기 어텐션 서브시스템 내 제1 어텐션 블록을 이용해 펩타이드 서열 세트를 나
타내는 펩타이드 표현 세트 및 초기 어텐션 서브시스템 내 제2 어텐션 블록을 이용해 IPC 서열을 나타내는 면역
단백질 복합체(IPC) 표현을 처리하여 출력을 생성하며, 상기 출력은 대응하는 펩타이드-IPC 결합체에 대한 상호
작용 예측, 상호작용 친화도 예측, 또는 면역원성 예측 중 적어도 하나를 포함하고,

상기 출력에 기초하여 리포트를 생성하고,

상기 사용자 장치로 상기 리포트를 전송하도록

구성되는 것인 단계, 및

상기 사용자 장치에서 상기 리포트를 수신하는 단계
를 포함하는 방법.

청구항 82

제81항에 있어서,

상기 대상체로부터의 질병 샘플을 수집하는 단계,

크로마토그래피 또는 질량 분석법 중 적어도 하나를 이용하여 상기 질병 샘플 내 MHC 분자로부터의 펩타이드 세
트를 포함하는 다중 펩타이드를 용리하는 단계,

펩타이드 세트를 시퀀싱하여 초기 서열 세트를 생성하는 단계,

상기 초기 서열 세트의 각각의 초기 서열을 참조 서열에 비교하는 단계, 및

비교에 기초하여 펩타이드 서열 세트를 정의하는 단계이며, 상기 펩타이드 서열 세트 내 각각의 펩타이드 서열
은 상기 참조 서열에 대한 변이를 포함하는 변이-코딩 서열인 단계

를 더 포함하는 방법.

청구항 83

대상체에 대한 치료제를 제조하기 위한 방법이며,

컴퓨팅 장치로부터 리포트를 수신하는 단계이며,

상기 컴퓨팅 장치는:

펩타이드 세트를 특징짓는 펩타이드 서열 세트를 액세스하며, 펩타이드 서열 세트의 각각의 펩타이드 서열은 대상체로부터의 질병 샘플을 처리함으로써 식별된 것이고,

대상체의 면역단백질 복합체(IPC: immunoprotein complex)에 대해 식별된 면역단백질 복합체(IPC) 서열을 액세스하고,

어텐션-기반 머신 러닝 모델의 초기 어텐션 서브시스템 내 제1 어텐션 블록을 이용해 펩타이드 서열 세트를 나타내는 펩타이드 표현 세트 및 초기 어텐션 서브시스템 내 제2 어텐션 블록을 이용해 IPC 서열을 나타내는 면역단백질 복합체(IPC) 표현을 처리하여 출력을 생성하며, 상기 출력은 대응하는 펩타이드-IPC 결합체에 대한 상호작용 예측, 상호작용 친화도 예측, 또는 면역원성 예측 중 적어도 하나를 포함하고,

상기 출력에 기초하여 리포트를 생성하도록

구성되는 것인 단계, 및

상기 리포트에 기초하여 치료제를 제조하기 위한 치료제 제조 계획을 생성하는 단계를 포함하는 방법.

청구항 84

제83항에 있어서,

상기 치료제 제조 계획에 기초하여 치료제를 제조하는 단계를 더 포함하는 방법.

청구항 85

복수의 돌연변이 펩타이드를 특징짓는 복수의 변이-코딩 서열을 어텐션-기반 머신 러닝 모델로 입력하는 단계이며, 복수의 변이-코딩 서열의 각각의 변이-코딩 서열은 대상체로부터의 질병 샘플을 처리함으로써 식별된 것인 단계,

상기 대상체의 면역단백질 복합체(IPC)에 대해 식별된 면역단백질 복합체(IPC) 서열을 어텐션-기반 머신 러닝 모델에 입력하는 단계이며,

상기 어텐션-기반 머신 러닝 모델은 어텐션-기반 머신 러닝 모델의 초기 어텐션 서브시스템 내 제1 어텐션 블록을 이용해 복수의 변이-코딩 서열을 나타내는 복수의 변이 표현 및 초기 어텐션 서브시스템 내 제2 어텐션 블록을 이용해 IPC 서열을 나타내는 면역단백질(IPC) 표현을 처리하여 출력을 생성하도록 구성되고,

상기 출력은 대응하는 돌연변이 펩타이드-IPC 결합체에 대한 상호작용 예측, 상호작용 친화도 예측, 또는 면역원성 예측 중 적어도 하나를 포함하는 것인 단계,

상기 출력에 기초하여 생성된 리포트를 수신하는 단계, 및

상기 리포트에 기초하여, 복수의 돌연변이 펩타이드 중 상기 대상체에 대한 치료제에서 사용되기 위한 서브세트를 선택하는 단계

를 포함하는 방법.

청구항 86

돌연변이 펩타이드를 특징짓는 펩타이드 서열을 수신하는 단계이며, 상기 펩타이드 서열은 대응하는 참조 서열에 대한 변이를 포함하는 것인 단계,

주 조직적합성 복합체(MHC)에 대해 식별된 MHC 서열을 수신하는 단계,

어텐션-기반 머신 러닝 모델 내 상이한 처리 경로를 이용해 펩타이드 서열 및 MHC 서열을 처리하여 출력을 생성하는 단계이며, 상기 출력은 돌연변이 펩타이드와 MHC 모두와 관련된 면역학적 활동에 대한 정보를 제공하는 것

인 단계, 및
상기 출력에 기초하여 리포트를 생성하는 단계
를 포함하는 방법.

청구항 87

제86항에 있어서, 상기 처리는

상기 어텐션-기반 머신 러닝 모델 내 펩타이드 처리 경로를 통해 펩타이드 서열을 처리하고, 상기 펩타이드 처리 경로는 제1 임베딩 블록 및 적어도 하나의 셀프-어텐션 층을 포함하는 제1 어텐션 블록을 포함하며,

상기 어텐션-기반 머신 러닝 모델 내 MHC 처리 경로를 통해 MHC 서열을 처리하고, 상기 MHC 처리 경로는 제2 임베딩 블록 및 적어도 하나의 셀프-어텐션 층을 포함하는 제2 어텐션 블록을 포함하는 것

을 포함하는 것인 방법.

청구항 88

제87항에 있어서,

T 세포 수용체(TCR)에 대해 식별된 TCR 서열을 수신하는 단계

를 더 포함하며,

상기 처리는

상기 어텐션-기반 머신 러닝 모델 내 TCR 처리 경로를 통해 TCR 서열을 처리하고, 상기 TCR 처리 경로는 제3 임베딩 블록 및 적어도 하나의 셀프-어텐션 층을 포함하는 제3 어텐션 블록을 포함하는 것

을 더 포함하는 것인 방법.

청구항 89

제86항 내지 제88항 중 어느 한 항에 있어서,

상기 면역학적 활동은 면역 반응을 포함하고,

상기 정보는 상기 면역 반응을 유발할 수 있는 돌연변이 펩타이드의 능력에 대한 예측을 포함하는 것인 방법.

청구항 90

제86항 내지 제89항 중 어느 한 항에 있어서, 상기 처리는

펩타이드 처리 경로를 통해 펩타이드 서열의 변환된 펩타이드 표현을 생성하고,

MHC 처리 경로를 통해 MHC 서열의 변환된 MHC 표현을 생성하고,

변환 펩타이드 표현 및 변환 MHC 표현을 이용해 복합 표현을 생성하고,

상기 복합 표현을 처리하여 출력을 생성하는 것

을 포함하는 것인 방법.

청구항 91

제86항 내지 제90항 중 어느 한 항에 있어서,

상기 면역학적 활동은 돌연변이 펩타이드의 MHC로의 결합을 포함하고,

상기 출력은 상기 돌연변이 펩타이드가 MHC에 결합하는지 여부에 대응하는 제1 예측 또는 결합과 연관된 친화도에 대응하는 제2 예측 중 적어도 하나를 포함하는 것인 방법.

청구항 92

제86항 내지 제91항 중 어느 한 항에 있어서,

상기 리포트에 기초하여 면역요법을 위한 표적으로서 상기 돌연변이 펩타이드를 포함하도록 결정하는 단계를 더 포함하는 방법.

청구항 93

제92항에 있어서, 상기 면역요법은 T 세포 요법, 개별화된 암 요법, 항원-특이적 면역요법, 항원-의존적 면역요법, 백신 및 자연 살해(NK) 세포 요법으로 구성된 군 중에서 선택되는 것인 방법.

청구항 94

제86항 내지 제93항 중 어느 한 항에 있어서,

상기 리포트에 기초하여 면역요법을 위한 표적으로서 상기 돌연변이 펩타이드를 배제하도록 결정하는 것 중 적어도 하나를 더 포함하는 방법.

청구항 95

제94항에 있어서, 상기 면역요법은 T 세포 요법, 개별화된 암 요법, 항원-특이적 면역요법, 항원-의존적 면역요법, 백신 및 자연 살해(NK) 세포 요법으로 구성된 군 중에서 선택되는 것인 방법.

청구항 96

제86항 내지 제95항 중 어느 한 항에 있어서,

상기 리포트에 기초하여, 상기 돌연변이 펩타이드, 상기 돌연변이 펩타이드의 전구체, 상기 돌연변이 펩타이드를 암호화하는 핵산, 또는 상기 돌연변이 펩타이드를 발현하는 복수의 세포 중 적어도 하나를 치료제에 포함하도록 결정하는 단계, 및

상기 치료제를 제조하는 단계

를 더 포함하는 방법.

청구항 97

제96항에 있어서,

상기 치료제로 대상체를 치료하는 단계

를 더 포함하는 방법.

청구항 98

제86항 내지 제97항 중 어느 한 항에 있어서,

돌연변이 펩타이드를 특징짓는 펩타이드 서열은 대상체로부터의 질병 샘플을 시퀀싱함으로써 식별되었고,

상기 펩타이드 서열은 대응하는 참조 서열에 비해 적어도 하나의 서열 변이를 가지며,

치료제는 상기 리포트에 기초하여 상기 대상체에 대해 설계된 것인 방법.

청구항 99

돌연변이 펩타이드를 특징짓는 펩타이드 서열을 수신하는 단계이며, 상기 펩타이드 서열은 대응하는 참조 서열에 대한 변이를 포함하는 것인 단계,

T 세포 수용체(TCR)에 대해 식별된 TCR 서열을 수신하는 단계,

어텐션-기반 머신 러닝 모델 내 상이한 처리 경로를 이용해 펩타이드 서열 및 TCR 서열을 처리하여 출력을 생성하는 단계이며, 상기 출력은 돌연변이 펩타이드와 TCR 모두와 관련된 면역학적 활동에 대한 정보를 제공하는 것인 단계, 및

상기 출력에 기초하여 리포트를 생성하는 단계

를 포함하는 방법.

청구항 100

제99항에 있어서, 상기 처리는

상기 어텐션-기반 머신 러닝 모델 내 펩타이드 처리 경로를 통해 펩타이드 서열을 처리하고, 상기 펩타이드 처리 경로는 제1 임베딩 블록 및 제1 어텐션 블록을 포함하고,

상기 어텐션-기반 머신 러닝 모델 내 TCR 처리 경로를 통해 TCR 서열을 처리하고, 상기 TCR 처리 경로는 제2 임베딩 블록 및 제2 어텐션 블록을 포함하는 것

을 포함하는 것인 방법.

청구항 101

제100항에 있어서,

주 조직적합성 복합체(MHC)에 대해 식별된 MHC 서열을 수신하는 단계

를 더 포함하며,

상기 처리는

상기 어텐션-기반 머신 러닝 모델 내 MHC 처리 경로를 통해 MHC 서열을 처리하고, 상기 MHC 처리 경로는 제3 임베딩 블록 및 MHC 제3 블록을 포함하는 것

을 더 포함하는 것인 방법.

청구항 102

제99항 내지 제101항 중 어느 한 항에 있어서,

상기 면역학적 활동은 면역 반응을 포함하고,

상기 정보는 상기 면역 반응을 유발할 수 있는 돌연변이 펩타이드의 능력에 대한 예측을 포함하는 것인 방법.

청구항 103

제99항 내지 제102항 중 어느 한 항에 있어서, 상기 처리는

펩타이드 처리 경로를 통해 펩타이드 서열의 변환된 펩타이드 표현을 생성하고,

TCR 처리 경로를 통해 TCR 서열의 변환된 TCR 표현을 생성하고,

변환 펩타이드 표현 및 변환 TCR 표현을 이용해 복합 표현을 생성하고,

상기 복합 표현을 처리하여 출력을 생성하는 것

을 포함하는 것인 방법.

청구항 104

제99항 내지 제103항 중 어느 한 항에 있어서,

상기 면역학적 활동은 돌연변이 펩타이드의 MHC로의 결합을 포함하고,

상기 출력은 상기 돌연변이 펩타이드가 MHC에 결합하는지 여부에 대응하는 제1 예측 또는 결합과 연관된 친화도에 대응하는 제2 예측 중 적어도 하나를 포함하는 것인 방법.

청구항 105

제99항 내지 제104항 중 어느 한 항에 있어서,

상기 리포트에 기초하여 면역요법을 위한 표적으로서 상기 돌연변이 펩타이드를 포함하도록 결정하는 단계를 더 포함하는 방법.

청구항 106

제105항에 있어서, 상기 면역요법은 T 세포 요법, 개별화된 암 요법, 항원-특이적 면역요법, 항원-의존적 면역요법, 백신 및 자연 살해(NK) 세포 요법으로 구성된 군 중에서 선택되는 것인 방법.

청구항 107

제99항 내지 제106항 중 어느 한 항에 있어서,

상기 리포트에 기초하여 면역요법을 위한 표적으로서 상기 돌연변이 펩타이드를 배제하도록 결정하는 것 중 적어도 하나를 더 포함하는 방법.

청구항 108

제107항에 있어서, 상기 면역요법은 T 세포 요법, 개별화된 암 요법, 항원-특이적 면역요법, 항원-의존적 면역요법, 백신 및 자연 살해(NK) 세포 요법으로 구성된 군 중에서 선택되는 것인 방법.

청구항 109

제99항 내지 제108항 중 어느 한 항에 있어서,

상기 리포트에 기초하여, 상기 돌연변이 펩타이드, 상기 돌연변이 펩타이드의 전구체, 상기 돌연변이 펩타이드를 암호화하는 핵산, 또는 상기 돌연변이 펩타이드를 발현하는 복수의 세포 중 적어도 하나를 치료제에 포함하도록 결정하는 단계, 및

상기 치료제를 제조하는 단계

를 더 포함하는 방법.

청구항 110

제109항에 있어서,

상기 치료제로 대상체를 치료하는 단계

를 더 포함하는 방법.

청구항 111

제99항 내지 제110항 중 어느 한 항에 있어서,

돌연변이 펩타이드를 특징짓는 펩타이드 서열은 대상체로부터의 질병 샘플을 시퀀싱함으로써 식별되었고,

상기 펩타이드 서열은 대응하는 참조 서열에 비해 적어도 하나의 서열 변이를 가지며,

치료제는 상기 리포트에 기초하여 상기 대상체에 대해 설계된 것인 방법.

청구항 112

하나 이상의 데이터 프로세서, 및

하나 이상의 데이터 프로세서 상에서 실행될 때, 상기 하나 이상의 데이터 프로세서로 하여금 청구항 제1항 내지 제49항, 제77항 내지 제81항, 제83항, 제85항 내지 제95항, 및 제99항 내지 제108항 중 어느 하나를 수행하게 하는 명령을 포함하는 비-일시적 컴퓨터 판독형 저장 매체

를 포함하는 시스템.

청구항 113

하나 이상의 데이터 프로세서로 하여금 청구항 제1항 내지 제49항, 제77항 내지 제81항, 제83항, 제85항 내지 제95항, 및 제99항 내지 제108항 중 어느 한 항을 수행하게 하도록 구성된 명령을 포함하는 비일시적 기계 판독형 저장 매체에서 유형적으로 구현되는 컴퓨터 프로그램 프로덕트.

발명의 설명

기술 분야

[0001] 발명자:

[0002] Kai Liu, Nicolas Lounsbury, William Thrift, Adric Broadwell, Lulia Delamarre,

[0003] Suchit Jhunjhunwala

관련 출원의 교차 참조

[0005] 본 출원은 2020년 07월 17일에 출원된 미국 가특허 출원 번호 63/053,307의 우선권, 발명의 명칭 "Attention-Based Neural Network to Predict Peptide Binding, Presentation, and Immunogenicity"을 주장하고, 동일자로 출원된 미국 특허 출원 번호 ____, 대리인 사건 번호 59868.23US01, 발명의 명칭 "Attention-Based Neural Network to Predict Peptide Binding, Presentation, and Immunogenicity"과 관련이 있으며, 둘 모두 그 전체가 참조로서 본 명세서에 포함된다.

기술분야

[0007] 본 개시내용은 일반적으로 (가령, 어텐션 메커니즘을 포함하는) 머신 러닝 모델을 이용하여 관심 펩타이드(가령, 돌연변이 펩타이드)가 면역단백질 복합체(IPC)와 표적 상호작용(들)(가령, MHC 분자에 결합, MHC 분자에 의해 제시, TCR에 결합 등)을 경험할지 여부, 이러한 표적 상호작용(들)과 연관된 친화도, 및/또는 면역 반응을 트리거할 수 있는 펩타이드의 능력과 관련된 예측을 생성하는 것과 관련된다. 본 개시내용은 또한 이러한 예측에 기초하여 선택된 특정 돌연변이 펩타이드(또는 연관된 전구체 또는 서열)를 포함하는 조성물 및 이를 치료에 이용하는 방법과 관련된다.

배경 기술

[0008] 신생항원 백신이 개별화된 암 치료를 제공하기 위한 비교적 새로운 방식이다. 신생항원은 종양의 체세포 돌연변이에서 파생되고 대상체의 암세포와 항원 제시 세포에 의해 제시되는 종양 특이적 항원이다.

[0009] 신생항원 백신은 하나 이상의 특정 종양 신생항원을 발현하는 암 세포를 인식하고 공격하도록 대상체의 T 세포를 프라이밍할 수 있다. 이 접근법은 종양 세포를 표적으로 삼으면서 건강한 세포를 피하는 종양 특이적 면역 반응을 생성한다. 그러나 어떤 신생항원이 대상체의 종양 세포에 의해 생성되고 대상체의 구조적 적합성 복합체(MHC) 분자에 의해 제시되는지에 대해 대상체에 따라 높은 가변성이 있다. 따라서 개별화된 백신이 개발되어 특정 대상체를 위해 사용될 수 있다. 개별화된 백신은 대상체-특이적 종양 프로파일에 기초하여 가공되거나 선택될 수 있다. 종양 프로파일은 대상체의 종양 세포에서 DNA 및/또는 RNA 서열을 결정하고 상기 서열을 이용해 종양 세포에는 존재하지만 정상 세포에는 없는 항원을 식별함으로써 정의될 수 있다.

[0010] 많은 경우에, 종양 세포에서 검출되는 대부분의 돌연변이 서열이 종양 세포 표면 상에 실제로 제시되지 않는 신생항원에 대응한다. 이러한 신생항원은 개별화된 백신에 대한 좋지 않은 후보가 될 것이다. 예를 들어, 검출된 펩타이드 서열은 세포내 생산되지만 MHC-I 또는 MHC-II 분자와 결합하지 못하거나 (세포 표면에서) 제시되지 않을 돌연변이 펩타이드 내 아미노산을 식별할 수 있습니다. 대안으로, MHC-I 또는 MHC-II 분자에 의해 제시될 수 있는 돌연변이 펩타이드는 세포내 생성되지 않을 수 있다. 어느 경우라도, 돌연변이 펩타이드는 예를 들어 MHC-I 분자의 경우 CD8+ 세포독성 T 림프구에 의해, 또는 MHC-II 분자의 경우 CD4+ 보조 T-세포에 의해 면역학적 반응을 트리거하지 못할 것이다.

[0011] 따라서, 돌연변이 펩타이드 서열을 검출하거나 어떤 돌연변이 펩타이드 서열에 대해 단일 생물학적 상호작용이 일어날지(예를 들어, 펩타이드가 분자에 결합할지 여부)를 예측하는 것에만 초점을 맞추는 백신에 대한 신생항원 후보를 식별하기 위한 서열 분석이 많은 위 양성을 생성할 수 있다. 이러한 유형의 서열 분석은 면역학적 반응을 프라이밍하도록 의도된 개별화된 백신을 개발하는 데 비효율적일 것이다.

[0012] 따라서, 어떤 신생항원이 주어진 대상체의 종양 세포에 의해 제시되는지 및/또는 신생항원을 포함하는 백신이 강한 면역학적 반응을 트리거할 것인지를 예측하는 것이 바람직할 수 있다.

발명의 내용

해결하려는 과제

과제의 해결 수단

- [0013] 하나 이상의 실시예에서, 방법이 제공된다. 상기 방법은 펩타이드의 세트를 특징짓는 펩타이드 서열의 세트를 액세스하는 단계를 포함하고, 펩타이드 서열의 세트의 각각의 펩타이드 서열은 대상체로부터의 질병 샘플을 처리함으로써 식별되었다. 상기 방법은 대상체의 면역단백질 복합체(IPC: immunoprotein complex)에 대해 식별된 면역단백질 복합체(IPC) 서열을 액세스하는 단계를 포함한다. 상기 방법은 어텐션-기반 머신 러닝 모델의 초기 어텐션 서브시스템 내 제1 어텐션 블록을 이용해 펩타이드 서열의 세트를 나타내는 펩타이드 표현의 세트 및 초기 어텐션 서브시스템 내 제2 어텐션 블록을 이용해 IPC 서열을 나타내는 면역단백질 복합체(IPC) 표현을 처리하여 출력을 생성하는 단계를 포함하며, 상기 출력은 대응하는 펩타이드-IPC 결합체에 대한 상호작용 예측, 상호작용 친화도 예측, 또는 면역원성 예측 중 적어도 하나를 포함한다. 상기 방법은 상기 출력에 기초하여 리포트를 생성하는 단계를 포함한다.
- [0014] 하나 이상의 실시예에서, 백신은 하나 이상의 펩타이드, 하나 이상의 펩타이드를 암호화하는 복수의 핵산, 또는 하나 이상의 펩타이드를 발현하는 복수의 세포를 포함한다. 하나 이상의 펩타이드는 본 명세서에 개시된 하나 이상의 방법 중 일부 또는 전부에 의해 생성되는 리포트에 기초하여 펩타이드의 세트 중에서 선택된다. 상기 하나 이상의 펩타이드는 상기 펩타이드의 세트의 불완전한 서브세트이다.
- [0015] 하나 이상의 실시예에서, 백신을 제조하기 위한 방법이 제공된다. 상기 방법은 백신을 생성하는 단계를 포함하며, 상기 백신은 하나 이상의 펩타이드, 하나 이상의 펩타이드를 암호화하는 복수의 핵산, 또는 하나 이상의 펩타이드를 발현하는 복수의 세포를 포함한다. 하나 이상의 펩타이드는 본 명세서에 개시된 하나 이상의 방법 중 일부 또는 전부에 의해 생성되는 리포트에 기초하여 펩타이드의 세트 중에서 선택된다. 상기 하나 이상의 펩타이드는 상기 펩타이드의 세트의 불완전한 서브세트이다.
- [0016] 하나 이상의 실시예에서, 본 명세서에 개시된 하나 이상의 방법 중 일부 또는 전부에 의해 생성되는 리포트에 기초하여 펩타이드의 세트 중에서 선택된 하나 이상의 펩타이드를 포함하는 약학 조성물이 제공된다. 상기 하나 이상의 펩타이드는 상기 펩타이드의 세트의 불완전한 서브세트이다.
- [0017] 하나 이상의 실시예에서, 본 명세서에 개시된 하나 이상의 방법 중 일부 또는 전부에 의해 생성되는 리포트에 기초하여 펩타이드의 세트 중에서 선택된 하나 이상의 펩타이드를 암호화하는 핵산 서열을 포함하는 약학 조성물이 제공된다. 상기 하나 이상의 펩타이드는 상기 펩타이드의 세트의 불완전한 서브세트이다.
- [0018] 하나 이상의 실시예에서, 본 명세서에 개시된 하나 이상의 방법 중 일부 또는 전부에 의해 생성되는 리포트에 기초하여 식별된 면역원성 펩타이드가 제공된다.
- [0019] 하나 이상의 실시예에서, 본 명세서에 개시된 하나 이상의 방법 중 일부 또는 전부에 의해 생성되는 리포트에 기초하여 식별된 핵산 서열이 제공된다.
- [0020] 하나 이상의 실시예에서, 대상체를 치료하는 방법이 제공된다. 상기 방법은 본 명세서에 개시된 하나 이상의 방법 중 일부 또는 전부에 의해 생성된 리포트에 기초하여 식별된 하나 이상의 펩타이드, 하나 이상의 약학 조성물, 또는 하나 이상의 핵산 서열 중 적어도 하나를 투여하는 단계를 포함한다.
- [0021] 하나 이상의 실시예에서, 대상체로부터 획득된 생물학적 샘플의 세트를 처리하여 펩타이드의 세트를 특징짓는 펩타이드 서열의 세트를 생성하는 단계를 포함하는 방법이 제공된다. 상기 방법은 상기 대상체로부터 획득된 생물학적 샘플의 세트를 처리하여 대상체의 면역단백질 복합체(IPC)에 대해 식별된 면역단백질 복합체(IPC) 서열을 생성하는 단계를 포함한다. 상기 방법은 어텐션-기반 머신 러닝 모델의 초기 어텐션 서브시스템 내 제1 어텐션 블록을 이용해 펩타이드 서열의 세트를 나타내는 펩타이드 표현의 세트를 생성하는 단계를 포함한다. 상기 방법은 상기 초기 어텐션 서브시스템 내 제2 어텐션 블록을 이용해 IPC 서열을 나타내는 면역단백질 복합체(IPC) 표현을 생성하는 단계를 포함한다. 상기 방법은 펩타이드 표현의 세트 및 IPC 표현을 처리하여 출력을 생성하는 단계를 포함하고, 상기 출력은 대응하는 펩타이드-IPC 결합체에 대한 상호작용 예측, 상호작용 친화도 예측, 면역원성 예측 중 적어도 하나를 포함하고, 대응하는 펩타이드-IPC 결합체는 상기 펩타이드의 세트의 펩타이드를 포함한다.
- [0022] 하나 이상의 실시예에서, 방법이 제공된다. 상기 방법은 사용자 장치에서, 대상체에 대한 개별화된 백신을 설계

하기 위한 요청을 수신하는 단계를 포함한다. 상기 방법은 상기 사용자 장치로부터, 원격 시스템으로 통신을 전송하는 단계를 포함하고, 상기 통신은 상기 대상체의 식별자를 포함한다. 원격 시스템은, 펩타이드의 세트를 특징짓는 펩타이드 서열의 세트를 액세스하고 — 상기 펩타이드 서열의 세트의 각각의 펩타이드 서열은 대상체로부터의 질병 샘플을 처리함으로써 식별되었음 —, 대상체의 면역단백질 복합체(IPC)에 대해 식별된 면역단백질 복합체(IPC) 서열을 액세스하며, 어텐션-기반 머신 러닝 모델의 초기 어텐션 서브시스템 내 제1 어텐션 블록을 이용해 펩타이드 서열의 세트를 나타내는 펩타이드 표현의 세트 및 초기 어텐션 서브시스템 내 제2 어텐션 블록을 이용해 IPC 서열을 나타내는 면역단백질 복합체(IPC) 표현을 처리하여 출력을 생성하도록 구성된다. 상기 출력은 상호작용 예측, 상호작용 친화도 예측, 또는 대응하는 펩타이드-IPC 결합체에 대한 면역원성 예측 중 적어도 하나를 포함한다. 상기 원격 시스템은 상기 출력에 기초하여 리포트를 생성하고, 상기 리포트를 사용자 장치로 전송하도록 구성된다.

[0023] 상기 방법은 사용자 장치에서 상기 리포트를 수신하는 단계를 포함한다.

[0024] 하나 이상의 실시예에서, 대상체를 위한 치료제를 제조하기 위한 방법이 제공된다. 상기 방법은 컴퓨팅 장치로부터 리포트를 수신하는 단계를 포함한다. 컴퓨팅 장치는, 펩타이드의 세트를 특징짓는 펩타이드 서열의 세트를 액세스하고 — 상기 펩타이드 서열의 세트의 각각의 펩타이드 서열은 대상체로부터의 질병 샘플을 처리함으로써 식별되었음 —, 대상체의 면역단백질 복합체(IPC)에 대해 식별된 면역단백질 복합체(IPC) 서열을 액세스하며, 어텐션-기반 머신 러닝 모델의 초기 어텐션 서브시스템 내 제1 어텐션 블록을 이용해 펩타이드 서열의 세트를 나타내는 펩타이드 표현의 세트 및 초기 어텐션 서브시스템 내 제2 어텐션 블록을 이용해 IPC 서열을 나타내는 면역단백질 복합체(IPC) 표현을 처리하여 출력을 생성하도록 구성된다. 상기 출력은 상호작용 예측, 상호작용 친화도 예측, 또는 대응하는 펩타이드-IPC 결합체에 대한 면역원성 예측 중 적어도 하나를 포함한다. 컴퓨팅 장치는 상기 출력에 기초하여 리포트를 생성하도록 구성된다. 상기 방법은 상기 리포트에 기초하여 치료제를 제조하기 위한 치료제 제조 계획을 생성하는 단계를 더 포함한다.

[0025] 하나 이상의 실시예에서, 복수의 돌연변이 펩타이드를 특징짓는 복수의 변이-코딩 서열을 어텐션-기반 머신 러닝 모델로 입력하는 단계를 포함하는 방법이 제공되며, 복수의 변이-코딩 서열의 각각의 변이-코딩 서열은 대상체로부터의 질병 샘플을 처리함으로써 식별되었다. 상기 방법은 상기 대상체의 면역단백질 복합체(IPC)에 대해 식별된 면역단백질 복합체(IPC) 서열을 어텐션-기반 머신 러닝 모델에 입력하는 단계를 포함한다. 상기 어텐션-기반 머신 러닝 모델은 어텐션-기반 머신 러닝 모델의 초기 어텐션 서브시스템 내 제1 어텐션 블록을 이용해 복수의 변이-코딩 서열을 나타내는 복수의 변이 표현을 처리하고 초기 어텐션 서브시스템 내 제2 어텐션 블록을 이용해 IPC 서열을 나타내는 면역단백질(IPC) 표현을 처리하여 출력을 생성하도록 구성된다. 상기 출력은 상호작용 예측, 상호작용 친화도 예측, 또는 대응하는 돌연변이 펩타이드-IPC 결합체에 대한 면역원성 예측 중 적어도 하나를 포함한다. 상기 방법은 상기 출력에 기초하여 생성된 리포트를 수신하는 단계, 및 상기 리포트에 기초하여, 복수의 돌연변이 펩타이드 중 상기 대상체에 대한 치료제에서 사용되기 위한 서브세트를 선택하는 단계를 포함한다.

[0026] 하나 이상의 실시예에서, 돌연변이 펩타이드를 특징짓는 펩타이드 서열을 수신하는 단계 — 펩타이드 서열은 대응하는 참조 서열에 대한 변이를 포함함 —, 주 조직적합성 복합체(MHC)에 대해 식별된 MHC 서열을 수신하는 단계, 어텐션-기반 머신 러닝 모델 내 상이한 처리 경로를 이용해 펩타이드 서열 및 MHC 서열을 처리하여, 출력을 생성하는 단계 — 출력은 돌연변이 펩타이드 및 MHC 모두와 관련된 면역학적 활동에 대한 정보를 제공함 —, 및 출력에 기초하여 리포트를 생성하는 단계를 포함하는 방법이 제공된다.

[0027] 하나 이상의 실시예에서, 돌연변이 펩타이드를 특징짓는 펩타이드 서열을 수신하는 단계 — 펩타이드 서열은 대응하는 참조 서열에 대한 변이를 포함함 —, T 세포 수용체(TCR)에 대해 식별된 TCR 서열을 수신하는 단계, 어텐션-기반 머신 러닝 모델 내 상이한 처리 경로를 이용해 펩타이드 서열 및 TCR 서열을 처리하여, 출력을 생성하는 단계 — 출력은 돌연변이 펩타이드 및 TCR 모두와 관련된 면역학적 활동에 대한 정보를 제공함 —, 및 출력에 기초하여 리포트를 생성하는 단계를 포함하는 방법이 제공된다.

[0028] 일부 실시예에서, 하나 이상의 데이터 프로세서 및 하나 이상의 데이터 프로세서 상에서 실행될 때, 상기 하나 이상의 데이터 프로세서로 하여금 여기에 개시된 하나 이상의 방법 중 일부 또는 전부를 수행하게 하는 명령을 포함하는 비일시적 컴퓨터 판독형 저장 매체를 포함하는 시스템이 제공된다.

[0029] 일부 실시예에서, 하나 이상의 데이터 프로세서로 하여금 여기에 개시된 하나 이상의 방법 중 일부 또는 전부를 수행하게 하는 명령을 포함하는 비일시적 기계 판독형 저장 매체로 유형으로(tangibly) 구현되는 컴퓨터 프로그램 프로덕트가 제공된다.

[0030] 본 개시내용의 일부 실시예는 하나 이상의 데이터 프로세서를 포함하는 시스템을 포함한다. 일부 실시예에서, 시스템은 하나 이상의 데이터 프로세서 상에서 실행될 때 하나 이상의 데이터 프로세서로 하여금 여기서 개시된 하나 이상의 방법 중 일부 또는 전부 및/또는 하나 이상의 프로세스 중 일부 또는 전부를 수행하게 하는 명령을 포함하는 비밀시적 컴퓨터 판독형 저장 매체를 포함한다. 본 개시내용의 일부 실시예는 하나 이상의 데이터 프로세서 상에서 실행될 때 하나 이상의 데이터 프로세서로 하여금 여기서 개시된 하나 이상의 방법 중 일부 또는 전부 및/또는 여기서 개시된 하나 이상의 프로세스의 일부 또는 전부를 수행하게 하는 명령을 포함하는 비밀시적 기계 판독형 저장 매체로 유형으로 구현되는 컴퓨터-프로그램 프로덕트를 포함한다.

[0031] 채용된 용어 및 표현은 한정이 아닌 설명 측면에서 사용되며, 이러한 용어 및 표현의 사용에 도시되고 기재된 특징부 또는 이의 일부분의 임의의 균등물을 배제하는 어떠한 의도도 없고, 청구된 발명의 범위 내에서 다양한 수정이 가능함이 자명하다. 따라서, 본 발명이 실시예 및 선택적 특징으로 특정하게 개시되었지만 본 명세서에 개시된 개념의 수정, 및 변형이 해당 분야의 통상의 기술자에 의해 이용될 수 있고, 이러한 수정 및 변형이 첨부된 청구항에 의해 정의된 본 발명의 범위 내에 있는 것으로 간주됨이 이해될 것이다.

도면의 간단한 설명

[0032] 본 개시내용은 첨부된 도면과 함께 설명된다:

도 1은 다양한 실시예에 따르는 예측 시스템의 블록도이다.

도 2는 하나 이상의 실시예에 따르는 머신 러닝 모델을 이용해 예측을 생성하기 위한 프로세스의 흐름도이다.

도 3은 하나 이상의 실시예에 따르는 도 1로부터의 머신 러닝 모델에 대한 하나의 구성의 개략도이다.

도 4a는 하나 이상의 실시예에 따르는 머신 러닝 모델(400)의 개략도이다.

도 4b는 하나 이상의 실시예에 따르는 머신 러닝 모델(400)에 대한 상이한 구성의 개략도이다.

도 4c는 하나 이상의 실시예에 따르는 머신 러닝 모델(400)에 대한 상이한 구성의 개략도이다.

도 5는 하나 이상의 실시예에 따르는 어텐션 블록(500)의 개략도이다.

도 6은 하나 이상의 실시예에 따르는 예시적 셀프-어텐션 층을 이용한 서열 표현을 처리하기 위한 프로세스의 흐름도이다.

도 7은 하나 이상의 실시예에 따르는 도 6에 기재된 프로세스(600)를 예시하는 개략도이다.

도 8은 다양한 펩타이드의 면역학적 활동에 대한 정보를 생성하기 위한 프로세스의 흐름도이다.

도 9는 다양한 펩타이드의 면역학적 활동에 대한 정보를 생성하기 위한 프로세스의 흐름도이다.

도 10은 하나 이상의 실시예에 따르는 펩타이드 및 MHC와 관련된 예측을 생성하도록 머신 러닝 모델을 훈련하고 훈련된 머신 러닝 모델을 이용하기 위한 프로세스의 흐름도이다.

도 11은 하나 이상의 실시예에 따라 훈련 데이터의 표를 포함하는 예시이다.

도 12는 하나 이상의 실시예에 따르는 신생항원 후보 및 대응하는 가능한 신생항원결정기 후보의 예시이다.

도 13은 하나 이상의 실시예에 따르는 펩타이드 및 TCR와 관련된 예측을 생성하도록 머신 러닝 모델을 훈련하고 훈련된 머신 러닝 모델을 이용하기 위한 프로세스의 흐름도이다.

도 14a, 14b 및 14c는 하나 이상의 실시예에 따른 예시적 정밀도-재현율(RR) 곡선을 갖는 플롯이다.

도 15는 하나 이상의 실시예에 따른 테스트 데이터 세트에서 각각의 대립유전자에 대한 모델 A 및 P-MHC-I 모델의 용리-리간드 출력의 예시적인 평균 정밀도 값을 비교하는 플롯이다.

도 16a 및 16b는 하나 이상의 실시예에 따르는 쥐 데이터세트에 대한 P-MHC-I 모델의 성능에 인간 데이터세트에 대한 P-MHC-I 모델의 성능을 비교하는 플롯이다.

도 17a 및 17b는 하나 이상의 실시예에 따르는 제시 데이터에 대한 Model C에 P-MHC-II 모델의 성능을 비교하는 플롯이다.

도 18a 및 18b는 하나 이상의 실시예에 따르는 홀드아웃 데이터세트에 대한 Model C에 P-MHC-II 모델의 성능을

비교하는 플롯이다.

도 19는 하나 이상의 실시예에 따르는 테스트 데이터세트에 대한 모델 C에 P-MHC-II 모델에 대한 평균 정밀도의 유전자형별 비교를 보여주는 플롯이다.

도 20은 하나 이상의 실시예에 따르는 CD 8 멀티머 어세이 데이터(제1 테스트 면역원성 데이터세트)에 대한 P-MHC-I 모델(EL 출력), 모델 A(EL 출력) 및 모델 B(BA 출력)의 성능을 예시하는 수신기 작동 특성(ROC) 곡선의 플롯이다.

도 21a-d은 하나 이상의 실시예에 따르는 ELISpot 어세이(제1 테스트 면역원성 데이터세트)에 대한 P-MHC-I 모델(EL 출력), 모델 A(EL 출력) 및 모델 B(BA 출력)의 성능을 예시하는 플롯이다.

도 22a-d은 하나 이상의 실시예에 따르는 모델 A(BA 출력), 모델 A(EL 출력), 모델 C(BA 출력) 및 P-MHC-I 모델(EL 출력)의 성능을 예시하는 플롯이다.

도 23은 하나 이상의 실시예에 따르는 TESLA 멀티머 어세이 데이터를 이용한 모델 A(EL 출력) 및 모델 B(BA 출력), 및 P-MHC-I 모델(EL 출력)에 대한 ROC 곡선을 비교하는 플롯의 예시이다.

첨부된 도면에서, 유사한 구성요소 및/또는 특징은 동일한 참조 라벨을 가질 수 있다. 또한, 동일한 유형의 다양한 구성요소는 참조 라벨 뒤의 대시와 유사한 구성요소를 구별하는 두 번째 라벨로 구분할 수 있다. 명세서에서 첫 번째 참조 라벨만 사용되는 경우 두 번째 참조 라벨과 관계없이 동일한 첫 번째 참조 라벨을 가진 유사한 구성요소 중 하나에 설명이 적용된다.

발명을 실시하기 위한 구체적인 내용

I. 개요

개별화된 백신에 대한 후보로서 어떤 돌연변이 펩타이드(예를 들어, 신생항원)를 선택할지 예측할 수 있는 것의 중요성을 인식하여, 본 명세서에 기재된 실시예는 이러한 예측을 현재 이용 가능한 다양한 방법 및 시스템보다 더 정확하게 하기 위한 방법 및 시스템을 제공한다. 본 명세서에 기재된 실시예는 머신 러닝 방법 및 시스템을 사용하여, 비제한적 예를 들어 돌연변이 펩타이드를 특징짓는 서열을 분석하여 백신 후보로서 이러한 돌연변이 펩타이드의 성공 가능성을 결정할 때 생성되는 위양성의 수를 줄임으로써 예측 성능을 향상시킨다.

예를 들어, 본 명세서에 기재된 실시예는 머신 러닝 모델 및/또는 머신 러닝 모델에 의해 생성된 출력을 사용하여 대상체로부터의 질병 샘플로부터 식별된 서열을 분석하는 다양한 방법을 제공한다. 질병 샘플에서 검출된 돌연변이 펩타이드가 주 조직적합성 복합체(MHC: major histocompatibility complex) 분자(가령, MHC-I, MHC-II)와 상호 작용하는지 여부를 예측하거나, 돌연변이 펩타이드가 MHC 분자와 상호 작용하는 정도를 예측하거나, 둘 모두를 예측하기 위해, 머신 러닝 모델은 초기에 MHC 분자에 대응하는 MHC 서열의 표현의 처리와 별도로 돌연변이 펩타이드를 특징짓는 서열의 표현을 처리한다. 돌연변이 펩타이드를 특징짓는 서열은 변이-코딩 서열(variant-coding sequence)로 지칭될 수 있다. MHC 서열은 MHC 분자의 전체 서열의 적어도 일부분(예를 들어, 전체 서열, 펩타이드-결합 포켓과 상호작용하는 부분인 MHC 분자의 의사서열(pseudosequence), 의사서열을 포함하는 다른 일부 부분, 등)으로 구성될 수 있다.

머신 러닝 모델은 다양한 처리 서브시스템을 포함한다. 머신 러닝 모델은 예를 들어 표현 서브시스템, 표현 어텐션 서브시스템, 복합 서브시스템, 복합 어텐션 서브시스템 및 출력 서브시스템을 포함할 수 있다. 각각의 "서브시스템"은 하나 이상의 블록으로 구성될 수 있으며, 각각의 블록은 하나 이상의 서브블록 및/또는 층으로 구성된다. 서브블록은 임의의 수의 층(또는 유닛)로 구성될 수 있다.

표현 서브시스템은 펩타이드 서열(변이-코딩 서열을 포함할 수 있음)의 펩타이드 표현 및 MHC 서열의 MHC 표현을 생성하기 위해 사용될 수 있다. 표현 어텐션 서브시스템은 MHC 서열의 표현과 독립적으로 또는 별도로(예를 들어, 병렬로) 펩타이드 서열의 표현을 처리하는 데 사용된다. 이들 두 개의 병렬 처리 경로는 유사하거나 상이하게 구성될 수 있지만 각각은 적어도 하나의 어텐션 메커니즘을 포함한다. 이들 병렬 처리 경로를 통해 펩타이드 서열 및 MHC 서열의 표현을 처리하면 머신 러닝 모델의 예측 성능이 향상된다.

또한, 본 명세서에 기재된 실시예는 일련의 생물학적 이벤트에 대응하는 모델을 훈련시키는 것이 단일 생물학적 이벤트에 대응하는 모델을 훈련시키는 것보다 훨씬 더 많은 데이터를 요구할 수 있음을 인식하고 고려한다. 서열 분석을 위한 모델을 훈련시키는 것은 잠재적으로 관찰 가능한 서열의 수가 많기 때문에 특히 복잡할 수 있다. 수백만 개의 잠재적 신생항원이 있을 뿐만 아니라, 예를 들어 MHC 클래스-I 분자에 대한 단백질을 암호화

하는 유전자도 매우 다형적이다: 클래스-I 인간 MHC에는 거의 20,000개의 대립유전자가 있다. 따라서, 본 명세서에 기재된 실시예는 훈련의 복잡도를 줄이고 훈련 성능을 향상시키는 머신 러닝 모델을 훈련하기 위한 방법 및 시스템을 제공한다. 예를 들어, 훈련에 사용되는 변이-코딩 서열은 임계 아미노산 길이(가령, 14개의 아미노산) 이하의 아미노산 길이를 갖는 변이-코딩 서열을 사용하여 훈련이 수행되도록 선택 및/또는 트리밍될 수 있다. 임계 아미노산 길이 이하의 길이를 갖는 변이-코딩 서열을 포함하는 훈련 데이터셋을 생성하면 훈련의 전반적인 복잡성을 줄일 수 있을 뿐만 아니라 훈련 및/또는 예측 성능을 향상시킬 수도 있다(가령, 에포크당 성능 메트릭의 변동 감소에 의한 예측 성능을 향상).

[0039] 따라서, 본 명세서에 개시된 기술은 돌연변이 펩타이드와 같은 펩타이드와 관련된 면역학적 활동에 관한 예측을 생성하기 위한 머신 러닝 기반 접근법을 포함한다. 하나 이상의 예측을 포함하는 출력을 생성하는 머신 러닝 모델이 제공된다. 예를 들어 출력은 하나 이상의 상호작용 예측, 하나 이상의 상호작용 친화도 예측, 하나 이상의 면역원성 예측 또는 이들의 조합을 생성할 수 있다. 상호작용 예측은 펩타이드(예를 들어, 주어진 변이-코딩 서열에 의해 식별되는 아미노산의 주어진 정렬된 세트를 포함하는 돌연변이 펩타이드)가 하나 이상의 표적 상호작용을 경험하는지 여부에 관한 예측을 포함할 수 있다. 표적 상호작용은 예를 들어, IPC에 결합(가령, MHC 분자, TCR), 세포 표면에서 MHC 분자에 의해 제시되는 것, 또는 또 다른 유형의 표적 상호작용일 수 있다. 상호작용 친화도 예측은 하나 이상의 표적 상호작용에 대한 친화도의 예측을 포함할 수 있다. 예를 들어, 상호작용 친화도 예측은 펩타이드-MHC 결합에 대한 결합 친화도를 나타낼 수 있다. 상호작용(가령, 결합) 친화도는 상호작용(가령, 결합)의 경향, 강도 및/또는 안정성에 기초하여 결정될 수 있다.

[0040] 또한, 출력은 펩타이드의 면역원성을 포함하거나 나타낼 수 있다. 예를 들어, 출력은 펩타이드가 특정 대상체 또는 대상체 그룹에서 면역 반응을 트리거할지 여부를 예측할 수 있다. 이들 예측은 다수의 돌연변이 펩타이드 각각에 대해 생성될 수 있고, 예측은 백신에 포함 및/또는 치료에 사용될 하나 이상의 돌연변이 펩타이드를 선택하는 데 사용될 수 있다. 비제한적 예를 들면, 높은 예상 결합 친화도, 종양 세포 표면에 제시될 높은 확률 및/또는 높은 예측 면역원성과 연관된 돌연변이 펩타이드는 백신에 포함되거나 치료에서 사용되기 위해 선택될 수 있다.

[0041] 본 명세서에 기재된 실시예는 펩타이드 및 면역단백질 복합체(IPC)와 관련된 면역학적 활동에 대한 예측을 생성하기 위해 어텐션-기반 머신 러닝 모델을 사용하기 위한 방법 및 시스템을 제공한다. IPC는 MHC 또는 TCR일 수 있다. 펩타이드의 세트를 특징짓는 펩타이드 서열의 세트가 액세스될 수 있고, 펩타이드 서열의 세트의 각각의 펩타이드 서열은 대상체로부터의 질병 샘플을 처리함으로써 식별되었다. 대상체의 면역단백질 복합체(IPC)에 대해 면역단백질 복합체(IPC) 서열이 식별될 수 있다. 어텐션-기반 머신 러닝 모델의 초기 어텐션 서브시스템 내 제1 어텐션 블록을 이용해 펩타이드 서열의 세트를 나타내는 펩타이드 표현의 세트가 처리되고 초기 어텐션 서브시스템 내 제2 어텐션 블록을 이용해 IPC 서열을 나타내는 면역단백질 복합체(IPC) 표현이 처리되어 출력을 생성할 수 있다. 상기 출력은 상호작용 예측, 상호작용 친화도 예측, 또는 대응하는 펩타이드-IPC 결합체에 대한 면역원성 예측 중 적어도 하나를 포함한다. 리포트가 상기 출력에 기반하여 생성된다.

[0042] 아래의 설명은 이러한 방법 및 시스템의 예시적인 구현 및 생성된 리포트가 치료를 계획, 설계 및/또는 제조하는 데 사용될 수 있는 방식을 제공한다.

[0043] II. 어텐션-기반 머신 러닝 모델링을 사용한 돌연변이 펩타이드와 관련된 면역학적 활동과 관련된 예측

[0044] II.A. 개요

[0045] 이제 도면을 참조하면, 도 1은 다양한 실시예에 따른 예측 시스템(100)의 블록도이다. 예측 시스템(100)은 펩타이드, 특히, 돌연변이 펩타이드의 면역학적 활동에 관한 예측을 생성하는 데 사용된다. 예측 시스템(100)은 컴퓨팅 플랫폼(102), 데이터 저장소(104) 및 디스플레이 시스템(106)을 포함한다. 컴퓨팅 플랫폼(102)은 다양한 형태를 취할 수 있다. 하나 이상의 실시예에서, 컴퓨팅 플랫폼(102)은 단일 컴퓨터(또는 컴퓨터 시스템) 또는 서로 통신하는 다수의 컴퓨터를 포함한다. 다른 예에서, 컴퓨팅 플랫폼(102)은 클라우드 컴퓨팅 플랫폼의 형태를 취한다.

[0046] 데이터 저장소(104) 및 디스플레이 시스템(106)은 각각 컴퓨팅 플랫폼(102)과 통신한다. 일부 예에서, 데이터 저장소(104), 디스플레이 시스템(106) 또는 둘 모두는 컴퓨팅 플랫폼(102)의 일부로 간주되거나 이와 통합될 수 있다. 따라서, 일부 예에서, 컴퓨팅 플랫폼(102), 데이터 저장소(104), 및 디스플레이 시스템(106)은 서로 통신하는 개별적인 구성요소일 수 있지만, 다른 예에서는, 이들 구성요소의 일부 조합이 함께 통합될 수 있다. 상이한 구성요소들 간의 통신은 임의의 개수의 유선 통신 링크, 무선 통신 링크, 광학 통신 링크, 또는 이들의 조합을

사용하여 구현될 수 있다.

- [0047] 예측 시스템(100)은 하드웨어, 소프트웨어, 펌웨어 또는 이들의 조합을 사용하여 구현될 수 있는 서열 분석기(108)를 포함한다. 하나 이상의 실시예에서, 서열 분석기(108)는 컴퓨팅 플랫폼(102)에서 구현된다. 서열 분석기(108)는 처리를 위해 서열 데이터(110)를 수신한다. 예를 들어, 서열 데이터(110)는 서열 분석기(108)에 입력으로 전송되거나, 데이터 저장소(104) 또는 일부 다른 유형의 저장소(가령, 클라우드 저장소)로부터 불러와지거나, 그 밖의 다른 일부 방식으로 획득될 수 있다. 일부 경우에, 서열 데이터(110)는 입력 장치를 통해 사용자에게 의해 입력된 사용자 입력을 수신하는 것에 응답하여 데이터 저장소(104)로부터 불러와질 수 있다.
- [0048] 서열 데이터(110)는 샘플 세트(112)의 처리로부터 생성될 수 있다. 샘플 세트(112)는 하나 이상의 대상체로부터의 하나 이상의 생물학적 샘플의 형태를 취할 수 있다(예를 들어, 질병 샘플, 건강한 샘플, 이들의 조합). 샘플 세트(112)는 대상체의 종양으로부터 얻은 샘플을 포함할 수 있다. 종양은 예를 들어, 폐암, 흑색종, 유방암, 난소암, 전립선암 신장암, 위암, 결장암, 고환암, 두경부암, 췌장암, 뇌암, B-세포 림프종, 급성 골수성 백혈병, 만성 골수성 백혈병, 만성 림프구성 백혈병, T 세포 림프구성 백혈병, 비소세포 폐암, 소세포 폐암, 또는 이들의 조합의 정후일 수 있다.
- [0049] 샘플 세트(112)의 샘플은 예를 들어 다양한 면역단백질 복합체(IPC) 분자 및 다양한 펩타이드, 또는 이들의 조합을 포함할 수 있다. 샘플 세트(112)가 질병 샘플을 포함할 때, 펩타이드는 하나 이상의 돌연변이 펩타이드(가령, 신생항원)를 포함할 수 있다. IPC 분자는 예를 들어 다양한 MHC 분자, 다양한 TCR 분자 또는 이들의 조합을 포함할 수 있다.
- [0050] 하나 이상의 실시예에서, 샘플 세트(112)는 면역단백질 복합체(IPC)(114)(예를 들어, MHC 클래스 I 분자, MHC 클래스 II 분자, TCR 등) 및 아미노산 사슬(116)을 포함한다. 아미노산 사슬(116)은 펩타이드(118), N-플랭크(120) 및 C-플랭크(122)를 포함하는 아미노산 사슬일 수 있다. 펩타이드(118)는 펩타이드(118)와 N-플랭크(120) 사이의 N-말단을 포함하거나 제외하는 것으로 정의될 수 있고 펩타이드(118)와 C-플랭크(122) 사이의 C-말단을 포함하거나 제외하는 것으로 정의될 수 있다. 펩타이드(118)는 대응하는 참조 서열과 비교할 때 펩타이드(118)가 하나 이상의 변이(예를 들어, 하나 이상의 서열 변이)를 포함할 때 돌연변이 펩타이드로 간주된다. 일부 실시예에서, 샘플 세트(112)는 또한 면역단백질 복합체(123)(예를 들어, MHC 클래스 I 분자, MHC 클래스 II 분자, TCR 등)를 포함한다.
- [0051] 샘플 세트(112)는 서열 데이터(110)를 생성하기 위해 처리될 수 있다. 일부 실시예에서, 샘플 세트(112) 내의 다수의 샘플은 상이한 시간에 처리될 수 있다. 일부 실시예에서, 예측 시스템(110)은 서열 데이터(110)를 생성하기 위해 샘플 세트(112)를 처리하는 데 사용되는 샘플 분석기를 포함한다. 서열 데이터(110)는 예를 들어 적어도 하나의 면역단백질 복합체(IPC) 서열(124)(가령, 면역단백질 복합체(114)에 대응하는 하나의 IPC 서열(124)) 및 적어도 하나의 펩타이드 서열(126)(가령, 펩타이드(118)에 대응하는 하나의 펩타이드 서열(126))을 포함한다. 서열 데이터(110)는 또한 적어도 하나의 N-플랭크 서열(128)(예를 들어, N-플랭크(120)에 대응하는 하나의 N-플랭크 서열(128)), 적어도 하나의 C-플랭크 서열(130)(예를 들어, C-플랭크(122)에 대응하는 하나의 C-플랭크 서열(130)), 또는 각자의 펩타이드 서열(126)에 대응하는 둘 모두를 포함할 수 있다.
- [0052] 면역단백질 복합체(114)가 MHC의 형태를 취할 때, IPC 서열(124)은 예를 들어 MHC의 적어도 일부분을 특징짓는 MHC 서열일 수 있다. 면역단백질 복합체(114)가 TCR의 형태를 취할 때, IPC 서열(124)은 예를 들어 TCR의 적어도 일부분을 특징짓는 TCR 서열일 수 있다. 또 다른 실시예에서, IPC 서열(124)은 TCR 서열과 TCR 분자의 적어도 일부분 및 TCR 분자에 펩타이드를 제시할 수 있는 MHC 분자의 적어도 일부분을 각각 특징짓는 MHC 서열 모두를 포함할 수 있다. 일부 실시예에서, 서열 데이터(110)는 샘플 세트(112)에서 MHC 형태의 면역단백질 복합체(114)의 적어도 일부분을 특징짓는 MHC 서열의 형태로 IPC 서열(124) 뿐만 아니라 TCR의 적어도 일부분(예를 들어, 면역단백질 복합체(123))을 특징짓는 별개의 TCR 서열(131)까지 포함할 수 있다.
- [0053] 펩타이드 서열(126)은 펩타이드(118)의 적어도 일부분을 특징짓는다. N-플랭크 서열(128)은 N-플랭크(120)의 적어도 일부분을 특징짓는다. 예를 들어, N-말단으로부터 업스트림인 아미노산(또는 아미노산 잔기)의 수가 클 수 있기 때문에, N-플랭크(120)에 대한 대응하는 서열은 트리밍되어 N-플랭크 서열(128)을 생성할 수 있다. C-플랭크 서열(130)은 C-플랭크(122)의 적어도 일부를 특징짓는다. C-말단으로부터 다운스트림인 아미노산(또는 아미노산 잔기)의 수가 클 때, C-플랭크(122)에 대한 대응하는 서열이 트리밍되어 C-플랭크 서열(130)을 생성할 수 있다.
- [0054] 서열 분석기(108)는 처리를 위한 입력으로서 서열 데이터(110)를 수신한다. 서열 분석기(108)는 서열 데이터

(110)를 처리하는 머신 러닝 모델(132)을 포함한다. 일부 실시예에서, 서열 분석기(108)는 처리를 위해 머신 러닝 모델(132)로 직접 전송된다. 또 다른 실시예에서, 서열 분석기(108)는 처리를 위해 서열 데이터(110)를 머신 러닝 모델(132)로 전송하기 전에 서열 데이터(110)를 전처리한다.

[0055] 머신 러닝 모델(132)은 다수의 상이한 방식 중 임의의 방식으로 구현될 수 있다. 하나 이상의 실시예에서, 머신 러닝 모델(132)은 어텐션-기반 머신 러닝 모델의 형태를 취한다. 머신 러닝 모델(132)은 훈련 모드 또는 예측 모드에서 사용될 수 있다. 훈련 모드에서, 머신 러닝 모델(132)은 훈련 데이터세트(133)를 사용하여 훈련된다. 훈련 데이터세트를 형성할 수 있는 데이터의 예는 이하의 섹션 II.E에서 더 설명된다. 머신 러닝 모델(132)은 예측 모드에서 사용될 수 있도록 훈련된다.

[0056] 머신 러닝 모델(132)은 IPC 처리 경로(134)를 통해 IPC 서열(124)을 처리하고 펩타이드 처리 경로(136)를 통해 펩타이드 서열(126)을 처리한다. IPC 및 펩타이드에 대한 이들 두 경로의 분리는 머신 러닝 모델(132)의 개선된 예측 성능을 가능하게 한다. 일부 실시예에서, 머신 러닝 모델(132)은 N-플랭크 처리 경로(138)를 통한 N-플랭크 서열(128), C-플랭크 처리 경로(140)를 통한 C-플랭크 서열(130), 또는 둘 모두를 추가로 처리한다.

[0057] IPC 처리 경로(134)는 하나 이상의 상이한 경로로 구성될 수 있다. 예를 들어, 어떤 경우에, IPC 처리 경로(134)는 예를 들어 MHC 서열 형태의 IPC 서열(124)을 처리하기 위한 MHC 처리 경로의 형태를 취한다. 또 다른 경우에, IPC 처리 경로(134)는 예를 들어 TCR 서열 형태의 IPC 서열(124)을 처리하기 위한 TCR 처리 경로를 포함한다. 또 다른 경우에, IPC 처리 경로(134)는 MHC 서열과 TCR 서열 모두를 포함하는 IPC 서열(124)을 처리하기 위한 처리 경로를 포함한다. 일부 실시예에서, IPC 처리 경로(134)가 MHC 처리 경로의 형태를 취할 때, 머신 러닝 모델(132)은 예를 들어 TCR 서열(131)을 처리하기 위한 TCR 처리 경로(142)를 또한 포함한다. 이들 상이한 처리 경로에 대한 구현예가 이하에서 더 상세히 기재된다.

[0058] 머신 러닝 모델(132)은 리포트(144)를 생성하는 데 사용되는 출력을 생성하기 위해 서열 데이터(110)를 처리한다. 리포트(144)는 머신 러닝 모델(132)의 정확한 출력을 포함할 수 있고, 출력의 변환 또는 필터링된 버전을 포함할 수 있으며, 또는 둘 모두를 포함할 수 있다. 일부 경우에, 서열 분석기(108)는 머신 러닝 모델(132)의 출력에 기초하여, 통지, 추천, 경고 또는 그 밖의 다른 정보를 생성할 수 있으며, 이러한 추가 정보는 리포트(144)에 포함된다.

[0059] 리포트(144)는 예를 들어 하나 이상의 펩타이드(예를 들어, 하나 이상의 돌연변이 펩타이드)에 대한 관심 면역학적 활동에 대한 정보를 포함하는 출력일 수 있다. 예를 들어, 리포트(144)는 펩타이드(118) 및 면역단백질 복합체(114)(예를 들어, MHC), 펩타이드 및 면역단백질 복합체(123)(예를 들어, TCR) 또는 둘 다에 관련된 면역활성에 대한 정보를 포함할 수 있다. 리포트(144)는 예를 들어 상호작용 정보(146), 면역원성 정보(148) 또는 둘 모두를 포함할 수 있다. 상호작용 정보(146)는 펩타이드(118)와 면역단백질 복합체(114) 사이, 펩타이드(118)와 면역단백질 복합체(123) 사이, 또는 둘 다 사이의 선택된 상호작용 세트에 대한 예측을 제공할 수 있다. 면역원성 정보(148)는 펩타이드(118)의 면역원성에 대한 예측을 제공할 수 있다.

[0060] 하나 이상의 실시예에서, 리포트(144)는 디스플레이 시스템(106) 상의 그래픽 사용자 인터페이스(150) 상에 디스플레이될 수 있다. 사용자는 리포트(144)를 조회 및/또는 그래픽 사용자 인터페이스(150)를 통해 리포트(144)와 상호작용할 수 있고 리포트(144)를 사용하여 샘플 세트(112) 중 적어도 하나가 획득(수집)된 대상체의 치료에 대한 결정을 내릴 수 있다.

[0061] 일부 실시예에서, 예측 시스템(100)은 리포트(144)를 원격 시스템(152)으로 (예를 들어, 무선으로) 전송한다. 원격 시스템(152)은 클라우드 컴퓨팅 플랫폼, 클라우드 저장소, 또 다른 컴퓨터 시스템, 사용자 장치(예를 들어, 스마트폰, 태블릿, 랩톱 등) 또는 일부 다른 유형의 플랫폼일 수 있다. 일부 실시예에서, 원격 시스템(152)은 치료제 제조 시스템(또는 기계) 또는 그 일부일 수 있다.

[0062] 도 2는 하나 이상의 실시예에 따르는 머신 러닝 모델을 이용해 예측을 생성하기 위한 프로세스의 흐름도이다. 프로세스(200)는 도 1에 기재된 예측 시스템(100)을 사용하여 구현될 수 있다. 예를 들어, 프로세스(200)는 도 1의 서열 분석기(108) 및 머신 러닝 모델(132)을 사용하여 구현될 수 있다.

[0063] 프로세스(200)는 예를 들어 단계(202)를 포함할 수 있다. 단계(202)는 훈련 펩타이드 서열 데이터, 훈련 면역단백질 복합체(IPC) 데이터, 및 훈련 면역학적 활동 데이터를 포함하는 훈련 데이터 세트를 사용하여 어텐션-기반 머신 러닝 모델을 훈련하는 것을 포함한다.

[0064] 단계(204)는 펩타이드의 세트를 특징 짓는 펩타이드 서열의 세트를 액세스하는 단계를 포함하며, 펩타이드 서열

의 세트의 각각의 펩타이드 서열은 대상체로부터의 질병 샘플을 처리함으로써 식별되었다.

- [0065] 단계(206)는 대상체의 면역단백질 복합체(IPC: immunoprotein complex)에 대해 식별된 면역단백질 복합체(IPC) 서열을 액세스하는 단계를 포함한다.
- [0066] 단계(208)는 어텐션-기반 머신 러닝 모델의 초기 어텐션 서브시스템 내 제1 어텐션 블록을 이용해 펩타이드 서열의 세트를 나타내는 펩타이드 표현의 세트 및 초기 어텐션 서브시스템 내 제2 어텐션 블록을 이용해 IPC 서열을 나타내는 면역단백질 복합체(IPC) 표현을 처리하여 출력을 생성하는 단계를 포함하며, 상기 출력은 대응하는 펩타이드-IPC 결합체에 대한 상호작용 예측, 상호작용 친화도 예측, 또는 면역원성 예측 중 적어도 하나를 포함한다. 제1 어텐션 블록은 제2 어텐션 블록에 독립적이다.
- [0067] 단계(210)는 상기 출력에 기초하여 리포트를 생성하는 단계를 포함한다. 리포트는 치료제의 설계 및/또는 제조 및/또는 치료 계획을 촉진시키는 데 사용될 수 있다. 예를 들어, 리포트는 펩타이드의 세트 중 펩타이드의 서브세트를 식별하거나 대상체에 대한 치료제를 생성하는 데 사용되기 위해 펩타이드의 서브세트에 대해 어떤 것을 선택할 것인지에 대한 표시자를 제공할 수 있다. 예를 들어, 치료제는 펩타이드의 서브세트, 펩타이드의 서브세트의 각각에 대한 전구체, 또는 일부 다른 형태일 수 있다.
- [0068] II.B. 머신 러닝 모델의 예시적 아키텍처
- [0069] II.B.1. 일반 특성 및 구현 고려사항
- [0070] 전술한 바와 같이, 다양한 실시예에서, 본 명세서에 기재된 실시예의 머신 러닝 모델, 예를 들어, 머신 러닝 모델(132)은 (예를 들어, 하나 이상의 어텐션 층을 포함하는) 어텐션-기반 머신 러닝 모델일 수 있다. 머신 러닝 모델(132)은 예를 들어 하나 이상의 셀프-어텐션 층을 구현할 수 있다. 머신 러닝 모델(132)은 셀프-어텐션 메커니즘, 글로벌 어텐션 메커니즘, 소프트 어텐션 메커니즘, 로컬 어텐션 메커니즘, 및/또는 하드 어텐션 메커니즘을 사용할 수 있다.
- [0071] 일부 예에서, 어텐션-기반 머신 러닝 모델은 정렬(가령, 펩타이드 서열과 MHC 서열 간 정렬)을 학습하도록 구성될 수 있다. 정렬은 어텐션-기반 정렬 점수 함수, 예를 들어 콘텐츠-기반 함수, 가산 함수, 위치-기반 함수, 내적 함수 및/또는 스케일링된 내적 함수를 이용해 학습 및 수행될 수 있다. 머신 러닝 모델(132)은 하나 이상의 인코더, 하나 이상의 변환기 및/또는 하나 이상의 변환기 인코더를 포함할 수 있다. 일부 실시예에서, 머신 러닝 모델(132)은 Vaswani, A, et al., "Attention is All You Need." *31st Conference on Neural Information Systems*, <http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>, 2017에 기재된 하나 이상의 특징(가령, 하나 이상의 인코더)을 사용할 수 있다.
- [0072] 머신 러닝 모델(132)은 예를 들어 입력(예를 들어, 아미노산 서열, 핵산 서열, 코돈 서열 등을 나타내는 서열 표현)을 더 높은 차원 공간으로 변환하도록 구성된 하나 이상의 인코더를 포함할 수 있다. 인코더는 변압기 인코더일 수 있다. 인코더는 어텐션-기반 기술을 구현 및/또는 하나 이상의 어텐션 층(예를 들어, 하나 이상의 셀프-어텐션 층)을 포함하도록 구성될 수 있다.
- [0073] 일부 실시예에서, 머신 러닝 모델(132)은 컨볼루션 층, 장기-단기 메모리 유닛, 반복 구조 및/또는 반복 구성요소를 사용하거나 생략할 수 있다. 예를 들어, 어떤 경우에, 머신 러닝 모델(132)은 임의의 컨볼루션 층, 임의의 반복 구조, 임의의 장단기 메모리(LSTM) 유닛 및/또는 임의의 반복 구성요소를 포함하지 않는다. 일부 경우에, 머신 러닝 모델(132)은 반복적인 머신 러닝 모델이 아니거나 및/또는 반복적인 신경망을 포함하지 않는다. 일부 경우, 머신 러닝 모델은 반복 신경망을 포함 및/또는 위치 암호화를 사용하여 하나 이상의 서열에 걸친 시간 정보를 제공할 수 있다. 어떤 경우에는, 머신 러닝 모델(132)은 컨볼루션 머신 러닝 모델이 아니거나 및/또는 컨볼루션 신경망을 포함하지 않는다.
- [0074] 머신 러닝 모델(132)은 다수의 서브시스템(또는 서브망)을 포함할 수 있다. 다수의 서브시스템 각각은 인코더, 변환기 인코더, 하나 이상의 어텐션 층 및/또는 하나 이상의 셀프-어텐션 층을 포함할 수 있다. 머신 러닝 모델(132)은 IPC 서열(예를 들어, MHC 의사서열)의 일부 또는 전부의 IPC 표현을 처리하는 데 사용되는 제2 어텐션 블록과 독립적으로 펩타이드 표현을 처리하는 데 사용되는 제1 어텐션 블록을 갖는 어텐션 블록을 포함할 수 있다. 이들 어텐션 블록의 독립성은 머신 러닝 모델을 사용할 때 병렬 처리를 용이하게 할 수 있다. 또한, 독립성은 머신 러닝 모델(132)의 성능(예를 들어, 예측 정확도)을 개선할 수 있다.
- [0075] 머신 러닝 모델(132) 내에서, 임의의 주어진 층에서의 출력 값이 대응하는 입력 값뿐만 아니라 하나 이상의 또는 모든 다른 입력 값에 의존하도록 어텐션-기반 메커니즘이 구성될 수 있다. 따라서 머신 러닝 모델(132), 순

실 함수, 및/또는 최적화 함수가 (대응하는 입력에 의해 표현되는) 주어진 MHC 분자가 (또 다른 대응하는 입력에 의해 표현되는) 주어진 펩타이드에 결합 및/또는 제시 및/또는 주어진 펩타이드에 반응하여 면역원성을 트리거할 정도를 나타내는 단일 위치에 대응하는 출력을 최적화하도록 구성될 수 있다. 어떤 경우에는, 변환기 인코더의 복수의 출력 중 임의의 출력이 그러한 발생 확률을 나타낼 수 있거나 및/또는 모델이 그에 따라 훈련될 수 있다. 어떤 경우에는, 서열 시작 요소와 같은 종료점(가령, 잉여 종료점)은 결합, 제시 및/또는 면역원성 확률을 (훈련에 대한 반응으로) 나타낼 수 있다. 통합된 출력은 예를 들어 다른 층 및/또는 다른 서브시스템 또는 어텐션 블록(가령, 어텐션 층 및/또는 셀프-어텐션 층을 포함 및/또는 변환기 인코더 및/또는 인코더임)에 공급될 수 있다.

[0076] 일부 경우에, 다른 층 및/또는 다른 서브시스템 또는 어텐션 블록으로부터의 출력의 하나, 둘 또는 모든 차원은 다른 층 및/또는 다른 서브시스템 또는 어텐션 블록에 공급되는 입력과 동일한 크기를 가진다. 일부 경우에, 이 다른 층 및/또는 다른 서브시스템 또는 어텐션 블록에 공급되는 입력은 IPC 서열의 아미노산의 수, 펩타이드 서열의 아미노산의 수, 및 N-플랭크 및 C-플랭크 중 하나 이상 내 아미노산의 수의 합 이상인 하나의 축을 따른 길이를 가진다. 경우에 따라, 입력에 대한 차원의 길이는 총 아미노산 수보다 하나 더 길다. 예를 들어 추가 특징 벡터(가령, 서열-시작부분(Beginning-of-Sequence) 특징 벡터)가 아미노산-특이적(amino-acid-specific) 특징 값에 추가될 때 한 축을 따른 입력 길이는 아미노산의 합계 수를 초과할 수 있다. 입력의 다른 차원은 (가령, 하이퍼파라미터를 통해 정의되는) 다수의 특징을 포함할 수 있다. 다른 층 및/또는 다른 서브시스템 또는 어텐션 블록에 의해 생성된 출력은 입력과 동일한 크기를 가질 수 있다.

[0077] 다른 층 및/또는 다른 서브망에 의해 생성된 출력 값의 서브세트는 다른 신경망(예를 들어, 완전 연결된 피드포워드 망)에 의해 추가로 처리될 수 있다. 값의 서브세트는 특징 값의 하나의 세트에 대응할 수 있는 값의 1차원 벡터를 포함할 수 있다. 1차원 벡터는 서열-시작부분 특징 벡터와 연관된 특징 값에 대응할 수 있다.

[0078] 머신 러닝 모델(132) 내의 신경망은 하나 이상의 결과를 출력하도록 구성될 수 있다. 하나 이상의 결과는 예를 들어 숫자 결과, 이진 결과 및/또는 범주 결과를 포함할 수 있다. 하나 이상의 결과 각각은 IPC와 펩타이드가 특정 유형의 반응(가령, 함께 결합)을 겪는지 여부 및/또는 정도를 예측할 수 있다. 머신 러닝 모델(132)은 표적 시간의 결과를 생성하기 위해(예를 들어, 실수 중간 값을 이진 및/또는 범주 출력으로 변환하기 위해) 하나 이상의 활성화 층(activation layer)을 포함할 수 있다. 머신 러닝 모델(132)은 다수 유형의 예측(가령, 상호작용 예측, 상호작용 친화도 예측 및/또는 면역원성 예측)을 생성하도록 훈련될 수 있다. 어떤 경우에는, 예측이 이진 또는 범주형일 수 있다. 또 다른 예측은 비-이진 또는 비-범주형일 수 있다. 예를 들어, 예측은 스칼라일 수 있다.

[0079] 머신 러닝 모델(132)은 앙상블 모델(ensemble model)을 포함 및/또는 앙상블 모델에 포함될 수 있다. 앙상블 모델은 훈련 데이터 세트의 상이한 부분을 사용하여 훈련될 수 있는 다수의(예를 들어, 동일한) 서브-모델을 포함할 수 있다.

[0080] II.B.2. 머신 러닝 모델을 위한 예시적인 구성

[0081] 도 3은 하나 이상의 실시예에 따르는 도 1로부터의 머신 러닝 모델(132)에 대한 하나의 구성의 개략도이다. 머신 러닝 모델(132)은 도 1을 계속 참조하여 기재된다. 머신 러닝 모델(132)은 구성(300)을 가진다. 구성(300)으로, 머신 러닝 모델(132)은 표현 서브시스템(302), 초기 어텐션 서브시스템(304), 복합 서브시스템(306), 복합 어텐션 서브시스템(308) 및 출력 서브시스템(310)을 포함한다. 머신 러닝 모델(132) 내의 각각의 "서브시스템"은 하나 이상의 블록, 하나 이상의 서브블록, 하나 이상의 층, 또는 이들의 조합으로 구성될 수 있다. 머신 러닝 모델(132) 내의 각각의 "블록"은 하나 이상의 서브블록, 하나 이상의 층, 또는 이들의 조합으로 구성될 수 있다. 머신 러닝 모델(132)의 각각의 "서브블록"은 하나 이상의 층(또는 유닛)으로 구성될 수 있다.

[0082] 표현 서브시스템(302)은 서열 데이터(110)를 입력으로서 수신하고 서열 데이터(110) 내 다양한 서열에 대한 표현을 생성한다. "표현"은 예를 들어, 요소의 세트(가령, 각각의 요소가 하나 이상의 값을 포함함)를 포함할 수 있고, 각각의 요소는 표현의 모 서열 내 하나 이상의 아미노산 또는 하나 이상의 핵산을 표현하거나 식별한다. 예를 들어, 모 서열 내 각각의 아미노산은 타 아미노산을 나타내는 이진 문자열 및/또는 벡터와 구별되는 고유한 값의 이진 문자열 및/또는 벡터에 의해 표현될 수 있다.

[0083] 초기 어텐션 서브시스템(304)은 이들 표현을 입력으로서 수신하고, 이들 표현을 처리하며, 복합 서브시스템(306)으로 전송되는 변환된 표현을 생성한다. 초기 어텐션 서브시스템(304)은 다양한 어텐션 블록으로 구성되며, 각각의 어텐션 블록은 적어도 하나의 셀프-어텐션 층을 포함한다.

- [0084] 하나 이상의 실시예에서, 표현 서브시스템(302)은 펩타이드 서열(126)을 처리하여 펩타이드 표현(312)을 생성할 수 있으며, 이는 이후 초기 어텐션 서브시스템(304) 내 어텐션 블록(314)에 의해 처리되어 변환된 펩타이드 표현(316)을 생성할 수 있다. 이 처리는 도 1의 펩타이드 처리 경로(136)의 적어도 일부분을 형성할 수 있다. 또한, 표현 서브시스템(302)은 IPC 서열(124)을 처리하여 IPC 표현(318)을 생성할 수 있으며, 이는 이후 초기 어텐션 서브시스템(304) 내 어텐션 블록(320)에 의해 처리되어 변환된 IPC 표현(322)을 생성할 수 있다. 이 처리는 도 1의 IPC 처리 경로(134)의 적어도 일부분을 형성할 수 있다. IPC 서열(124)이 MHC 서열일 때, IPC 표현(318)은 MHC 표현이라 지칭되고 변환된 IPC 표현(322)은 변환된 MHC 표현이라 지칭된다. IPC 서열(124)이 TCR 서열일 때, IPC 표현(318)은 TCR 표현이라 지칭되고 변환된 IPC 표현(322)은 변환된 TCR 표현으로 지칭된다.
- [0085] 일부 실시예에서, 표현 서브시스템(302)은 N-플랭크 서열(128)을 처리하여 N-플랭크 표현(324)을 생성할 수 있고, 이는 이후 초기 어텐션 서브시스템(304)의 어텐션 블록(326)에 의해 처리되어 변환된 N-플랭크 표현(328)을 생성할 수 있다. 이 처리는 도 1의 N-플랭크 처리 경로(138)의 적어도 일부분을 형성할 수 있다. 일부 실시예에서, 표현 서브시스템(302)은 C-플랭크 서열(130)을 처리하여 C-플랭크 표현(330)을 생성할 수 있고, 이는 이후 초기 어텐션 서브시스템(304) 내 어텐션 블록(332)에 의해 처리되어 변환된 C-플랭크 표현(334)을 생성할 수 있다. 이 처리는 도 1의 C-플랭크 처리 경로(140)의 적어도 일부분을 형성할 수 있다.
- [0086] 머신 러닝 모델(132)이 또한 TCR 처리 경로(142)를 포함할 때, 표현 서브시스템(302)은 TCR 서열(131)을 처리하여 TCR 표현(336)을 생성할 수 있고, 이는 이후 초기 어텐션 서브시스템(304) 내 어텐션 블록(338)에 의해 처리되어 변환된 TCR 표현(340)을 생성할 수 있다. 이 처리는 도 1의 TCR 처리 경로(142)의 적어도 일부분을 형성할 수 있다.
- [0087] 복합 서브시스템(306)은 초기 어텐션 서브시스템(304)으로부터 출력된 변환된 표현(가령, 변환된 펩타이드 표현(316), 변환된 IPC 표현(322), 변환된 N-플랭크 표현(328), 변환된 C-플랭크 표현(334), 변환된 TCR 표현(340) 또는 이들의 조합)을 수신하고 복합 표현(342)을 생성하기 위한 하나 이상의 동작을 수행한다. 복합 표현(342)은, 예를 들어, 초기 어텐션 서브시스템(304)으로부터 출력되는 변환된 표현의 집합일 수 있다. 하나 이상의 실시예에서, 복합 표현은 초기 어텐션 서브시스템(304)으로부터 출력되는 변환된 표현들을 연결하는 연결 층(concatenation layer)을 포함할 수 있다. 일부 실시예에서, 합성 표현(342)은 하나 이상의 추가 특징 벡터(예를 들어, 변환된 표현의 시작부분 또는 끝부분에 추가될 수 있음)를 포함한다. 추가 특징 벡터는 예를 들어 각각의 모 서열에 나타난 각각의 개별 아미노산에 대응하는 특징의 수와 동일한 길이를 가질 수 있다. 부가적인 특징은 예를 들어, 서열 시작부분(BoS) 요소를 포함할 수 있다.
- [0088] 복합 표현(342)은 복합 어텐션 서브시스템(308)으로의 입력으로서 전송된다. 복합 어텐션 서브시스템(308)은 복합 표현(342)을 처리하기 위한 하나 이상의 어텐션 블록을 포함한다. 예를 들어, 복합 어텐션 서브시스템(308)은 복합 표현(342)을 수신 및 처리하는 (복합 어텐션 블록으로 지칭될 수 있는) 어텐션 블록(344)을 포함할 수 있다. 복합 어텐션 서브시스템(308)의 출력은 처리를 위해 출력 서브시스템(310)으로 전송되며, 이는 도 1에서 전술한 바와 같이 리포트(144)를 생성한다.
- [0089] 도 4a-4c는 하나 이상의 실시예에 따른 머신 러닝 모델(400)에 대한 상이한 구성의 개략도이다.
- [0090] 도 4a는 하나 이상의 실시예에 따르는 머신 러닝 모델(400)의 개략도이다. 머신 러닝 모델(400)은 도 1 및 도 3의 머신 러닝 모델(132)에 대한 구현의 한 예이다. 머신 러닝 모델(400)은 어텐션-기반 머신 러닝 모델이다. 머신 러닝 모델(400)은 표현 서브시스템(401), 초기 어텐션 서브시스템(403), 복합 서브시스템(405), 복합 어텐션 서브시스템(407), 및 출력 서브시스템(409)을 포함하며, 이들은 도 3의 표현 서브시스템(302), 초기 어텐션 서브시스템(304), 복합 서브시스템(306), 복합 어텐션 서브시스템(308), 및 출력 서브시스템(310)에 대한 각각의 구현 예이다.
- [0091] 표현 서브시스템(401)은 펩타이드 표현 블록(402) 및 IPC 표현 블록(404)을 포함한다. 일부 실시예에서, 표현 서브시스템(401)은 N-플랭크 표현 블록(406), C-플랭크 표현 블록(408) 또는 둘 모두를 더 포함한다. 일부 실시예에서, IPC 표현 블록(404)이 MHC에 대응하고 MHC 표현 블록으로서 사용될 때, 표현 서브시스템(401)은 또한 TCR 표현 블록(410)을 더 포함할 수 있다. 이들 상이한 표현 블록 각각은 적어도 하나의 임베딩 층을 포함하고 예를 들어 위치 인코더를 포함할 수 있다.
- [0092] 임베딩 층은 예를 들어 초기 비슷자 표현(가령, 아미노산 식별자의 문자열)을 숫자 표현으로 변환하여 임베딩된 표현을 생성함으로써 서열을 임베딩할 수 있다. 임베딩은 예를 들어 원-핫 암호화(one-hot encoding), 진화-동기 암호화, 가령, BLOSUM, 랜덤 또는 의사랜덤 초기화 학습된 임베딩, 또는 이들의 조합을 이용하여 수행될 수

있다. 임베딩된 표현은 위치적으로 암호화되어 암호화된 표현을 생성할 수 있다. 표현 블록에 의해 생성된 서열 표현은 암호화된 표현 또는 암호화된 표현과 임베딩된 표현의 통합(가령, 연결 또는 합산)일 수 있다.

[0093] 어떤 경우에는, 다양한 어텐션 메커니즘이 입력 데이터 세트의 값 순서에 의해 전달되는 잠재적인 정보를 검출하지 못할 수 있다. 학습되거나 고정된 암호화 알고리즘을 사용하는 위치 암호화와 함께 위치 인코더가 사용되고 임베딩된 표현에 추가될 수 있다. 예를 들어, 고정된 위치 암호화는 사인 및/또는 코사인 함수(예를 들어, 독립 변수로서 서열내 위치 및/또는 차원을 가짐)를 사용하여 정의될 수 있다. 위치 암호화는 암호화된 표현과 동일한 차원을 가질 수 있다. 위치 암호화는 초기 어텐션 서브시스템(403)으로 공급되는 서열의 위치-표시 임베딩된 표현을 생성하기 위해 임베딩된 표현과 합해질 수 있다.

[0094] 예를 들어, 펩타이드 표현 블록(402)은 펩타이드 서열(가령, 도 1의 펩타이드 서열(126))을 임베딩하여 임베딩된 펩타이드 표현을 생성하는 임베딩 층(412), 및 임베딩된 펩타이드 표현을 위치적으로 암호화하여 펩타이드 서열을 나타내는 펩타이드 표현(가령, 도 3의 펩타이드 표현(312))을 생성하는 위치 인코더(414)를 포함할 수 있다. IPC 표현 블록(404)은 임베딩된 IPC 표현을 생성하기 위해 IPC 서열(가령, 도 1의 IPC 서열(124))을 임베딩하는 임베딩 층(416), 및 IPC 서열을 표현하는 IPC 표현(가령, 도 3의 IPC 표현(318))을 생성하기 위해 임베딩된 IPC 표현을 위치적으로 암호화하는 위치 인코더(418)를 포함할 수 있다.

[0095] 또한, N-플랭크 표현 블록(406)은 임베딩된 N-플랭크 표현을 생성하기 위해 N-플랭크 서열(가령, 도 1의 N-플랭크 서열(128))을 임베딩하는 임베딩 층(420), 및 N-플랭크 서열을 나타내는 N-플랭크 표현(가령, 도 3의 N-플랭크 표현(324))을 생성하기 위해 임베딩된 N-플랭크 표현을 위치적으로 암호화하는 위치 인코더(422)를 포함할 수 있다. C-플랭크 표현(408)은 임베딩된 C-플랭크 표현을 생성하기 위해 C-플랭크 서열(예를 들어, 도 1의 C-플랭크 서열(130))을 임베딩하는 임베딩 층(424), 및 C-플랭크 서열을 표현하는 C-플랭크 표현(가령, 도 3의 C-플랭크 표현(330))을 생성하기 위해 임베딩된 C-플랭크 표현을 위치적으로 암호화하는 위치 인코더(426)를 포함할 수 있다.

[0096] 또한, TCR 표현 블록(410)은 임베딩된 TCR 표현을 생성하기 위해 TCR 서열(예를 들어, 도 1의 TCR 서열(131))을 임베딩하는 임베딩 층(428), 및 TCR 서열을 표현하는 TCR 표현(가령, 도 3의 TCR 표현(336))을 생성하기 위해 임베딩된 TCR 표현을 위치적으로 암호화하는 위치 인코더(430)를 포함할 수 있다.

[0097] 서열을 임베딩하는 것은 예를 들어 초기의 비-숫자 표현(예를 들어, 아미노산 식별자의 문자열을 포함함)을 숫자 표현으로 변환하는 것을 포함할 수 있다. 임베딩은 원-핫 암호화, 진화-동기 암호화, 가령, BLOSUM, 또는 랜덤 또는 의사랜덤 초기화 학습된 임베딩을 포함할 수 있다. 표현은 서열 및 임베딩된 서열의 위치 암호화의 합산 및/또는 통합(예를 들어, 연결(concatenation))을 포함할 수 있다.

[0098] 표현 서브시스템(401)에 의해 생성된 표현은 처리를 위해 초기 어텐션 서브시스템(403)으로의 입력으로서 전송된다. 초기 어텐션 서브시스템(403)은 표현 내 하나 이상의 또는 모든 위치 각각에 대해, 하나 이상의 다른 위치 각각의 값에 대한 어텐션 가중치(가령, 얼마나 많은 어텐션을 기울여야 하는지를 나타냄)를 결정하는 다양한 셀프-어텐션 메커니즘을 포함할 수 있다. 그런 다음 어텐션 가중치가 사용되어 위치에 대한 변환된 값을 생성할 수 있다.

[0099] 초기 어텐션 서브시스템(401)은 어텐션 블록(432) 및 어텐션 블록(434)을 포함한다. 초기 어텐션 서브시스템(401)은 또한 일부 실시예에서 어텐션 블록(436), 어텐션 블록(438), 어텐션 블록(440) 또는 이들의 조합을 포함할 수 있다. 어텐션 블록(432)은 펩타이드 표현 블록(402)으로부터 펩타이드 표현을 수신하고 어텐션 서브블록(442)의 세트를 이용해 펩타이드 표현을 처리하여 변환된 펩타이드 표현(가령, 도 3의 변환된 펩타이드 표현(316))을 생성할 수 있다. 어텐션 서브블록에 대한 한 구현예가 아래에서 도 6에 더 상세히 기재된다. 어텐션 블록(434)은 IPC 표현 블록(404)으로부터 IPC 표현을 수신하고 어텐션 서브블록(444)의 세트를 이용해 IPC 표현을 처리하여 변환된 IPC 표현(예를 들어, 도 3의 변환된 IPC 표현(322))을 생성할 수 있다.

[0100] 또한, 포함될 때, 어텐션 블록(436)은 N-플랭크 표현 블록(406)으로부터 N-플랭크 표현을 수신하고 어텐션 서브블록 세트(446)를 사용하여 N-플랭크 표현을 처리하여 변환된 N-플랭크 표현(예를 들어, 도 3의 변환된 N-플랭크 표현(328))을 생성할 수 있다. 어텐션 블록(438)은 C-플랭크 표현 블록(408)으로부터 C-플랭크 표현을 수신하고 어텐션 서브블록(448)의 세트를 이용해 C-플랭크 표현을 처리하여 변환된 C-플랭크 표현(가령, 도 3의 변환된 C-플랭크 표현(334))을 생성할 수 있다. 어텐션 블록(440)은 TCR 표현 블록(410)으로부터 TCR 표현을 수신하고 어텐션 서브블록(450)의 세트를 이용해 TCR 표현을 처리하여 변환된 TCR 표현(가령, 도 3의 변환된 TCR 표현(340))을 생성할 수 있다.

- [0101] 초기 어텐션 서브시스템(403)으로부터 출력된 변환된 표현은 처리를 위해 복합 서브시스템(405)으로 전송된다. 복합 서브시스템(405)은 복합 블록(452)을 포함한다. 복합 블록(452)은 초기 어텐션 서브시스템(403)으로부터 출력된 변환된 표현을 사용하여 복합 표현(예를 들어, 도 3의 복합 표현(342))을 형성할 수 있다. 예를 들어, 복합 블록(452)은 변환된 표현들을 통합, 연결, 또는 그 밖의 다른 방식으로 결합하여 초기 복합 표현을 형성할 수 있다. 일부 경우에, 복합 블록(452)은 또한 초기 복합 표현 내에 하나 이상의 추가적인 특징 벡터(예를 들어, BoS 벡터)를 추가한다.
- [0102] 일부 실시예에서, 복합 서브시스템(405)은 또한 위치 인코더(positional encoder)(454)를 포함할 수 있다. 위치 인코더(454)는 초기 복합 표현을 위치적으로 암호화하여 복합 어텐션 서브시스템(407)으로 출력되는 복합 표현을 생성할 수 있다. 위치 인코더(454)가 복합 서브시스템(405) 내에 존재하지 않는 경우, 복합 블록(452)에 의해 생성된 초기 복합 표현은 복합 어텐션 서브시스템(407)으로 출력되는 복합 표현일 수 있다.
- [0103] 복합 어텐션 서브시스템(407)은 어텐션 블록(456)(복합 어텐션 블록이라고도 지칭될 수 있음)을 포함할 수 있다. 어텐션 블록(456)은 어텐션 서브블록(458)의 세트를 포함한다. 어텐션 블록(456)은 복합 서브시스템(405)에 의해 생성된 복합 표현을 수신하고 어텐션 서브블록(458)의 세트를 사용하여 복합 표현을 처리하여 변환된 복합 표현을 생성할 수 있다. 이 변환된 합성 표현은 처리를 위해 출력 서브시스템(409)으로 출력된다.
- [0104] 복합 어텐션 서브시스템(407) 또는 복합 어텐션 서브시스템(407) 내의 어텐션 서브블록에 의해 생성된 출력의 크기는 복합 어텐션 서브시스템(407) 또는 복합 어텐션 서브시스템(407) 내의 어텐션 서브블록에 공급되는 입력의 크기와 동일할 수 있다. 크기는 예를 들어 $m \times n$ 일 수 있으며, 여기서 m 은 1로 간주되는 아미노산의 총 수(가령, 서열 시작부분 표현의 경우)와 같고 n 은 특징의 수(미리 결정된 값)와 같다. 추가 처리를 위해 (n 개 값을 갖는) 단일 열이 선택될 수 있다. 단일 열은 첫 번째 열 및/또는 서열 시작부분 표현과 연관된 열일 수 있다. 복합 어텐션 서브시스템(407) 또는 복합 어텐션 서브시스템(407) 내 어텐션 서브블록으로의 출력의 일부 분만이 출력 서브시스템(409)으로 공급되는 경우, 머신 러닝 모델(400)의 훈련이 IPC 서열과 펩타이드-관련 서열(들) 모두에 대한 관련 정보 및 서열 시작부분 표현에서 표현될 펩타이드-IPC 상호작용을 전달하는 학습된 파라미터 값을 도출할 수 있다. 다른 경우에, 통합된 표현은 단일 벡터를 생성하기 위해 복합 어텐션 서브시스템(407)으로부터 출력 후에 풀링될 수 있으며, 이는 이어서 출력 서브시스템(409)으로 공급될 수 있다.
- [0105] 출력 서브시스템(409)은 최종 출력을 생성하기 위해 다양한 블록, 서브블록, 층 또는 이들의 조합을 포함할 수 있다. 하나 이상의 실시예에서, 출력 서브시스템(409)은 드롭아웃 블록(460), 완전 연결된 블록(462) 및 출력 블록(464)을 포함한다. 드롭아웃 블록(460)은 예를 들어 하나 이상의 드롭아웃 층을 포함할 수 있다. 완전 연결 블록(462)은 예를 들어 하나 이상의 완전 연결 층을 포함할 수 있다. 출력 블록(464)은 예를 들어 출력을 필터링, 선택, 변환 또는 그 밖의 다른 방식으로 생성하기 위한 하나 이상의 층을 포함할 수 있다. 예를 들어, 출력 블록(464)은 예를 들어 선택된 임계값 또는 범위에 기초하여 출력 블록(464)에서 수신된 입력의 서브셋을 선택하도록 구성되는 적어도 하나의 최대 층(465)을 포함할 수 있다.
- [0106] 일부 경우에, 변환된 복합 표현은 완전 연결 블록(462)에 의해 수신되는 제1 출력을 생성하기 위해 드롭아웃 블록(460)에 의해 수신 및 처리된다. 완전 연결 블록(462)은 이 제1 출력을 수신 및 처리하여 제2 출력을 생성할 수 있으며, 그 중 적어도 일부는 출력 블록(464)에 의해 수신된다. 출력 블록(464)은 이의 입력을 수신하고 처리하여 상호작용 출력(466), 면역원성 출력(468) 또는 둘 모두를 생성할 수 있다.
- [0107] 일부 실시예에서, 완전히 연결된 블록(462)은 완전히 연결된 블록(462)에 공급된 차원보다 작은(예를 들어, 특징의 지정된 수보다 작은) 차원을 갖는 하나 이상의 출력을 생성하도록 구성될 수 있다. 예를 들어, 완전 연결 블록(462)의 출력은 단일 값, 2개의 값, 또는 3개의 값을 포함할 수 있으며, 각각은 표적 상호작용 또는 면역 반응과 관련된 예측에 대응한다. 완전 연결 블록(462)은 예를 들어 단일 은닉 층, 두 개의 은닉 층 또는 세 개 이상의 은닉 층을 포함할 수 있다. 초기 은닉 층 내 노드 수는 후속 은닉 층의 노드 수보다 클 수 있다. 예를 들어, 제1 은닉 층은 256개의 노드를 포함할 수 있고 제2 은닉 층은 126개의 노드를 포함할 수 있다. 다양한 실시예에서, 완전 연결 블록(462)으로부터의 각각의 출력은 예를 들어 (예를 들어, 훈련된 활성화 함수를 사용하여) 이진 및/또는 범주 결과로 변환 및/또는 스케일링된 값으로 변환될 수 있는 실수 점수를 포함할 수 있다. 예를 들어, 스케일링된 숫자는 0에서 1까지의 스케일에 대한 확률을 포함할 수 있다.
- [0108] 상호작용 출력(466)은 예를 들어, 상호작용 예측(470)의 세트, 상호작용 친화도 예측(472)의 세트, 또는 하나 이상의 표적 상호작용에 대해 둘 모두를 포함할 수 있다. 상호작용 예측은 예를 들어 IPC(예를 들어, MHC, TCR)가 펩타이드에 결합할지 여부에 대한 대응하는 펩타이드-IPC(가령, 펩타이드-MHC, 펩타이드-TCR) 결합체에 대한 예측을 포함할 수 있다. 상호작용 예측은 예를 들어 IPC(예를 들어, MHC)가 세포 표면에서 펩타이드를 제시

할지 여부에 대한 대응하는 펩타이드-IPC(가령, 펩타이드-MHC) 결합체에 대한 예측을 포함할 수 있다. 또한, 상호작용 친화도 예측은 예를 들어 대응하는 펩타이드-IPC(예를 들어, 펩타이드-MHC, 펩타이드-TCR) 결합체에 대한 표적 상호작용에 대한 친화도의 예측을 포함할 수 있다. 표적 상호작용은 예를 들어 펩타이드와 IPC의 결합일 수 있다. 예를 들어, 결합 친화도일 수 있는 표적 상호작용에 대한 친화도는 펩타이드와 IPC 간의 결합의 강도, 경향 및/또는 안정성을 나타낸다.

[0109] 면역원성 출력(466)은 면역원성 예측의 세트를 포함한다. 면역원성 예측은 예를 들어 대응하는 펩타이드-IPC 결합체에 대한 면역원성 예측을 포함할 수 있다. 예를 들어, 면역원성 예측은 특정 관심 IPC(가령, TCR 또는 MHC 및 TCR 복합체)에 대해 면역 반응을 유발할 수 있는 펩타이드의 능력을 나타낼 수 있다.

[0110] 일부 경우에, 완전 연결 블록(462)으로부터의 출력의 제1 부분은 출력 블록(464)으로 전송되는 반면, 완전 연결 블록(462)으로부터의 출력의 제2 부분은 최종 형태이고 상호작용 친화도 예측(472)의 세트로서 사용된다.

[0111] 다른 실시예에서, 출력 서브시스템(409)에서 수신된 변환된 복합 표현은 완전 연결 블록(462)에 의해 수신되고 처리되며, 이는 드롭아웃 블록(460)으로 전송되는 제1 출력을 생성하기 위해 변환된 복합 표현을 처리한다. 그 후 처리를 위해 출력 블록(460)의 출력 또는 그 일부가 출력 블록(464)으로 전송될 수 있다.

[0112] 일부 실시예에서, 출력 서브시스템(409)으로부터의 출력은 각각의 IPC(예를 들어, MHC) 대립유전자에 대해, 펩타이드가 IPC 대립유전자에 결합하는지 여부 및/또는 결합 확률에 대한 예측을 포함하는 다수의 결과를 포함할 수 있다. 대립유전자 특이적 예측이 출력될 수 있거나, 어떤 경우에는 대립유전자 특이적 예측의 최댓값을 결정하기 위해 맥스 층(max layer)(465)이 사용될 수 있고, 최댓값이 출력될 수 있다.

[0113] 이러한 방식으로, 출력 서브시스템(409)은 상호작용 출력(466), 면역원성 출력(468), 또는 둘 모두의 생성을 가능하게 하는 임의의 개수의 상이한 블록, 서브블록, 및/또는 층에 의해, 여러 다른 방식 중 임의의 방식으로 구현될 수 있다. 복합 서브시스템(405) 이전에 IPC 서열(예를 들어, MHC 서열, TCR 서열, 결합된 MHC-TCR 서열 등)의 처리와 별개로 펩타이드 서열을 처리하는 것은 머신 러닝 모델(400)의 예측 성능을 증가시킨다. 예를 들어, 복합 표현을 생성하기 전에 IPC 표현 블록(404) 및 어텐션 블록(434)을 이용해 변환된 IPC 표현을 생성하는 것과 분리(그리고 적용 가능하다면, TCR 표현 블록(410) 및 어텐션 블록(440)을 이용해 변환된 TCR 표현의 생성과 분리)되는 경로를 따라 펩타이드 표현 블록(402) 및 어텐션 블록(432)을 이용해 변환된 펩타이드 표현을 생성하는 것이 출력 서브시스템(409)에 의해 생성된 출력의 정확도를 증가시킨다. 또한, 다수의 펩타이드-IPC(및 펩타이드-TCR) 결합체가 모듈식으로 고려될 수 있기 때문에 이러한 처리가 (가령, 감소된 컴퓨팅 자원, 더 빠른 처리 등을 이용해) 효율적인 처리를 가능하게 할 수 있다.

[0114] 다양한 실시예에서, 머신 러닝 모델(400)은 어느 특정 IPC 대립유전자가 펩타이드에 결합하여 제시하는 것으로 예측되는지에 대한 자동화된 결정을 용이하게 할 수 있다. 예를 들어, MHC 분자가 6개의 MHC 대립유전자를 포함하는 경우(인간의 경우와 같이), 신경망 처리의 적어도 일부분의 6회 반복이 — 각각의 대립유전자에 하나씩 — (가령, 병렬로) 수행될 수 있다. 각각의 처리는, 입력으로서 MHC 대립유전자에 대한 MHC 서열의 MHC 표현 및 펩타이드의 서열의 적어도 일부분의 펩타이드 표현을 사용할 수 있다. 각각의 처리는 펩타이드가 MHC 대립유전자에 결합 및/또는 제시될지 여부에 대한 예측에 대응하는 출력을 생성할 수 있다. 대립유전자에 걸쳐 가장 높은 예측 값과 연관된(예를 들어, 가장 가능성 있는 결합 및/또는 제시 예측을 나타내는) 펩타이드가 펩타이드가 결합할 것 및 펩타이드를 제시할 것이라고 추론할 수 있다.

[0115] 일부 예에서, 6개의 MHC 대립유전자에 대해, 동일한 IPC 표현 블록(404)을 통해 6개의 상이한 MHC 대립유전자 서열을 실행하고 각각의 대립유전자-펩타이드 결합체에 대한 복합 표현을 생성함으로써 6개의 복합 표현이 생성될 수 있다. 일부 실시예에서, 6개의 복합 표현 각각은 임베딩 층에 의해 임베딩된 서열 시작부분 토큰(벡터)과 함께, 통합(예를 들어, 연결)될 수 있다. 그런 다음 6개의 복합 표현 각각은 전술한 바와 같이 복합 서브시스템(407)을 통해 공급될 수 있다.

[0116] 일부 실시예에서, 처리된 서열 시작부분 토큰이 추출되어 완전 연결 블록(462)으로 공급되어 머신 러닝 모델(400)의 최종 노드로 직접 출력될 수 있다. 이 BoS 토큰은 노드 제시 가능성을 나타낼 수 있다. 일부 경우에, 완전 연결 블록(462) 내의 각각의 완전 연결 서브블록은 드롭아웃 적용될 수 있고 배치 정규화 층이 뒤따를 수 있다. 일부 실시예에서, 출력 블록(464)은 디콘볼루션에 사용되어, (가령, 소프트맥스 함수를 포함할 수 있는 맥스 층(465)을 통해) ~6 제시 예측에 활성화 함수를 적용함으로써 ~6 쌍을 이룬 펩타이드-MHC 상호작용이 단일 선택된 MHC 대립유전자에 대응할 것이다. 훈련 동안, 선택된 펩타이드-MHC 상호작용 출력은 0과 1 사이의 값으로 정규화될 수 있고 손실 함수(가령, 이진 손실 함수)를 이용하여 진실한 표시 값에 비교되어 모델 파라미터를

튜닝하기 위한 오차를 생성할 수 있다.

- [0117] 또 다른 실시예에서, 머신 러닝 모델(400)에 포함된 어텐션 블록 또는 어텐션 서브블록 중 하나 이상이 또 다른 유형의 망 및/또는 처리 유닛으로 대체되어 하나 이상의 서열의 표현을 변환할 수 있다. 변환은 (특정 위치에서의) 다양한 아미노산이 결합 친화도 및/또는 제시 확률에 영향을 미치는 것으로 예측되는 정도 및/또는 (특정 위치에서의) 아미노산의 다양한 특정 조합이, 단일 서열에서 또는 서열들에 걸쳐 발생하는, 결합 친화도 및/또는 제시에 영향을 미칠 것으로 예측된다. 예를 들어, 하나 이상의 어텐션 서브블록은 하나 이상의 게이트 반복 단위로 대체될 수 있다.
- [0118] 도 4b는 하나 이상의 실시예에 따르는 머신 러닝 모델(400)에 대한 상이한 구성의 개략도이다. 도 4b에 도시된 구성으로, 표현 서브시스템(401)은 통합 표현 블록(480)을 포함한다. 통합 표현 블록(480)은 통합 서열, 가령, 펩타이드 서열의 통합(예를 들어, 도 1의 펩타이드 서열(126)) 및 N-플랭크 서열(예를 들어, 도 1의 N-플랭크 서열(128)) 및/또는 C-플랭크 서열(예를 들어, 도 1의 C-플랭크 서열(130))을 수신한다.
- [0119] 통합 표현 블록(480)은 예를 들어 통합 서열을 처리하여, 임베딩된 통합 표현을 위치적으로 암호화하는 위치 인코더(483)에 의해 수신될 수 있는 임베딩된 통합 표현을 형성하는 임베딩 층(482)을 포함하여 통합 표현(484)을 생성할 수 있다. 예를 들어, 통합 표현(484)은 모 펩타이드 서열의 펩타이드 표현(485) 및 모 N-플랭크 서열의 N-플랭크 표현(486) 및/또는 모 C-플랭크 서열의 C-플랭크 표현(487)을 포함할 수 있다.
- [0120] 통합 표현(484)은 통합 표현 블록(480)으로부터 출력되고 처리를 위해 초기 어텐션 서브시스템(403)의 어텐션 블록(488)으로 전송된다. 어텐션 블록(488)은 처리를 위해 복합 블록(452)으로 전송되는 변환된 통합 표현을 생성하기 위해 통합 표현(484)을 처리하는 어텐션 서브블록(489)의 세트를 포함한다.
- [0121] 일부 실시예에서, 통합 표현 블록(480)으로 전송된 통합 서열이 N-플랭크 서열 또는 C-플랭크 서열을 포함하지만 다른 것은 포함하지 않는 경우, 머신 러닝 모델(400)은 또한 통합 서열에 포함되지 않는 서열에 대한 대응하는 표현 블록(예를 들어, N-플랭크 표현 블록(406) 또는 C-플랭크 표현 블록(408)) 및 대응하는 어텐션 블록(가령, 어텐션 블록(436) 또는 어텐션 블록(438))을 포함할 수 있다.
- [0122] 도 4c는 하나 이상의 실시예에 따르는 머신 러닝 모델(400)에 대한 상이한 구성의 개략도이다. 도 4c에 도시된 구성으로, 표현 서브시스템(401)에 의해 생성된 펩타이드 표현 및 N-플랭크 표현, 및 선택적으로 C-플랭크 표현은 통합 블록(490)으로 전송된다. 통합 블록(490)은 이들 표현들을 통합(가령, 연결)하여 어텐션 블록(492)으로 전송되는 통합 표현을 형성할 수 있다. 어텐션 블록(492)은 처리를 위해 복합 블록(452)으로 전송되는 변환된 통합 표현을 생성하기 위해 통합 표현을 처리하는 어텐션 서브블록(494)의 세트를 포함한다.
- [0123] 도 4a-4c에 도시된 바와 같이, 머신 러닝 모델(400)은 다양한 서브시스템 내의 블록, 서브블록 및/또는 층의 임의의 수 또는 조합을 사용하여 다양한 방식으로 구현될 수 있다. 따라서 머신 러닝 모델(400)은 모듈식이며 주어진 작업에 대해 맞춤 구성될 수 있다.
- [0124] 도 5는 하나 이상의 실시예에 따르는 어텐션 블록(500)의 개략도이다. 어텐션 블록(500)은 도 3의 초기 어텐션 서브시스템(304), 도 3의 복합 어텐션 서브시스템(308), 또는 도 4a-c의 초기 어텐션 서브시스템(403) 내 어텐션 블록에 대한 구현의 한 예일 수 있다. 또한, 어텐션 블록(500)은 도 4a-4c의 어텐션 블록(456)에 대한 구현의 한 예일 수 있다.
- [0125] 어텐션 블록(500)은 하나 이상의 어텐션 서브블록을 포함한다. 예를 들어, 어텐션 블록(500)은 어텐션 서브블록 1(501) 및 선택적으로 어텐션 서브블록 n (504)까지의 하나 이상의 다른 어텐션 서브블록을 포함할 수 있다. 다수의 어텐션 서브블록이 어텐션 블록(500)에서 제시될 때, 이들 어텐션 서브블록은 직렬로 연결될 수 있다(가령, 최종 출력을 생성하기 위해 함께 데이터-체인 연결될 수 있다).
- [0126] 어텐션 서브블록 1(501)은 다양한 방식으로 구현될 수 있다. 하나 이상의 실시예에서, 어텐션 서브블록 1(501)은 예를 들어, 셀프-어텐션 층(506), 추가 및 정규화 층(508), 피드 포워드 층(510), 및 추가 및 정규화 층(512)을 포함한다. 어텐션 서브블록(501)에 대한 이 구성을 이용해, 어텐션 서브블록 1(501)은 변환기 인코더로도 지칭될 수 있다. 셀프-어텐션 층(506)은 예를 들어, 원-헤드 어텐션 유닛 또는 멀티-헤드 어텐션 유닛을 사용하여 구현될 수 있다. 존재한다면, 어텐션 블록(500) 내 어텐션 서브블록 n 까지의 하나 이상의 다른 어텐션 서브블록이 어텐션 서브블록 1(501)과 유사한 방식으로 구현될 수 있다.
- [0127] 추가 및 정규화 층에서, 변환된 표현은 (간기 연결을 통해) 서열의 위치 표시 임베딩된 표현에 추가될 수 있으며 합산된 표현이 정규화될 수 있다. 정규화된 데이터는 대응하는 피드 포워드 층(510)(예를 들어, 완전 연결

피드포워드 망)에 공급될 수 있다. 피드포워드 망은 (예를 들어) 각각의 위치에 대해, 하나, 둘, 셋 또는 그 이상의 선형 변환에 영향을 미칠 수 있고 및/또는 각각의 선형 변환 사이에 활성화(가령, ReLU 활성화)를 포함할 수 있다. 예를 들어 피드포워드 층은 다음과 같이 나타낼 수 있다:

$$FF(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

여기서 x 는 층으로의 입력이고 W_1 및 W_2 는 선형 변환의 기울기이고 b_1 및 b_2 는 선형 변환의 절편이다. 특정 어텐션 서브블록의 피드 포워드 층의 출력의 차원은 어텐션 서브블록의 피드 포워드 층으로의 입력의 차원과 동일할 수 있다. 따라서 어떤 경우에는 다양한 유형의 정보의 표현을 보존하기 위해 입력 및 출력이 (가령, 다른 추가 및 정규화 층을 통한 다른 잔기 연결을 통해) 합산되고 정규화할 수 있다.

II.B.3. 셀프-어텐션을 위한 예시적 메커니즘

도 6은 하나 이상의 실시예에 따르는 예시적 셀프-어텐션 층을 이용한 서열 표현을 처리하기 위한 프로세스의 흐름도이다. 프로세스(600)는 예를 들어 도 1 및 3의 머신 러닝 모델(132) 내에 존재하는 어텐션 블록 중 하나 이상, 도 4a-4c 내 머신 러닝 모델(400) 내에 존재하는 어텐션 블록 중 하나 이상 및/또는 도 5의 어텐션 블록(500)에 의해 사용될 수 있다.

단계(602)는 복수의 요소를 포함하는 서열 표현을 수신하는 것을 포함한다. 서열 표현은 유전자 서열 내 아미노산 서열 또는 유전적 핵산 서열, 또는 코돈 서열을 나타낸다. 하나 이상의 실시예에서, 서열 표현 내 복수의 요소의 각각의 요소는 아미노산(또는 아미노산 잔기), 핵산, 코돈 등을 나타낸다. 또한, 각각의 요소는 서열에서 고유한 위치와 연관된다.

서열 표현은 예를 들어 펩타이드 표현, IPC 표현, N-플랭크 표현, C-플랭크 표현, MHC 표현, TCR 표현, 통합 표현 또는 다른 유형의 표현일 수 있다. 예를 들어, 서열 표현은 변이-코딩 서열, 야생형 또는 돌연변이 펩타이드를 암호화하는 서열의 일부 또는 전부, (가령, 변이를 포함하는) 항원결정기 서열, 후보 신생항원결정기 서열, 신생항원 서열 중 일부 또는 전부, 펩타이드의 말단(가령, N-플랭크 또는 C-플랭크)에서 시작하거나 종료하는 서열, MHC 서열(가령, MHC 의사서열) 중 일부 또는 전부를 나타낼 수 있다. 서열 표현은 예를 들어 도 3의 표현 서브시스템(302) 또는 도 4a-4c의 표현 서브시스템(401)을 이용해 생성될 수 있다.

단계(304)는 각각 키 가중치의 세트, 값 가중치의 세트 및 쿼리 가중치의 세트를 사용하여 서열 표현의 각각의 요소에 대한 키 벡터, 값 벡터 및 쿼리 벡터를 결정하는 것을 포함한다. 예를 들어, 서열 표현에 나타난 서열이 예를 들어 20개의 아미노산을 포함하는 경우, 20개의 키 벡터, 20개의 값 벡터 및 20개의 쿼리 벡터가 생성될 수 있다. 서열 표현의 요소는 예를 들어 2차원 서열 표현의 행 또는 열에 대응할 수 있다(예를 들어, 여기서 제 1 차원은 서열 내 상이한 아미노산을 나타내고 제2 차원은 예를 들어 개별 아미노산을 특징짓는 상이한 성분을 나타낸다).

일부 실시예에서, 키 가중치의 세트는 키 가중치 매트릭스의 형태이다. 특정 요소에 대한 키 가중치 행렬은 요소의 길이에 키 벡터의 길이를 곱한 것과 동일한 크기를 가질 수 있다. 예를 들어, 요소는 20의 길이를 가질 수 있으며(예를 들어, 각각의 값이 서열 내 아미노산이 21개 아미노산 중 특정 1개와 동일한지 여부에 대한 이진 표시자에 대응함), 키 벡터의 길이가 5인 경우(가령, 5개의 구성요소 또는 특징을 나타냄), 키 가중치 행렬이 [5, 21]의 크기를 가질 수 있다. 키 가중치 행렬은 훈련 중에 학습될 수 있다(가령, 훈련 시작 시 무작위로 초기화됨).

요소에 대한 값 벡터는 요소에 대한 키 벡터와 동일한 크기를 가질 수 있다. 값 벡터는 훈련 중에 학습될 수 있고 값 가중치 행렬 내에 포함될 수 있는 값 가중치 세트를 사용하여 결정될 수 있다. 주어진 요소에 대한 값 가중치 행렬은 키 가중치 행렬의 크기를 가질 수 있고 및/또는 그 요소의 길이 및 값 벡터의 길이에 기초하여 정의된 크기를 가질 수 있다.

요소에 대한 쿼리 벡터는 요소에 대한 키 벡터 및/또는 값 벡터와 동일한 크기를 가질 수 있다. 쿼리 벡터는 훈련 중에 학습될 수 있고 쿼리 가중치 행렬 내에 포함될 수 있는 쿼리 가중치의 세트를 이용해 결정될 수 있다. 요소에 대한 쿼리 가중치 행렬은 키 가중치 행렬 및/또는 값 가중치 행렬의 크기를 가질 수 있고 및/또는 요소의 길이 및 쿼리 벡터의 길이에 기초하여 정의된 크기를 가질 수 있다.

단계(606)는 서열 표현 내 각각의 요소에 대해, 요소의 쿼리 벡터(쿼리 가중치 및 서열 표현을 사용하여 생성됨) 및 다중 요소의 키 벡터(키 가중치 및 서열 표현을 이용해 생성됨)를 사용하여 요소 중심 어텐션 점수

세트를 생성하는 것을 포함한다. 주어진 요소에 대해, 요소 중심 어텐션 점수의 세트는 주어진 요소의 값 벡터에 얼마나 많은 가중치를 부여할지를 나타낼 수 있다. 키 벡터가 서열 표현 내 선택된 요소에 대한 요소 중심 어텐션 점수의 세트를 생성하는 데 사용되는 다수의 요소가 서열 표현 내 일부 또는 모든 요소(가령, 아미노산 표현 중 일부 또는 전체 표현)를 포함할 수 있다. 다수의 요소는 초점 요소(예를 들어, 요소 중심 어텐션 점수의 세트가 결정되는 특정 아미노산)를 포함할 수 있다.

[0139] 요소 중심 어텐션 점수의 세트는 서열 표현의 각각의 요소에 대해, 초점 요소(제1 요소)와 동일하거나 상이한 요소(제2 요소)의 각 쌍에 대한 어텐션 점수를 생성함으로써 생성된다. 이 페어링에 대한 어텐션 점수는 제1 요소의 쿼리 벡터와 제2 요소의 키 벡터의 곱으로 정의될 수 있다.

[0140] 일부 경우에, 단계(606)는 활성화 함수 및/또는 정규화를 구현하는 것을 포함할 수 있다. 정규화는 키 벡터(또는 쿼리 벡터)의 차원을 기반으로 할 수 있다. 예를 들어 정규화는 키 벡터 길이의 제곱근으로 정의될 수 있다. 활성화 함수는 소프트맥스(softmax) 함수를 포함할 수 있다. 경우에 따라 정규화가 활성화 함수 전에 적용된다.

[0141] 단계(608)는 복수의 요소의 변환을 수행하여 복수의 수정된 요소를 형성하는 단계를 포함하며, 여기서 변환은 복수의 요소의 각각에 대해 생성된 요소-중심 어텐션 점수의 세트 및 복수의 요소의 각각에 대해 결정된 값 벡터를 사용하여 수행된다. 예를 들어, 서열 표현이 11개의 요소(예를 들어, 11개의 아미노산을 나타냄)를 포함하는 경우, 그리고 요소의 모든 쌍별 조합에 대해 어텐션 점수가 결정되는 경우, 복수의 변이된 요소를 포함하는 변이된 서열 표현이 생성되며, 여기서, 변이된 요소는 (가중을 위한 어텐션 점수 사용) 모든 요소의 값 벡터의 가중 평균으로 정의될 수 있다.

[0142] 단계(610)는 변환된 서열 표현, 초기 서열 표현, 및 피드포워드 망을 사용하여 서열의 암호화를 생성하는 것을 포함한다. 예를 들어, 변환된 서열 표현과 초기 서열 표현이 합산될 수 있다. 이 결과는 여전히 여러 요소를 포함할 수 있다(가령, 변환, 합산 및 정규화를 통해 각각 업데이트된 것). 그런 다음 피드포워드 신경망은 (가령, 하나, 둘 또는 그 이상의 선형 변환을 수행 및/또는 하나 이상의 활성화 함수를 구현함으로써) 합산된 표현을 처리할 수 있다. 표현을 합산하면 변환된 서열 표현에서 (주어진 요소에 대해 변환된 값 벡터를 생성할 때 다른 요소의 값에 어텐션을 기울이기 때문에) 가려질 수 있는 위치 정보를 다시 도입시킬 수 있다.

[0143] 피드포워드 신경망은 (예를 들어, 동일한 기술 및/또는 동일한 파라미터의 세트를 사용하여) 업데이트된 다중 요소의 각각을 개별적으로 처리하도록 구성될 수 있다. 따라서, 피드포워드 망에 대한 입력은 단일 요소, 단일 아미노산 및/또는 단일 서열 위치에 대응하는 벡터를 포함할 수 있다. 피드포워드 망은 피드포워드 망의 출력이 피드포워드 망으로의 입력과 동일한 크기가 되도록 구성될 수 있다. 일부 경우에, 피드포워드 망을 사용하여 변환된 서열 표현 및 초기 서열 표현을 처리하는 대신에, 컨볼루션(가령, 1차원 컨볼루션)이 대신 채용되어 위치/요소에 걸쳐 동일하게 작동하는 국소화된 변환을 수행할 수 있다. 1차원 컨볼루션은 피드포워드 신경망의 기능을 해석하는 또 다른 방법으로 사용될 수 있다.

[0144] 도 6에 설명된 기법은 싱글-헤드 어텐션(single-head attention)(키 벡터, 값 벡터 및 쿼리 벡터가 어텐션 점수를 계산하는 데 사용됨)와 관련이 있다. 멀티-헤드 어텐션(multi-head attention)이 대안으로 사용될 수 있다. 멀티-헤드 어텐션 내 각각의 어텐션 헤드는 키 가중치의 고유 세트, 값 가중치의 고유 세트, 및 쿼리 가중치의 고유 세트와 연관될 수 있다. 그런 다음 멀티-헤드 어텐션 내 각각의 어텐션 헤드는 별개의 키 벡터, 별개의 값 벡터 및 별개의 쿼리 벡터를 생성할 수 있다. 멀티-헤드 어텐션 내 각각의 어텐션 헤드는 이들 별개의 벡터를 사용하여 각각의 요소에 대한 어텐션 점수 및 변환된 값을 생성할 수 있다. 변환된 값들이 연결되고 투영될 수 있다.

[0145] 도 6이 다양한 벡터의 계산 및 사용을 나타내지만, 행렬 표현이 대신 사용될 수 있음을 추가로 이해해야 한다. 행렬 표현은 다양한 벡터를 개별적으로 반복적으로 계산하는 것과는 반대로 요소들에 걸쳐 계산을 효율적으로 수행하는 데 도움이 될 수 있다.

[0146] 도 7은 하나 이상의 실시예에 따르는 도 6에 기재된 프로세스(600)를 예시하는 개략도이다. 도 7에서, 표현 및 어텐션 프로세스(700)는 입력으로서 서열(702)을 수신한다. 서열(702)은 예를 들어 아미노산 서열일 수 있다.

[0147] 도 7의 예시적인 예에서, 서열(702)은 복수의 아미노산(704)(4개의 아미노산: x^1-x^4)을 포함한다. 복수의 요소 a^1-a^4 를 포함하는 서열 표현(706)은 임베딩과, 일부 실시예에서는 위치 암호화를 통해 생성된다. 각각의 요소 a^i 는 예를 들어 숫자 벡터를 가질 수 있다. 서열 표현(706)은 도 6의 단계(602)에서 수신된 서열 표현의 한 예일 수 있다.

- [0148] 벡터(708)(예를 들어, 쿼리 벡터 q^i , 키 벡터 k^i 및 값 벡터 v^i)는 각각의 요소 a^i 에 대해 생성될 수 있다. 벡터 (708)는 도 6의 단계(604)에서 생성된 벡터에 대한 구현의 예일 수 있다. 도시된 예는 제1 요소 a_1 에 초점을 두고 선택 요소-중심 어텐션 점수(710), $\hat{a}_{1,i}$ 를 생성하는 것에 대응한다. 요소-중심 어텐션 점수(710)는 도 6의 단계(606)에서 특정 요소에 대해 생성된 요소-중심 어텐션 점수의 하나의 세트의 예이다. 요소-중심 어텐션 점수 $\hat{a}_{1,i}$ 의 각각은 q^1 과 k^i 의 내적하도록 정의된다. 가중치가 $\hat{a}_{1,i}$ 로 설정될 때 값 벡터 v^i 의 가중 합은 수정 요소 (712) b^1 를 생성한 변환을 수행하기 위해 계산된다. 수정 요소(712)는 도 6의 단계(608)에서 생성된 수정 요소의 한 예이다. 유사한 변환이 서열 표현(706)의 다른 요소에 대해 수행될 수 있다.
- [0149] II.C. 머신 러닝 모델을 사용한 예시적인 방법론
- [0150] 도 1 및 도 3의 머신 러닝 모델(132) 및 도 4a-4c의 머신 러닝 모델(400)은 돌연변이 펩타이드(예를 들어, 신생 항원)를 포함하는 다양한 펩타이드와 연관된 면역학적 활동(예를 들어, 예측된 결합, 결합 친화도, 예측된 제시 발생, 면역원성 등)에 대한 예측을 생성하기 위해 다양한 방식으로 사용될 수 있다.
- [0151] 도 8은 다양한 펩타이드의 면역학적 활동에 대한 정보를 생성하기 위한 프로세스의 흐름도이다. 프로세스(800)의 적어도 일부분은, 비제한적 예를 들어, 도 1에 기재된 예측 시스템(100)을 사용하여 구현될 수 있다. 예를 들어 프로세스(800)의 적어도 일부분은, 비제한적 예를 들어, 도 1 및 3의 머신 러닝 모델(132) 또는 도 4a-4c의 머신 러닝 모델(400)을 이용해 구현될 수 있다.
- [0152] 단계(802)는 돌연변이 펩타이드를 특징 짓는 펩타이드 서열을 수신하는 단계를 포함하며, 상기 펩타이드 서열은 대응하는 참조 서열에 대한 변이를 포함한다. 펩타이드 서열은 돌연변이 펩타이드의 적어도 일부분을 특징지움으로써 돌연변이 펩타이드를 특징짓는다. 돌연변이 펩타이드는 예를 들어 신생항원일 수 있다. 단계(802)는 예를 들어 데이터 저장소(예를 들어, 도 1의 데이터 저장소(104), 클라우드 저장소, 서버 또는 서버 시스템 등)로부터 펩타이드 서열을 불러옴으로써 수행될 수 있다. 일부 구현예에서, 펩타이드 서열은 머신 러닝 모델을 통해 처리되는 복수의 펩타이드 서열 중 하나일 수 있다.
- [0153] 단계(804)는 면역단백질 복합체(IPC)에 대해 식별된 면역단백질 복합체(IPC) 서열을 수신하는 것을 포함한다. IPC는 예를 들어 MHC, TCR 또는 MHC-TCR 복합체일 수 있다. 따라서, IPC 서열은 MHC 서열, TCR 서열 또는 MHC-TCR 서열일 수 있다. IPC 서열은 IPC의 적어도 일부분을 특징지움으로써 IPC를 특징짓는다. 단계(802)는 예를 들어 데이터 저장소(예를 들어, 도 1의 데이터 저장소(104), 클라우드 저장소, 서버 또는 서버 시스템 등)로부터 IPC 서열을 불러옴으로써 수행될 수 있다. 일부 실시예에서, IPS 서열은 머신 러닝 모델을 통해 처리되는 복수의 IPC 서열 중 하나일 수 있다.
- [0154] 단계(806)는 출력을 생성하기 위해 어텐션-기반 머신 러닝 모델 내에서 상이한 처리 경로를 사용하여 펩타이드 서열 및 IPC 서열을 처리하는 것을 포함하며, 여기서 출력은 돌연변이 펩타이드 및 IPC 모두와 관련된 면역학적 활동에 대한 정보를 제공한다. 단계(806)는 예를 들어, 펩타이드 서열을 나타내는 변환된 펩타이드 표현을 생성하기 위해 대응하는 어텐션 블록을 통해 처리되는 펩타이드 표현을 생성하기 위해 대응하는 표현 블록을 통해 펩타이드 서열을 처리하는 것을 포함한다. 이 펩타이드 처리 경로는, IPC 서열을 나타내는 변환된 IPC 표현(가령, 변환된 MHC 표현, 변환된 TCR 표현, 변환된 MHC-TCR 표현)을 생성하기 위해 대응하는 어텐션 블록을 통해 처리되는 IPC 표현(가령, MHC 표현, TCR 표현, MHC-TCR 표현)을 생성하기 위해 IPC 서열이 대응하는 표현 블록을 통해 처리되는 IPC 처리 경로와 분리된다.
- [0155] 일부 구현예에서, 펩타이드 표현은 N-플랭크 서열에 대한 N-플랭크 표현 및/또는 C-플랭크 서열에 대한 C-플랭크 표현을 또한 포함하는 통합 표현의 일부이다. 이러한 실시예에서, 통합 처리 경로(본질적으로 펩타이드 처리 경로를 포함할 것임)는 IPC 처리 경로와 별개로 유지된다.
- [0156] 다양한 실시예에서, 단계(806)에서, 변환된 펩타이드 표현 및 변환된 IPC 표현은 출력을 생성하기 위해 추가로 처리되는 복합 표현을 형성하기 위해 사용된다. 예를 들어, 복합 표현은 어텐션 블록을 사용하여 변환되어 출력을 생성하도록 처리되는 변환된 복합 표현을 생성할 수 있다. 출력은, 비제한적 예를 들면, 상호작용 예측의 세트, 상호작용 친화도 예측의 세트, 면역원성 예측의 세트 또는 이들의 조합을 포함할 수 있다.
- [0157] 단계(808)는 출력에 기초하여 리포트를 생성하는 것을 포함한다. 리포트는 출력을 포함할 수 있다. 다른 실시예에서, 리포트는 출력의 변환되거나 필터링된 버전을 포함한다. 또 다른 실시예에서, 리포트는 출력의 요약, 시놉시스 또는 시각적 표현을 포함한다.

- [0158] 일부 실시예에서, 프로세스(800)는 단계(810)를 더 포함한다. 단계(810)는 리포트에 기초하여 작업의 세트를 수행하는 것을 포함한다. 작업의 세트는 리포트를 기반으로 하는 치료제의 설계 및/또는 제조와 관련된 다양한 작업을 포함할 수 있다.
- [0159] 도 9는 다양한 펩타이드의 면역학적 활동에 대한 정보를 생성하기 위한 프로세스의 흐름도이다. 프로세스(900)의 적어도 일부분은 비제한적 예를 들어 도 1에 기재된 예측 시스템(100)을 사용하여 구현될 수 있다. 예를 들어 프로세스(900)의 적어도 일부분은 비제한적 예를 들어, 도 1 및 3의 머신 러닝 모델(132) 또는 도 4a-4c의 머신 러닝 모델(400)을 사용하여 구현될 수 있다.
- [0160] 단계(902)는 복수의 펩타이드 서열 및 복수의 IPC 서열을 포함하는 서열 데이터를 수신하는 것을 포함한다.
- [0161] 단계(904)는 펩타이드 서열 및 IPC 서열을 사용하여 복수의 펩타이드-IPC 결합체를 생성하는 것을 포함한다. 각각의 펩타이드-IPC 결합체는 고유한 결합체이다.
- [0162] 단계(906)는 각각의 펩타이드-IPC 결합체에 대해, 펩타이드-IPC 결합체에 대응하는 펩타이드 서열을 머신 러닝 모델의 펩타이드 처리 경로에 입력하며 펩타이드-IPC 결합체에 해당하는 IPC 서열을 머신 러닝 모델의 IPC 처리 경로에 입력하는 것을 포함한다.
- [0163] 단계(908)는 각각의 펩타이드-IPC 결합체에 대해 제1 어텐션 블록을 사용하여 펩타이드 서열의 펩타이드 표현을 처리하고 제2 어텐션 블록을 사용하여 IPC 서열의 IPC 표현을 처리하여 변환된 펩타이드 표현 및 변환된 IPC 표현을 각각 생성하는 것을 포함한다.
- [0164] 단계(910)는 각각의 펩타이드-IPC 결합체에 대해 변환된 펩타이드 표현 및 변환된 IPC 표현을 사용하여 복합 표현을 생성하는 것을 포함한다.
- [0165] 단계(912)는, 각각의 펩타이드-IPC 결합체에 대해, 제3 어텐션 블록을 이용해 복합 표현을 처리하여 변환된 복합 표현을 생성하는 것을 포함한다.
- [0166] 단계(914)는 변환된 복합 표현에 기초하여 출력을 생성하는 것을 포함한다. 출력은 펩타이드 서열 중 어느 것이 치료제를 생성하는 데 사용될 수 있는지에 대한 표시자를 제공할 수 있다. 예를 들어, 출력은 어느 펩타이드 서열(그리고 따라서 그 펩타이드 서열을 포함하는 펩타이드)이 MHC와 결합할 높은 가능성, MHC에 의해 제시될 높은 가능성, 펩타이드-MHC 결합에 대한 높은 상호작용 친화도, 및/또는 면역 반응을 트리거하기 위한 면역원성일 높은 가능성을 가지는지에 대한 표시자를 제공할 수 있다.
- [0167] II.C.1. 예시적인 방법론: 펩타이드 및 MHC
- [0168] 도 10은 하나 이상의 실시예에 따르는 펩타이드 및 MHC와 관련된 예측을 생성하도록 머신 러닝 모델을 훈련하고 훈련된 머신 러닝 모델을 이용하기 위한 프로세스의 흐름도이다. 프로세스(1000)는 도 1의 예측 시스템(100)을 사용하여 수행될 수 있다. 예를 들어, 프로세스(1000)는 도 1 및 3의 머신 러닝 모델(132) 또는 도 4a-4c의 머신 러닝 모델(400)을 이용해 구현될 수 있다. 일부 사례에서, 프로세스(1000)의 일부 또는 전부가 사용자 장치 및/또는 검사실과 관련하여 원격인 원격 컴퓨팅 시스템에서 수행될 수 있다. 원격 컴퓨팅 시스템은 클라우드 컴퓨팅 시스템일 수 있다.
- [0169] 단계(1002)는 훈련 펩타이드 서열 데이터, 훈련 MHC 서열 데이터, 및 훈련 면역학적 활동 데이터를 식별하는 훈련 요소로 훈련 데이터 세트에 접근하는 것을 포함한다. 훈련 데이터 세트는 도 1의 훈련 데이터(133)에 대한 구현의 한 예일 수 있다. 훈련 면역학적 활동 데이터는 예를 들어 상호작용 표시자를 포함할 수 있다.
- [0170] 훈련 펩타이드 서열 데이터는 예를 들어 훈련을 위한 하나 이상의 펩타이드 서열(변이-코딩 서열을 포함할 수 있음)을 포함할 수 있다. 펩타이드 서열은 펩타이드(가령, 신생항원) 내에서 정렬된 아미노산 세트를 식별할 수 있다. 펩타이드 서열은 펩타이드의 (예를 들어, 변이를 포함 및/또는 신생항원결정기를 포함하거나 신생항원결정기인) 항원결정기 내의 아미노산을 식별할 수 있다. 일부 실시예에서, 펩타이드 서열은 N-플랭크 서열(예를 들어, 대응하는 펩타이드의 N-말단에서의 아미노산의 사슬을 특징지음) 또는 C-플랭크 서열(예를 들어, 대응하는 펩타이드의 C-말단에 있는 아미노산의 사슬을 특징지음)을 더 포함하는 통합 서열 내에 있다. N-플랭크와 C-플랭크 모두 MHC 분자에 결합하지 않지만 각각은 MHC 분자에 의해 제시되는지 여부에 영향을 미칠 수 있다.
- [0171] 훈련 MHC 서열 데이터는 훈련을 위한 하나 이상의 MHC 서열을 포함할 수 있다. 예를 들어, MHC 서열은 MHC 분자(예를 들어, MHC-I 분자 또는 MHC-II 분자)의 일부 또는 전부 내의 아미노산을 식별할 수 있다. MHC 서열은 MHC 의사서열(예를 들어, 34개의 아미노산을 포함함)을 포함할 수 있다. MHC 서열은 예를 들어 MHC-I에 대한 1, 2,

3, 4, 5 또는 6개의 MHC 대립유전자 또는 MHC-II에 대한 1,2,3,4,5,6,7,8,9,10,11,12개의 MHC 동종이형 내 아미노산을 식별할 수 있다. MHC 서열은 HLA 분자의 일부 또는 전부를 구성하는 아미노산을 식별할 수 있다.

[0172] 훈련 면역학적 활동 데이터는 예를 들어 하나 이상의 펩타이드-MHC 결합체에 대한 하나 이상의 상호작용 표시자를 포함할 수 있다. 예를 들어, 훈련 데이터 세트는 훈련 요소를 포함할 수 있고, 여기서 각각의 훈련 요소는 훈련을 위한 펩타이드 서열 및 MHC 서열뿐만 아니라 대응하는 펩타이드-MHC 결합체에 대한 하나 이상의 상호작용 표시자까지 포함한다. 상호작용 표시자는 표적 상호작용(예를 들어, 펩타이드와 MHC의 결합, MHC에 의한 세포 표면 상의 펩타이드 제시)이 펩타이드와 MHC 사이에 발생하는지 또는 표적 상호작용에 대한 친화도를 나타낼 수 있다.

[0173] 상호작용 표시자는 예를 들어 라벨일 수 있다. 음성 상호작용 라벨은 펩타이드가 MHC 분자에 결합하지 않음 및/또는 제시되지 않음을 나타낼 수 있다. 양성 상호작용 라벨은 펩타이드는 MHC 분자에 결합함 및/또는 제시됨을 나타낼 수 있다. 또한, 상호작용 라벨은 펩타이드가 MHC 분자에 결합할 확률, MHC 분자가 세포 표면에 펩타이드를 제시할 확률, 펩타이드-MHC 결합체에 대한 결합 친화도, 펩타이드와 MHC 분자 간 결합 강도, 펩타이드와 MHC 분자 간 결합 안정도, 펩타이드가 MHC와 결합하는 경향, 또는 MHC와 펩타이드 간 상호작용과 연관된 또 다른 메트릭 또는 특성을 포함할 수 있다.

[0174] 훈련 데이터 세트는 예를 들어 생체외 또는 생체내 실험을 통해 및/또는 의료 기록을 기반으로 생성되었을 수 있다. 훈련 데이터는 이하의 섹션 II.E에서 개시된 하나 이상의 기법에 기초하여 생성되었을 수 있다.

[0175] 훈련 데이터 세트를 액세스하는 것은, 예를 들어, 로컬 또는 원격 저장장치로부터 훈련 데이터 세트를 불러오는 것, 훈련 데이터 세트를 로딩하는 것, 및/또는 하나 이상의 데이터 저장소(가령, 클라우드 데이터 저장소, 서버 시스템, 또는 그 밖의 다른 일부 데이터 소스)로부터의 훈련 데이터 세트의 일부 또는 전부를 요청하는 것(그리고 수신하는 것)을 포함할 수 있다.

[0176] 일부 예에서, 초기 훈련 데이터 세트(예를 들어, 변이-코딩 서열을 포함함)는 우세하게 음성 데이터를 포함할 수 있으며, 여기서, 서열 결합체(예를 들어, 펩타이드-MHC 결합체)의 상대적으로 작은 부분이 실제 표적 상호작용과 연관됨이 발견된다. 훈련 데이터 세트는 음성 훈련 데이터 요소를 포함하도록 설계될 수 있다. 일부 실시예에서, 음성 훈련 데이터 요소는 양성 세트(관찰된 제시에 대응하는) 기원 단백질의 유사-무작위로 선택된 단편 내의 아미노산을 식별하도록 정의될 수 있다. 예를 들어, 음성 훈련 데이터 요소는 양성 세트에 기초하여 시뮬레이션될 수 있다. 단편은 지정 범위(예를 들어, 균일한 확률을 사용하여, MHC-I에 대해 8 내지 14개 아미노산 및 MHC-II에 대해 8-30개 아미노산) 내의 길이를 갖도록 선택될 수 있다. N-말단 및 C-말단 플랭킹 서열은 음성 훈련 데이터 요소 내에 유지되어, 잠재적으로 최대 길이(가령, 10개 아미노산)를 부과할 수 있다. 양성 펩타이드와 중복되는 임의의 펩타이드 단편(예를 들어, 적어도 9-mer)은 음성 훈련 데이터로부터 폐기될 수 있다.

[0177] 다양한 실시예에서, 음성 훈련 데이터 요소는 양성 데이터 요소에 기초하여 시뮬레이션된다. 또한, 훈련 기간의 에포크마다 음성 훈련 데이터 요소의 상이한 세트가 사용되도록 훈련 데이터가 선택된다. 예를 들어, 각각의 에포크에 대해, 음성 펩타이드 서열의 상이한 "음성 서브세트"는 펩타이드 서열의 양성 세트에 기초하여 식별된 이용 가능한 음성 펩타이드 서열의 전체 공간으로부터 선택될 수 있다. 각각의 에포크에 대해 선택된 음성 서브세트는 총 에포크 수 동안 음성 서브세트 중 어느 것에서도 음성 펩타이드 서열이 반복되지 않는다는 점에서 고유할 수 있다. 따라서, 훈련 기간의 각각의 에포크에 대해 사용되는 훈련 데이터는 펩타이드 서열의 동일한 양성 세트를 포함하지만 완전히 상이한 음성 펩타이드 서열 세트를 포함한다. "음성 세트 전환"이라고 할 수 있는 이 기법은 훈련에 전반적인 강건성을 제공할 수 있으며 머신 러닝 모델에 의한 위음성(가령, 위음성 표시자/예측)의 감소된 수를 보장하거나 어떠한 위음성도 두 번 이상 반복되지 않음을 보장할 수 있다. 또한, 이 기법에 의해, 머신 러닝 모델은 양성 펩타이드 서열의 수에 훈련 기간의 에포크 수를 곱한 것과 동일한 음성 펩타이드 서열의 총 수에 대해 훈련될 수 있다.

[0178] 단계(1004)는 훈련 데이터 세트를 사용하여 머신 러닝 모델을 훈련하는 것을 포함한다. 머신 러닝 모델은 예를 들어, 도 1 및 3의 머신 러닝 모델(132)일 수 있거나 머신 러닝 모델은 예를 들어 도 4a-4c의 머신 러닝 모델(400)일 수 있다.

[0179] 머신 러닝 모델(132)은 정적 또는 동적 학습물을 사용하여 훈련될 수 있다. 예를 들어 학습물 어닐링을 사용하여 동적 학습물이 생성될 수 있다. 훈련은 예를 들어 분류 손실 함수 및/또는 회귀 손실 함수를 사용하여 수행될 수 있다. 손실 함수는 예를 들어 평균 제곱 오차, 중앙값 제곱 오차, 평균 절대 오차, 중앙값 절대 오차, 엔트로피-기반 오차, 교차 엔트로피 오차, 및/또는 이진 교차 엔트로피 오차에 기초할 수 있다. 검증 데이터(예를

들어, 머신 러닝 모델(132)을 훈련하는 데 사용되는 훈련 데이터 세트의 분리된 서브세트는 머신 러닝 모델 (132)이 훈련되는 대로 그 성능을 평가하는 데 사용될 수 있다. 표적 성능이 획득하는 경우 및/또는 때 및/또는 최대 수의 훈련 반복구간이 완료될 때 훈련이 종료될 수 있다.

[0180] 단계(1006)는 돌연변이 펩타이드의 세트에 대응하는 변이-코딩 서열의 대상체 특이적 세트를 액세스하는 것을 포함한다. 전술한 바와 같이, 변이-코딩 서열은 펩타이드 서열의 한 예이다. 변이-코딩 서열의 대상체 특이적 세트는 돌연변이 펩타이드의 세트에 대응할 수 있으며, 이로써 변이-코딩 서열의 대상체 특이적 세트의 각각이 돌연변이 펩타이드의 세트의 대응하는 돌연변이 펩타이드 내 아미노산을 식별할 수 있거나 및/또는 변이-코딩 서열의 대상체 특이적 세트의 각각이 돌연변이 내 하나 이상의 아미노산을 식별할 수 있다. 변이-코딩 서열의 대상체 특이적 세트의 각각은 특정 대상체(예를 들어, 인간 대상체)와 연관될 수 있다. 특정 대상체는 특정 의학적 상태(가령, 암)로 진단 받았을 수 있거나, 및/또는 이와 연관된 증상을 경험했거나 및/또는 검사 결과를 받았을 수 있다. 예를 들어, 변이-코딩 서열의 대상체 특이적 세트는 종양으로부터의 샘플을 처리함으로써 식별 되었을 수 있다. 샘플은 예를 들어 도 1의 샘플(112)의 세트일 수 있거나 그 내에 포함될 수 있다.

[0181] 변이-코딩 서열의 대상체 특이적 세트는 본 명세서(가령, 섹션 II.D)에 개시된 기술을 이용해 식별될 수 있다. 예를 들어, 변이-코딩 시퀀싱의 대상체 특이적 세트는 질병 샘플에서 펩타이드를 식별하기 위한 시퀀싱 기술을 수행하고 식별된 펩타이드를 건강한 샘플 또는 참조 데이터베이스에서 검출된 것과 비교하여 고유한 서열을 식별함으로써 식별되었을 수 있다. 일부 구현예에서, 고유한 서열이 핵산 서열인 경우, 각각의 고유한 핵산 서열은 아미노산 서열로 변환될 수 있다.

[0182] 변이-코딩 서열의 대상체 특징적 세트의 각각은 (신생항원의 신생항원결정기 내 아미노산일 수 있는) 펩타이드 내의 아미노산을 식별할 수 있다. 어떤 경우에는, 변이-코딩 서열의 대상체 특이적 세트의 하나 이상 또는 모두의 각각은 펩타이드의 N-플랭크에 있는 서열 및/또는 펩타이드의 C-플랭크에 있는 서열을 추가로 포함하는 대응하는 통합 서열의 일부일 수 있다.

[0183] 변이-코딩 서열의 대상체 특이적 세트를 액세스하는 것은 예를 들어 로컬 또는 원격 저장소로부터 변이-코딩 서열의 대상체 특이적 세트를 불러오는 것 및/또는 또 다른 장치로부터 변이-코딩 서열의 대상체 특이적 세트를 요청하는 것을 포함할 수 있다. 변이-코딩 서열의 대상체 특이적 세트를 액세스하는 것은 변이-코딩 서열의 대상체 특이적 세트를 결정하는 것을 포함하거나 및/또는 상기 결정하는 것과 함께 수행될 수 있다.

[0184] 변이-코딩 서열의 대상체 특이적 세트는 대상체의 질병 샘플 내의 펩타이드 서열을 식별하고 어떤 펩타이드 서열이 참조, 건강한 샘플 및/또는 야생형 서열 세트 내에서 표현되지 않는지를 결정함으로써 획득되었을 수 있다. 비교를 위해 건강한 샘플이 사용되는 경우, 건강한 샘플은 대상체로부터 수집되었을 수 있다(그러나 반드시 그럴 필요는 없다).

[0185] 단계(1008)는 MHC에 대응하는 MHC 서열을 액세스하는 것을 포함한다. MHC 서열은, 예를 들어 대상체로부터 수집된 샘플 내의 MHC(예를 들어, MHC 분자)의 의사서열을 포함할 수 있다. 일부 예에서, MHC 서열 및 변이-코딩 서열의 대상체 특이적 세트는 대상체로부터의 동일한 샘플 또는 대상체로부터의 다수의 샘플(예를 들어, 질병 샘플 및 건강한 샘플)로부터 식별된다. 일부 경우에, MHC 서열 및 변이-코딩 서열의 대상체 특정 세트는 대상체 및 하나 이상의 다른 대상체로부터의 샘플로부터 식별된다. 따라서 어떤 경우에는 MHC 서열은 대상체 특이적일 수 있다. MHC 서열은 예를 들어 시퀀싱 및/또는 질량 분석 기술을 사용하여 결정될 수 있거나 결정되었을 수 있다.

[0186] MHC 서열을 액세스하는 것은 예를 들어 로컬 또는 원격 저장소로부터 MHC 서열을 불러오는 것 및/또는 다른 장치로부터 대상체 특이적 MHC 서열을 요청하는 것을 포함할 수 있다. MHC 서열을 액세스하는 것은 MHC 서열을 결정하는 것을 포함하거나 및/또는 상기 결정하는 것과 함께 수행될 수 있다.

[0187] 단계(1010)는 예를 들어 출력을 생성하기 위해 훈련된 머신 러닝 모델을 이용해 대상체 특이적 변이-코딩 서열의 세트 및 MHC 서열을 처리하는 것을 포함한다. 단계(1010)는 대상체 특이적 변이-코딩 서열의 세트의 대상체 특이적 변이-코딩 서열과 MHC 서열의 고유의 결합체(가령, 변이-코딩-MHC 결합체 또는 펩타이드-MHC 결합체)를 처리하여 출력을 생성하는 것을 포함할 수 있다.

[0188] 머신 러닝 모델에 의해 생성된 출력은 머신 러닝 모델을 훈련시키는 데 사용되는 훈련 면역학적 활동 데이터에 포함된 것과 동일하거나 유사한 유형의 데이터를 포함할 수 있다. 각각의 고유한 결합체에 대해 머신 러닝 모델은 상호작용 예측의 세트 또는 상호작용 선호도 예측의 세트 중 적어도 하나를 포함하는 출력을 생성한다.

[0189] 상호작용 예측의 세트 내 상호작용 예측은 변이-코딩 서열을 포함하는 돌연변이 펩타이드와 MHC 서열을 포함하

는 MHC 사이의 표적 상호작용이 일어날 것인지 여부에 대한 예측을 포함한다. 예를 들어, 상호작용 예측은 대상체 특이적 변이-코딩 서열에 의해 표시되는 아미노산 구조를 갖는 돌연변이 펩타이드가 MHC 서열로 표시되는 아미노산 구조를 갖는 MHC 분자에 의해 제시 및/또는 결합될 것인지 여부에 대한 이진 또는 범주형 예측을 포함할 수 있다. 상호작용 친화도 예측의 세트 내 상호작용 친화도 예측은 표적 상호작용에 대한 친화도에 대한 예측을 포함한다. 이 친화도는 예를 들어 표적 상호작용의 강도, 경향 및/또는 안정성에 기초하여 정의될 수 있다. 예를 들어, 상호작용 친화도 예측은 대상체 특이적 변이-코딩 서열 내에서 식별된 아미노산을 포함하는 돌연변이 펩타이드 및 MHC 서열 내에서 식별된 아미노산을 포함하는 MHC 분자와 연관된 예측된 실수 결합 친화도를 포함할 수 있다.

[0190] 단계(1012)는 머신 러닝 모델의 출력에 기초하여 리포트를 생성하는 것을 포함한다. 리포트는 예를 들어 도 1 및 3의 리포트(144)로서 구현될 수 있다. 리포트는 출력이거나 출력을 포함할 수 있다. 경우에 따라, 리포트는 변환되거나 필터링된 출력 버전일 수 있다.

[0191] 하나 이상의 실시예에서, 변이-코딩 서열의 대상체 특이적 세트가 리포트에 포함시키기 위한 정보를 생성하기 위해 출력에 기초하여 필터링, 순위화 및/또는 다른 방식으로 처리된다. 예를 들어, 변이-코딩 서열의 대상체 특이적 세트는 필터링되어 예측된 상호작용 친화도(예를 들어, 결합 친화도)가 지정된 친화도 임계값 미만인 서열 및/또는 표적 상호작용(예를 들어, MHC 분자에 의한 제시 또는 결합)이 발생하지 않거나 발생할 가능성이 없는 서열을 제외할 수 있다. 어떤 경우에는 변이-코딩 서열의 대상체 특이적 세트의 지정된 수 및/또는 부분을 식별하기 위해 필터링이 수행된다. 예를 들어, 돌연변이 펩타이드가 MHC 분자에 결합 및/또는 제시될지 여부에 대해 상대적으로 높은 예측 확률과 연관된 10, 20, 40, 60, 80, 100, 500 또는 1,000개의 변이-코딩 서열을 (가령, 변이-코딩 서열의 대상체 특이적 세트에서 선택되지 않은 변이-코딩 서열에 대해) 식별하기 위해 필터링이 수행될 수 있다.

[0192] 리포트는 (예를 들어, 세트로부터 필터링되지 않은) 하나 이상의 변이-코딩 서열 및/또는 (예를 들어, 선택된 변이-코딩 서열과 연관된) 하나 이상의 돌연변이 펩타이드를 확인할 수 있다. 돌연변이 펩타이드는 예를 들어 이름, 서열 및/또는 대응하는 야생형 서열 및 변이-코딩 서열에 나타난 변이체 모두를 식별함으로써 식별될 수 있다.

[0193] 리포트는 일부 구현예에서 하나 이상의 변이-코딩 서열 또는 하나 이상의 돌연변이 펩타이드와 연관된 하나 이상의 예측을 식별할 수 있다. 리포트는 대상체의 이름을 포함할 수 있다. 리포트는 예를 들어 로컬하게(예를 들어, 사용자 장치의 디스플레이 시스템에 디스플레이하기 위해, 사용자 장치에 대한 알림으로 전송되는 등) 및/또는 다른 장치로 전송될 수 있다(예를 들어, 클라우드 컴퓨팅 시스템으로 전송, 클라우드 저장소로 전송, 의료 전문가 또는 실험실 전문가와 연관된 사용자 장치로 전송, 이메일로서 전송 등).

[0194] 도 11은 하나 이상의 실시예에 따라 훈련 데이터의 표를 포함하는 예시이다. 표(1100)는 훈련 데이터(1102)(예를 들어, 훈련 데이터 세트)를 포함한다. 훈련 데이터(1102)는 도 1의 훈련 데이터(133)의 일부분의 한 예일 수 있다. 훈련 데이터(1102)는 훈련 데이터 세트, 가령, 도 11의 단계(1002)에서 기재된 훈련 데이터세트의 일부분의 한 예일 수 있다.

[0195] 훈련 데이터(1102)는 대립유전자 식별자(1106), 훈련 N-플랭크 서열(1108), 훈련 펩타이드 서열(1110), 훈련 C-플랭크 서열(1112), 훈련 MHC 서열(1114)(예를 들어, MHC 의사서열), 결합 친화도(1116) 및 제시 표시자(1118)를 포함한다. 결합 친화도(1116)는 훈련 펩타이드 서열(1110)에 의해 특징지어지는 펩타이드 및 훈련 MHC 서열(1114)에 의해 특징지어지는 각각의 MHC의 결합에 대해 검출된(예를 들어, 관찰된) 결합 친화도를 나타낸다. 제시 표시자(1118)는 MHC에 의한 펩타이드의 결합 또는 제시가 검출(또는 관찰)되었는지 여부를 나타낸다.

[0196] 도 12는 하나 이상의 실시예에 따르는 신생항원 후보 및 대응하는 가능한 신생항원결정기 후보의 예시이다. 프로세스, 가령, 프로세스(1000)가 구현될 때, 돌연변이 펩타이드는 신생항원일 수 있다.

[0197] 신생항원 후보(1200)인 상대적으로 긴 돌연변이 펩타이드의 경우, 동일한 돌연변이 또는 변이체를 모두 포함하는 다중 항원결정기(신생항원결정기라고 지칭됨)가 MHC 분자에 의해 제시될 수 있다. 따라서, 신생항원 후보의 면역원성은 각각의 신생항원결정기 후보(1202)에 대해 생성된 예측에 기초하여 예측될 수 있다.

[0198] 면역원성은 예를 들어 주어진 신생항원으로부터 나타날 수 있는 모든 가능한 신생항원결정기의 목록을 생성하고 목록 내 (길이에서 최대 10개의 아미노산까지, 항원결정기의 N-말단의 업스트림 및 C-말단의 다운스트림의 나머지 아미노산을 이루는 플랭크를 갖는) 신생항원결정기 후보의 일부 또는 전부 각각에 대한 예측을 생성함으로써 예측될 수 있다. 이러한 제시 예측으로부터 MHC 후보(1204)와 관련하여 가장 큰 제시 가능성을 갖는 신생항원결

정기 후보가 전체 신생항원을 나타내도록 선택된다. 대안으로, 다중 후보 신생항원-MHC 쌍의 요약된 표현은 신생항원을 나타내는 요약된 점수를 얻기 위해 사용될 수 있다. 이러한 요약은 모든 후보 신생항원결정기-MHC 쌍을 고려하거나 MHC별 최상의 신생항원결정기를 고려한 다음 모든 MHC 분자에 걸쳐 요약함으로써 수행될 수 있다. 요약은 예를 들어 각각의 후보 신생항원결정기-HLA 쌍의 제시 또는 결합 친화도 점수의 산술 평균 또는 조화 평균을 취하는 것을 포함하는 여러 수학적 함수에 의해 수행될 수 있다.

[0199] 도 12가 신생항원 및 신생항원결정기와 관련하여 기재되어 있지만, 유사한 기법이 돌연변이 또는 변이체를 함유하고 다수의 가능한 항원결정기 후보를 갖는 비교적 긴 다른 유형의 돌연변이 펩타이드에 대해 사용될 수 있다. 일부 구현예에서, 이 기법은 항체 약물 서열과 함께 사용될 수 있다.

[0200] II.C.2. 예시적인 방법론: 펩타이드 및 TCR

[0201] 도 13은 하나 이상의 실시예에 따르는 펩타이드 및 TCR와 관련된 예측을 생성하도록 머신 러닝 모델을 훈련하고 훈련된 머신 러닝 모델을 이용하기 위한 프로세스의 흐름도이다. 프로세스(1300)는 도 1의 예측 시스템(130)을 사용하여 수행될 수 있다. 예를 들어, 프로세스(1300)는 도 1 및 3의 머신 러닝 모델(132) 또는 도 4a-4c의 머신 러닝 모델(400)을 이용해 구현될 수 있다. 일부 사례에서, 프로세스(1300)의 일부 또는 전부가 사용자 장치 및/또는 검사실과 관련하여 원격인 원격 컴퓨팅 시스템에서 수행될 수 있다. 원격 컴퓨팅 시스템은 클라우드 컴퓨팅 시스템일 수 있다. 단계(1302-1312)는 도 10의 단계(1002-1012)와 유사한 방식으로 구현될 수 있지만, TCR과 관련하여 구현될 수 있다.

[0202] 단계(1302)는 훈련 펩타이드 서열 데이터, 훈련 TCR 서열 데이터, 및 훈련 면역학적 활동 데이터를 식별하는 훈련 요소로 훈련 데이터 세트에 접근하는 것을 포함한다. 훈련 TCR 서열 데이터는 훈련을 위한 하나 이상의 TCR 서열을 포함할 수 있다. 예를 들어, TCR 서열은 TCR 분자의 일부 또는 전부 내의 아미노산을 식별할 수 있다.

[0203] 훈련 면역학적 활동 데이터는 예를 들어 하나 이상의 펩타이드-TCR 결합체 및/또는 하나 이상의 면역원성 예측에 대한 하나 이상의 상호작용 표시자를 포함할 수 있다. 면역원성 예측은 TCR에 대한 펩타이드의 면역원성을 예측할 수 있다. 예를 들어, 훈련 데이터 세트는 변이-코딩 서열에 의해 식별되는 아미노산을 갖는 돌연변이 펩타이드가 면역학적 반응을 트리거했는지 여부(예를 들어, 돌연변이 펩타이드가 면역원성인지 여부)를 나타내는 상호작용 라벨을 포함할 수 있다. 면역원성은 돌연변이 펩타이드가 T 세포 수용체(가령, CD8+ 세포독성 T 림프구 또는 CD4+ 보조 T 세포의 수용체)를 활성화 및/또는 면역학적 반응을 트리거했음을 나타낼 수 있다.

[0204] 훈련 데이터 세트는 예를 들어 샘플(예를 들어, 하나 이상의 수지상 세포)에서 다양한 돌연변이 펩타이드를 발현 및/또는 면역화 및/또는 백신에 의해 다양한 돌연변이 펩타이드를 (예를 들어, 샘플에 또는 이후에 샘플이 수집되는 대상체에) 도입함으로써 생성되었을 수 있다. 돌연변이 펩타이드는 개별적으로(예를 들어, 단일 돌연변이 펩타이드에 대한 각각의 실험에 집중) 또는 그룹으로 발현되거나 도입되었을 수 있다.

[0205] 예를 들어, 중앙 침윤 세포를 분석함으로써 면역원성이 시험되었을 수 있다. 예를 들어 돌연변이 펩타이드의 항원결정기가 (가령, 임계값을 초과하는 수량) 검출되는 경우, 인터페론 감마(IFN-) 또는 T 세포 면역글로불린 유신-3(TIM-3)의 측정된 레벨이 대응하는 임계값을 초과한 경우, 세포독성 T 세포(가령, 일반적으로 또는 돌연변이 펩타이드에 대응하는 항원결정기를 표시하는 세포독성 T 세포)의 검출된 양이 대응하는 임계값을 초과한 경우, 및/또는 적어도 임계 정도의 아포토시스가 관찰되는 경우, 돌연변이 펩타이드가 면역학적 반응을 트리거(따라서 면역원성임)했다고 결정되었을 수 있다. 다른 예로서, 돌연변이 펩타이드는 샘플(예를 들어, 하나 이상의 수지상 세포)에서 발현되었을 수 있다. 예를 들어, 제시된 펩타이드가 차후에 T 세포에 의해 인식되는 것으로 결정된다면, 돌연변이 펩타이드가 면역학적 반응을 트리거(따라서 면역원성임)한 것으로 결정되었을 수 있다. 일부 실시예는 (예를 들어, 본 명세서에 개시된 하나 이상의 실험 및/또는 분석을 수행함으로써) 훈련 데이터 세트의 적어도 일부분을 수집 및/또는 결정하는 것을 포함한다는 것을 이해할 것이다.

[0206] 훈련 데이터 세트를 액세스하는 것은, 예를 들어, 로컬 또는 원격 저장장치로부터 훈련 데이터 세트를 불러오는 것, 훈련 데이터 세트를 로딩하는 것, 및/또는 하나 이상의 데이터 저장소(가령, 클라우드 데이터 저장소, 서버 시스템, 또는 그 밖의 다른 일부 데이터 소스)로부터의 훈련 데이터 세트의 일부 또는 전부를 요청하는 것(그리고 수신하는 것)을 포함할 수 있다.

[0207] 단계(1304)는 훈련 데이터 세트를 사용하여 머신 러닝 모델을 훈련하는 것을 포함한다. 머신 러닝 모델은 예를 들어, 도 1 및 3의 머신 러닝 모델(132)일 수 있거나 머신 러닝 모델은 예를 들어 도 4a-4c의 머신 러닝 모델(400)일 수 있다.

[0208] 단계(1306)는 돌연변이 펩타이드의 세트에 대응하는 변이-코딩 서열의 대상체 특이적 세트를 액세스하는 것을

포함한다.

- [0209] 단계(1308)는 TCR에 대응하는 TCR 서열을 액세스하는 것을 포함한다. 일부 예에서, TCR 서열 및 변이-코딩 서열의 대상체 특이적 세트는 대상체로부터의 동일한 샘플 또는 대상체로부터의 다수의 샘플(예를 들어, 질병 샘플 및 건강한 샘플)로부터 식별된다. 일부 경우에, TCR 서열 및 변이-코딩 서열의 대상체 특정 세트는 대상체 및 하나 이상의 다른 대상체로부터의 샘플로부터 식별된다. 따라서 어떤 경우에는 TCR 서열은 대상체 특이적일 수 있다. TCR 서열은 예를 들어 시퀀싱 및/또는 질량 분석 기술을 사용하여 결정될 수 있거나 결정되었을 수 있다.
- [0210] TCR 서열을 액세스하는 것은 예를 들어 로컬 또는 원격 저장소로부터 TCR 서열을 불러오는 것 및/또는 다른 장치로부터 대상체 특이적 TCR 서열을 요청하는 것을 포함할 수 있다. TCR 서열을 액세스하는 것은 TCR 서열을 결정하는 것을 포함하거나 및/또는 상기 결정하는 것과 함께 수행될 수 있다.
- [0211] 단계(1310)는 예를 들어 출력을 생성하기 위해 훈련된 머신 러닝 모델을 이용해 대상체 특이적 변이-코딩 서열의 세트 및 TCR 서열을 처리하는 것을 포함한다. 단계(1310)는 대상체 특이적 변이-코딩 서열의 세트의 대상체 특이적 변이-코딩 서열과 TCR 서열의 고유의 결합체(가령, 변이-코딩-TCR 결합체 또는 펩타이드-TCR 결합체)를 처리하여 출력을 생성하는 것을 포함할 수 있다.
- [0212] 머신 러닝 모델에 의해 생성된 출력은 머신 러닝 모델을 훈련시키는 데 사용되는 훈련 면역학적 활동 데이터에 포함된 것과 동일하거나 유사한 유형의 데이터를 포함할 수 있다. 각각의 고유한 결합체에 대해 머신 러닝 모델은 면역원성 예측의 세트를 포함하는 출력을 생성한다. 면역원성 예측의 세트의 면역원성 예측은 돌연변이 펩타이드가 면역학적 반응을 트리거했는지(따라서 면역원성인지) 여부를 나타낼 수 있다. 어떤 경우에는 면역원성 예측이 면역원성의 정도를 나타낸다(가령, 낮음, 중간, 높음, 매우 높음 등).
- [0213] 단계(1312)는 머신 러닝 모델의 출력에 기초하여 리포트를 생성하는 것을 포함한다. 리포트는 예를 들어 도 1 및 3의 리포트(144)로서 구현될 수 있다. 단계(1312)는 도 10의 단계(1012)와 유사한 방식으로 구현될 수 있다.
- [0214] II.C.3. 예시적인 방법론: 머신 러닝 모델을 사용한 훈련 및 예측에 대한 추가 고려 사항
- [0215] 따라서, 본 명세서에 기재된 실시예는 돌연변이 펩타이드일 수 있는 펩타이드와 연관된 면역학적 활동에 대한 예측을 생성하기 위해 사용될 수 있는 머신 러닝 모델을 제공한다. 돌연변이 펩타이드를 특징짓는 펩타이드 서열 — 가령, 변이-코딩 서열 — 은 IPC를 특징짓는 IPC 서열과 함께 머신 러닝 모델에 의해 분석되어 펩타이드와 IPC 간 하나 이상의 표적 상호작용(관심 상호작용) 및/또는 펩타이드가 면역 반응을 유발할 수 있는 능력에 대한 하나 이상의 예측을 생성할 수 있다. 머신 러닝 모델에 의해 생성된 출력은 따라서 하나 이상의 표적 상호작용 및/또는 펩타이드의 면역원성에 대한 정보를 제공하는 하나 이상의 결과를 포함할 수 있다.
- [0216] 일부 실시예에서, 하나 이상의 변이-코딩 서열은 본 명세서에 기재된 하나 이상의 머신 러닝 모델로부터의 결과에 기초하여 변이-코딩 서열의 대상체 특이적 세트의 세트로부터 선택될 수 있다. 입력 데이터는 MHC 서열의 표현 및 돌연변이 펩타이드에 대응하는 변이-코딩 서열을 포함할 수 있다. 머신 러닝 모델은 MHC 분자에 의해 어떤 펩타이드가 제시되는지를 나타내는 결합 친화도 데이터 및 질량 분석 용출 데이터를 사용하여 훈련될 수 있다. 결합-친화도 데이터는 정성적 데이터(예를 들어, ELISA, 풀-다운 어세이 및/또는 겔 이동 어세이, 형광 공명 에너지 전달 어세이 및 질량분석법 어세이) 또는 정량적 데이터(예를 들어, 바이오센서 기반 방법, 가령, 표면 플라즈몬 공명(Surface Plasmon Resonance), 등온 적정 비색계(Isothermal Titration Colorimetry), 바이오레이어(BioLayer Interferometry) 간섭계 또는 마이크로스케일 열영동(MicroScale Thermophoresis)을 이용한 정량적 데이터)를 포함할 수 있다. 일부 경우, 결합 친화도 데이터는 경쟁적 결합 어세이로부터의 데이터, 면역항원결정기 데이터베이스로부터의 데이터 및/또는 면역 항원결정기 데이터베이스에 있는 유형의 데이터를 포함할 수 있다. 용리 데이터는 펩타이드-MHC 면역침전법, 그 후 용리 및 질량 분광법에 의한 제시된 MHC 리간드의 검출을 사용할 수 집될 수 있다. 훈련 데이터는 "양성" 사례(질량 분석 결과가 펩타이드가 MHC 분자에 의해 제시되었음을 나타냄) 및 "음성" 사례(예를 들어, 양성 사례와 동일하지만 질량분석법 평가에서 검출되지 않았던 단백질로부터 시뮬레이션된 길이-매칭된 n-mers(nmers)에 대응)를 포함했다.
- [0217] 일부 경우, 훈련 데이터의 양성 사례의 수량은 훈련 데이터의 음성 사례의 수량과 같다. 일부 경우, 양성 사례의 수량은 음성 사례의 수량보다 적거나 많다. 훈련 데이터 내 음성 사례 중 하나 이상 또는 전부 각각은 훈련 데이터 내 양성 사례에 길이-매칭될 수 있다. 경우에 따라, 훈련 데이터의 모든 서열 길이가 동일하다.
- [0218] 서열의 일부 또는 전부는 예를 들어 데이터 암호화를 사용하여 표현될 수 있다. 암호화는 알려진 및/또는 정적 규칙 또는 기술에 따라 및/또는 훈련된 망을 사용하여 수행될 수 있다. 예를 들어, 암호화는 각각의 암호화된

서열이 서열의 각각의 위치에 대해 그리고 (예를 들어, 21개의) 아미노산의 세트의 각각에 대해 특정 아미노산이 위치에 존재하는지 여부를 나타내도록 원-핫 암호화를 포함할 수 있다. 대안으로, 진화-동기 암호화, 가령, BLOSUM, 또는 학습된 암호화는 서열 내 아미노산을 표현하는 데 사용될 수 있다. 암호화는 위치 암호화(예를 들어, 학습 또는 고정 암호화)을 포함할 수 있다.

[0219] 일부 경우, 머신 러닝 모델은 서열 처리에 사용되는 하나 이상의 신경망을 포함한다. 신경망(들)은 예를 들어 인코더 신경망 및/또는 변환기 망의 일부 또는 전부를 추가로 또는 대안으로 포함할 수 있다.

[0220] 머신 러닝 모델은 어텐션 기반, 콘볼루션 층이 부재 및/또는 반복 층이 부재인 하나 이상의 신경망을 포함하는 어텐션-기반 머신 러닝 모델을 포함할 수 있다. 어텐션-기반 머신 러닝 모델은 어텐션-기반이 아님, 하나 이상의 콘볼루션 층을 포함, 및/또는 하나 이상의 반복 층을 포함하는 하나 이상의 다른 신경망을 더 포함할 수 있다(그러나 반드시 그럴 필요는 없다).

[0221] 어텐션-기반 망은 쿼리 가중치의 세트, 키 가중치의 세트, 및 값 가중치의 세트를 이용해 주어진 아미노산 표현에 대해, 주어진 아미노산 표현을 처리할 때 하나 이상의 다른 아미노산 표현 각각이 "어텐션"을 기울이는 정도를 결정할 수 있다. 셀프-어텐션 층은 예를 들어 인코더 또는 디코더가 인코더 또는 디코더의 이전 층에 있는 모든 위치에 어텐션을 기울일 수 있도록 동일한 층의 키, 값 및 쿼리를 사용할 수 있다.

[0222] 주어진 돌연변이 펩타이드가 특정 MHC 분자에 결합할 것인지 및/또는 이에 의해 제시될 것인지 여부를 예측할 때, 하나 이상의 변환기 인코더는 변이-코딩 서열 및/또는 MHC 서열의 상이한 부분 또는 전부의 표현을 개별적으로 처리할 수 있다. 각각의 변환기 인코더는 셀프-어텐션 층 및 피드-포워드 층을 포함할 수 있다. 각각의 어텐션 층은 예를 들어 위치 및/또는 비위치 임베딩을 수행하도록 구성된 하나 이상의 임베딩 구성요소를 더 포함할 수 있다. 일부 경우에, 돌연변이 펩타이드의 N-플랭크 영역, 돌연변이 펩타이드의 항원결정기 영역, 돌연변이 펩타이드의 C-플랭크 영역, 및 MHC 분자의 각각의 서열은 변환기 인코더의 상이한 반복구간을 개별적으로 처리한다. 서열의 암호화된 표현은 서열의 각각의 아미노산에 대해 아미노산을 나타내는 특징 벡터를 포함할 수 있다. 그런 다음 서열의 암호화된 표현들이 연결되어 변환기 인코더의 또 다른 반복구간에 공급될 수 있다. 따라서 연결은 변이-코딩 서열의 일부 또는 전부 내 각각의 아미노산 및 MHC 서열의 전부 또는 일부에 대한 특징 벡터를 포함할 수 있다.

[0223] 하나 이상의 추가 특징 벡터가 연결에 포함될 수 있다. 각각의 추가 특징은 예를 들어 특징 벡터에 대해 임의 또는 의사 랜덤 값을 할당 받을 수 있다. 연결된 표현(예를 들어, 추가적인 특징 벡터(들)을 포함함)은 암호화된 연결된 표현을 생성하기 위해 추가적인 변환기 인코더에 의해 처리될 수 있다. 서열 결합체의 이 암호화된 표현은 드롭아웃 및/또는 배치 정규화가 적용될 수 있는 피드포워드 망(예를 들어, 완전 연결 신경망)에 의해 처리될 수 있다. 일부 경우에, 추가적인 특징 벡터(들)의 암호화된 표현(들)은 선택적으로 피드포워드 망으로 전달된다(예를 들어, MHC 분자 및/또는 돌연변이 펩타이드의 개별 아미노산에 대응하는 특징 벡터는 그렇지 않음). 예를 들어, MHC 분자의 부분서열이 x_1 아미노산을 포함하고, 돌연변이 펩타이드의 부분서열(예를 들어, 하나 이상의 플랭크)이 x_2 아미노산을 포함하고, 특징 변환이 각 아미노산을 나타내는 y 특징 값을 식별한다고 가정한다. 따라서 1개의 추가 특징 벡터를 포함하는 연결된 표현은 $[(x_1+x_2+1), y]$ 의 크기를 가질 수 있다. 하나의 특징 벡터가 피드포워드 망에 의해 처리되도록 선택되는 경우 피드포워드 망에 입력되는 입력은 $[1, y]$ 의 크기를 가질 수 있다. 추가 요소 접근 방식을 사용할 때의 이점은 모델이 가변 길이의 서열을 처리할 수 있다는 것이다.

[0224] 피드포워드 망에 의해 생성된 결과는 돌연변이 펩타이드와 MHC 분자(예를 들어, 대상체의 MHC 분자) 사이의 결합 친화도 및/또는 돌연변이 펩타이드가 MHC 분자에 의해 제시될 것인지 여부에 대한 예측에 대응할 수 있다. 결합-친화도 예측은 예를 들어 수치적(예를 들어, 돌연변이 펩타이드가 MHC 분자에 결합할 예측된 확률, 예측된 결합 강도 및/또는 예측된 결합 안정성에 대응), 범주(예를 들어, 돌연변이 펩타이드와 MHC 분자 간 결합 안정성이 없거나, 낮거나, 높음) 또는 이진(가령, 돌연변이 펩타이드가 MHC 분자에 결합하는지 여부를 예측)일 수 있다.

[0225] 돌연변이 펩타이드와 연관되어 생성된 제시 예측은, 예를 들어, 숫자(가령, 대상체의 MHC 분자가 돌연변이 펩타이드를 제시하는 대상체 내 종양 세포의 세포 표면 또는 예측 부분에 돌연변이 펩타이드를 제시할 예측 확률에 대응), 범주(가령, 대상체의 MHC 분자에 의한 돌연변이 펩타이드의 제시 없음, 드물, 또는 빈번을 예측), 또는 이진(가령, 대상체 내 MHC 분자에 의해 돌연변이 펩타이드가 발현되는지 여부를 예측)일 수 있다. 제시 예측은 정규화되거나 조건부 예측을 나타낼 수 있다(그러나 반드시 그럴 필요는 없다). 예를 들어, 제시 예측은 돌연변이

이 펩타이드가 MHC 분자에 안정적으로 결합한 경우 대상체의 MHC 분자가 돌연변이 펩타이드를 제시하는지 여부에 대한 예측에 대응할 수 있다.

[0226] 일부 예에서, 머신 러닝 모델은 돌연변이 펩타이드와 MHC-I 분자 사이의 하나 이상의 잠재적인 상호작용에 대응하는 예측을 생성한다. 예를 들어, 머신 러닝 모델은 MHC-I 분자와 돌연변이 펩타이드의 결합 친화도 및/또는 MHC-I 분자가 돌연변이 펩타이드를 제시할지 여부를 예측할 수 있다. 머신 러닝 모델은 MHC-I 분자의 서열 또는 부분서열 및 돌연변이 펩타이드와 연관된 변이-코딩 서열을 입력으로서 수신할 수 있고 (예를 들어, 하나 이상의 셀프-어텐션 층을 이용해) 처리할 수 있다.

[0227] 일부 예에서, 머신 러닝 모델은 돌연변이 펩타이드와 MHC-II 분자 사이의 하나 이상의 잠재적인 상호작용에 대응하는 예측을 생성한다. 예를 들어, 머신 러닝 모델은 MHC-II 분자와 돌연변이 펩타이드의 결합 친화도 및/또는 MHC-II 분자가 돌연변이 펩타이드를 제시할지 여부를 예측할 수 있다. 머신 러닝 모델은 MHC-II 분자의 서열 또는 부분서열 및 돌연변이 펩타이드의 변이-코딩 서열을 입력으로서 수신할 수 있고 (예를 들어, 하나 이상의 셀프-어텐션 층을 이용해) 처리할 수 있다.

[0228] 일부 예에서, 머신 러닝 모델은 (가령, 돌연변이 펩타이드와 MHC 분자 간 하나 이상의 잠재적 상호작용에 대응하는 예측을 생성하는 대신 또는 이에 추가하여) 돌연변이 펩타이드, MHC 서열 또는 부분서열, 및 T-세포 수용체 간 하나 이상의 잠재적 상호작용에 대응하는 예측을 생성한다. 그런 다음 머신 러닝 모델은 예를 들어 돌연변이 펩타이드와 T-세포 수용체 사이의 결합 친화도 및/또는 돌연변이 펩타이드가 T-세포에서 면역 반응을 활성화 및/또는 트리거하는지 여부를 예측할 수 있다. 머신 러닝 모델은 T-세포 수용체의 서열 또는 부분서열, MHC의 서열 또는 부분서열, 및 돌연변이 펩타이드의 변이-코딩 서열을 입력으로서 수신할 수 있고 (예를 들어, 하나 이상의 셀프-어텐션 층을 이용해) 처리할 수 있다.

[0229] (예를 들어, 특정 대상체와 관련된) 돌연변이 펩타이드의 면역원성은 본 명세서에 개시된 머신 러닝 모델(예를 들어, 어텐션 기반 머신 러닝 모델)에 의해 생성된 하나 이상의 결과를 기반으로 예측될 수 있다. 예를 들어, 머신 러닝 모델 결과가 돌연변이 펩타이드가 MHC 분자와 낮은 결합 친화도를 가질 것이라고 예측, MHC 분자가 돌연변이 펩타이드를 제시하지 않을 것이거나 제시할 가능성이 낮다고 예측, 및/또는 돌연변이 펩타이드가 T-세포 수용체에 의해 면역 반응을 트리거하지 않을 것이라고 예측할 때, 대상체의 질병 샘플로부터 검출된 신생항원이 면역원성을 트리거하지 않을 것 또는 낮은 면역원성을 가질 것이라고 예측될 수 있다. 돌연변이 펩타이드와 연관되어 생성된 면역원성 예측이, 예를 들어, 숫자(가령, 돌연변이 펩타이드에 반응하여 면역원성 반응이 트리거될 예측 확률에 대응 및/또는 돌연변이 펩타이드에 대한 임의의 면역원성 반응의 예측 강도에 대응), 범주(가령, 면역학적 반응 없음, 낮음, 또는 높음), 또는 이진(가령, 대상체에서 주어진 돌연변이 펩타이드가 면역학적 반응을 트리거할지 여부를 예측)일 수 있다.

[0230] 예측된 면역원성은 추가로 하나 이상의 면역원성 인자의 예측 및/또는 실험적 표시에 기초할 수 있다. 면역원성을 지시하는 인자는 다음을 포함할 수 있다: i) 돌연변이-펩타이드 전구체의 단백질 수준; ii) 돌연변이-펩타이드 전구체를 암호화하는 전사체의 발현 수준; iii) 면역프로테아좀에 의한 돌연변이-펩타이드 전구체의 처리 효율; iv) 돌연변이-펩타이드 전구체를 암호화하는 전사체의 발현 시기; v) T-세포 수용체에 대한 돌연변이 펩타이드의 결합 친화도; vi) 돌연변이 펩타이드 내의 변이 아미노산의 위치; vii) MHC 분자에 결합될 때 돌연변이 펩타이드의 용매 노출; viii) MHC 분자에 결합될 때 변이체 아미노산의 용매 노출; ix) 펩타이드 내 방향족 잔기의 함량; x) 아생형 잔기와 비교할 때 변이 아미노산의 특성; 및/또는 xii) 돌연변이-펩타이드 전구체의 생성; xiii) 미생물 펩타이드를 알기 위한 돌연변이 펩타이드의 미생물 유사성; xiv) 아생형 프로테오에 대한 돌연변이 펩타이드의 자기유사성 또는 비유사성, xv) 아생형 펩타이드의 흥선 발현. 면역원성 인자는 다음을 더 포함할 수 있다: 단백질 서열 및/또는 돌연변이 펩타이드의 길이(예를 들어, 변이-코딩 서열 내에서 식별된 아미노산의 수로 표시됨) 및/또는 대상체 내 MHC 대립유전자의 발현 수준(가령, RNA-Seq 또는 질량분석법으로 측정).

[0231] 결합 친화도 예측 및/또는 돌연변이-펩타이드 제시가 (예를 들어, 대상체 내 하나 이상의 종양 세포 및/또는 하나 이상의 MHC 분자에 의해) 발생할지 여부(또는 그럴 확률)에 대한 예측이 (가령, 대상체로부터의 질병 샘플 내에서 검출된) 돌연변이 펩타이드의 세트의 각각에 대해 (가령, 어텐션-기반 머신 러닝 모델을 이용해) 본 명세서에 개시된 기법에 따라 생성될 수 있다. 이들 예측은 세트의 불완전한 서브세트(예를 들어, 세트의 50% 미만, 세트의 25% 미만, 세트의 10% 미만, 세트의 5% 미만 및/또는 세트의 1% 미만)를 선택하도록 사용될 수 있다. 불완전한 서브세트는 (예를 들어, MHC 분자와 가장 안정한 결합 및/또는 그룹의 다른 것에 비해 제시될 가능성이 가장 높은 세트 내에서 돌연변이 펩타이드를 식별하기 위한) 하나 이상의 상대 임계값 또는 하나 또는 더 많은 절대 임계값을 사용해 선택될 수 있다. 예를 들어, 각각의 선택된 돌연변이 펩타이드는 상대적으로 강

한 친화도 값(예를 들어, 세트 내 최고 50%, 최고 25%, 최고 10% 또는 최고 5% 친화도 값 내) 및/또는 절대적으로 강한 친화도 값(예를 들어, IC50 값의 경우 5000 nM, 1000 nM 또는 500 nM과 같은 지정된 임계값/컷오프보다 우수한 친화도 값을 갖는 것)을 갖고 MHC와 결합 친화도를 가질 수 있다. 세트의 불완전한 서브세트는 지정된 친화도 임계값/컷오프에 관계없이 1, 2, 3, 4, 5, 6, 7, 8, 9, 10개 이상의 돌연변이 펩타이드를 포함할 수 있다. 세트의 불완전한 서브세트는 20개 이상의 신생항원 또는 30개 이상의 돌연변이 펩타이드를 포함할 수 있다.

[0232] 각각의 선택된 돌연변이 펩타이드는 제조, 실험적으로 검사(가령, 결합 친화도, 제시 발생 및/또는 그 밖의 다른 면역학적 요인을 결정하기 위해), 조성물(가령, 약학적 조성물, 가령, 백신 및/또는 치료제)에 포함, 및/또는 대상체에게 투여될 수 있다.

[0233] 결합-친화도 및 제시 예측이 생성되는 돌연변이 펩타이드의 세트의 각각은 특정 대상체(예를 들어, 특정 인간 대상체)와 연관된 돌연변이 펩타이드를 포함할 수 있다. 돌연변이 펩타이드의 세트의 각각은 개체로부터의 질병 특이적 샘플을 사용하여 식별된 질병 특이적 면역원성 돌연변이 펩타이드일 수 있다. 개별 변이-코딩 서열은 질병 샘플에서 유전자 및/또는 핵산 서열(가령, DNA, RNA 및/또는 mRNA 서열)을 시퀀싱하고 각각의 식별된 유전자 및/또는 핵산 서열을 참조-샘플 서열과 비교함으로써 식별될 수 있다. 유전자 및/또는 핵산 서열 내의 코돈은 펩타이드 내 대응하는 아미노산의 존재를 나타낸다. 특히, 다수의 코돈 각각은 주어진 아미노산을 암호화할 수 있으므로 핵산 서열은 (예를 들어, 결정론적으로) 아미노산 서열을 나타낼 수 있지만, 동일한 아미노산 서열은 다른 핵산 서열에 의해 암호화될 수 있다.

[0234] 질병 샘플에서 식별된 서열 중 일부는 비-질병 펩타이드에 대응하는 비-질병 서열일 수 있다. 질병 특이적 핵산 서열 및/또는 질병 특이적 아미노산 서열을 식별하기 위해, 질병 특이적 샘플을 시퀀싱한 결과로서 검출된 각각의 서열에 대해, 참조 서열 데이터 세트에서도 서열이 확인되는지 여부가 결정될 수 있다. 참조 서열 데이터 세트는 서열이 질병(예를 들어, 임의의 질병 또는 주어진 질병)을 나타내거나 특성이 아니라고 알려지거나, 추론되거나, 추정되는 참조 서열 세트를 포함할 수 있다. 참조 서열 데이터 세트는 예를 들어, 질병 특이적 샘플이 수집된 동일한 대상체로부터 수집된 하나 이상의 참조 샘플 서열을 시퀀싱, 임의의 질병 또는 질병 특이적 샘플에 대응하는 질병으로 진단되지 않은 하나 이상의 다른 대상체로부터 수집된 하나 이상의 참조 샘플 서열을 시퀀싱 및/또는 특정 질병과 연관되지 않은 하나 이상의 세포주를 시퀀싱함으로써 식별된 서열을 포함할 수 있다. 일부 경우에, 참조 서열 데이터 세트는 하나 이상의 참조 데이터 저장소로부터 수집된 서열을 포함할 수 있다. 질병 특이적 샘플과 관련하여 검출되지만 참조 서열 데이터 세트에서 검출되지 않는(또는 미리 정의된 임계값 미만의 빈도로 검출되는) 서열은 (예를 들어, 일반적으로 또는 질병 특이적 샘플이 수집된 대상체에 대해) 변이-코딩 서열로 분류될 수 있다.

[0235] 일부 예에서, 다수의 변이-코딩 서열이 식별될 수 있고(예를 들어, 각각은 질병 샘플에서 검출되었지만 참조 샘플 서열에서는 나타나지 않음), 다수의 변이-코딩 서열의 각각의 표현이 결합 친화도 및/또는 프리젠테이션 예측을 예측하기 위해 본 명세서에 개시된 머신 러닝 모델(예를 들어, 어텐션 기반 머신 러닝 모델)을 사용하여 (예를 들어, 개별적으로, 순차적으로 및/또는 병렬로) 처리될 수 있다.

[0236] 질병 샘플은 예를 들어 조직(예를 들어, 고형 종양), 혈액 및/또는 세포 집합체(예를 들어 미세 바늘 흡입 또는 복강경 검사를 사용하여 수집되었을 수 있는 암세포)를 포함할 수 있다. 질병 샘플은 예를 들어, 폐암, 흑색종, 유방암, 난소암, 전립선암 신장암, 위암, 결장암, 고환암, 두경부암, 췌장암, 뇌암, B-세포 림프종, 급성 골수성 백혈병, 만성 골수성 백혈병, 만성 림프구성 백혈병, T 세포 림프구성 백혈병, 비소세포 폐암, 또는 소세포 폐암으로 진단되거나 및/또는 앓고 있는 대상체로부터 수집된 암성 세포를 포함할 수 있다.

[0237] 일부 예에서, 초기 샘플은 질병 샘플과 다른 나머지 샘플(예를 들어, 폐기되거나 참조 샘플로 사용될 수 있음)로 분리된다. 참조 샘플은 매칭된 비-질병 샘플을 포함할 수 있다. 질병 샘플 및 참조 샘플 각각은 동일한 대상체로부터 수집 및/또는 동일하거나 유사한 샘플 유형(가령, 조직 유형)을 포함하거나 그일 수 있다. 일부 예에서, 질병 샘플은 제1 대상체(예를 들어, 의학적 상태 또는 질병으로 진단된 사람)로부터 수집되고, 참조 샘플은 상이한 제2 대상체(예를 들어, 의학적 상태 또는 질병으로 진단되지 않은 사람)로부터 수집된다. 일부 경우에는 참조 샘플 서열이 유기체와 연관된 알려진 유전자의 데이터베이스에서 불러와진다.

[0238] 훈련 데이터는 MHC 분자에 결합된 각각의 펩타이드가 MHC 분자에 의해 제시되었는지 및/또는 면역학적 반응을 트리거했는지 여부에 대한 표시자와 함께 하나 이상의 펩타이드의 서열을 추가로 포함할 수 있다. 서열 데이터를 관찰된 제시 및/또는 결합 데이터와 연관시키는 훈련 데이터를 수집하기 위해, 질병 샘플(그리고 가능하다면 참조 샘플도 함께)은 (개별적으로) 처리되어(가령, MHC에 특이적인 항체를 이용해 면역침강법을 수행함으로써) MHC/펩타이드 복합체를 분리하고 및/또는 (가령, 크로마토그래피 및/또는 질량분석법을 이용해) 펩타이드를 MHC

분자로부터 용리(따라서 시퀀싱)할 수 있다. 일부 예에서, 참조 샘플 서열은 MHC 클래스-I 대립유전자 및/또는 MHC 클래스-II 대립유전자를 포함할 수 있는 하나 이상의 MHC 대립유전자(예를 들어, 질병 샘플에서 검출된 것)를 발현하도록 가공된 하나 이상의 세포주를 시퀀싱함으로써 제시 데이터를 생성하는 데 사용되도록 식별된다. 하나 이상의 세포주는 하나 이상의 대상체로부터 얻거나 유래된 하나 이상의 인간 세포주를 포함할 수 있다. 이 설명의 목적을 위해, 질병 샘플을 사용하여 식별되지만 참조 샘플 서열 세트에 나타나지 않는 펩타이드 서열은 변이-코딩 서열로 식별될 수 있다.

[0239] 일부 구현예에서, 훈련을 위해 사용되기 위한 면역원성-지시 메트릭을 수집하는 것은 HLA 유형 분석을 기반으로 할 수 있으며, 이는 대상체 특이적 MHC 분자 프로필을 식별할 수 있다. 대상체가 인간인 경우, HLA 복합체는 인간의 MHC 단백질을 암호화하는 유전자 복합체이기 때문에 이 프로파일은 인간 백혈구 항원(HLA) 프로파일이라고 할 수 있다. HLA 유형 분석은 대상체의 샘플(가령, 정상 조직 및/또는 비질병 샘플)을 사용하여 수행할 수 있다. 프로파일은 시퀀싱 기법, 가령, PCR 기반 시퀀싱, 직접 시퀀싱 및/또는 차세대 시퀀싱을 이용해 결정될 수 있다. HLA 유형 분석은 예를 들어 고해상도 유형(가령, 세포 표면에서 발현되지 않는 널 대립유전자(null alleles)를 나타내는 것을 제외함) 또는 대립유전자 수준 유형(가령, 정확한 뉴클레오타이드 서열 HLA-유전자 결정을 나타냄)을 포함할 수 있다. HLA 유형 분석은 저해상도 유형 분석 및/또는 더 광범위한 대립유전자 계열을 식별하는 HLA 수퍼유형 분석을 포함할 수 있다.

[0240] 임의의 유형의 시퀀싱과 관련하여(예를 들어, 샘플 내 서열, MHC 분자에 대한 펩타이드 결합, HLA 유형을 식별하기 위한 것), 결과는 하나 이상의 핵산 서열 또는 하나 이상의 아미노산 서열을 식별할 수 있다. 핵산 서열이 식별되고 어텐션 기반 모델(또는 기타 처리)이 아미노산 서열을 처리하도록 구성될 때, 기술(가령, 룩업 테이블)이 사용되어 핵산 서열 내의 개별 코돈을 개별 아미노산으로 전환할 수 있다.

[0241] 일부 실시예는 펩타이드(예를 들어, 펩타이드, 가령, 선택된 펩타이드를 암호화하는 핵산 서열을 사용함) 또는 선택된 펩타이드에 대한 전구체를 합성하는 것을 포함한다. 그런 다음 합성된 펩타이드 또는 전구체는 대응하는 제시 및/또는 결합 데이터를 식별하기 위해(예를 들어, 예측된 제시 및/또는 결합을 검증하거나 훈련을 위해 사용할 결과를 생성하기 위해) 실험에 사용될 수 있다. 예를 들어, 실험은 ELISA 폴다운 어세이, 겔-이동 어세이 또는 바이오센서 기반 방법론을 사용하여 특정 MHC 분자와 선택된 펩타이드의 결합 친화도를 평가하는 것을 포함할 수 있다. 또 다른 예로서, 실험은 펩타이드-MHC 면역침강법, 및 이에 뒤 따르는 질량분석법에 의한 제시된 MHC 리간드의 용리 및 검출을 이용함으로써 선택된 펩타이드가 MHC 분자에 의해 제시되었는지 여부를 나타내는 용리 데이터를 수집하는 것을 포함할 수 있다.

[0242] 개별 펩타이드가 개별 MHC에 결합 및/또는 제시되었는지 여부를 나타내는 훈련 또는 검증 데이터에 추가하여 또는 대신하여, 훈련 또는 검증 데이터가 개별 펩타이드가 면역원성을 트리거했는지 여부를 나타낼 수 있다. 면역원성 결과는 생체내 또는 생체외 검사를 이용해 결정될 수 있다. 하나 이상의 선택된 펩타이드를 검사하는 것은 하나 이상의 면역원성 인자(예를 들어, 주어진 사건이 발생하는지 여부 및/또는 정도를 결정하기 위해) 및/또는 면역원성(예를 들어, 펩타이드가 면역학적 반응을 트리거하는지 여부 및/또는 정도를 결정하기 위해)을 조사하도록 구성될 수 있다. 검사는 하나 이상의 펩타이드를 포함하는 조성물(예를 들어, 백신)을 주어진 대상체(예를 들어, 돌연변이-펩타이드 선택 동안 사용된 MHC 서열이 식별됨)에 투여하는 것이 의학적 상태(가령, 종양) 또는 질병(가령, 암)의 예방 또는 치료에 효과적인지 여부를 조사하도록 구성될 수 있다. 대상체는 인간 대상체일 수 있다.

[0243] 일부 실시예는 하나 이상의 선택된 돌연변이 펩타이드(또는 하나 이상의 선택된 돌연변이 펩타이드를 암호화하는 복수의 핵산)에 기초한 조성물을 제조하는 것을 포함한다. 예를 들어, 하나 이상의 선택된 돌연변이 펩타이드의 각각은 대상체에 대한 MHC 분자에 결합하고 이에 의해 제시될 것으로 예측되었을 수 있다(예를 들어, 적어도 임계값 정도로). 조성물은 하나 이상의 선택된 돌연변이 펩타이드, 하나 이상의 선택된 돌연변이 펩타이드에 대한 하나 이상의 전구체, 하나 이상의 선택된 돌연변이 펩타이드에 대응하는 하나 이상의 폴리펩타이드 서열, 하나 이상의 선택된 돌연변이 펩타이드에 대응하는 RNA(예를 들어, mRNA), 하나 이상의 선택된 돌연변이 펩타이드에 대응하는 DNA, 하나 이상의 선택된 돌연변이 펩타이드를 포함하는 세포(가령, 항원-제시 세포) 및/또는 이러한 펩타이드를 암호화하는 핵산(들), 하나 이상의 선택된 돌연변이 펩타이드에 대응하는 플라스미드 및/또는 하나 이상의 선택된 돌연변이 펩타이드에 대응하는 벡터 각각을 포함할 수 있다.

[0244] 조성물은 보조제, 부형제, 면역조절제, 관문 단백질, PD-1의 길항제(예를 들어, 항-PD-1 항체) 및/또는 PD-L1의 길항제(예를 들어, 항-PD-L1 항체)를 더 포함할 수 있다. 조성물은 종양 백신과 같은 백신일 수 있다. 조성물은 특정 대상체를 위해 제조되거나 선택된 개별화된 백신일 수 있다.

- [0245] 조성물은 폴리뉴클레오타이드 작제물(예를 들어, DNA 작제물 또는 RNA 작제물)을 포함할 수 있다. 폴리뉴클레오타이드 작제물은 표적 조직 또는 세포에 '이식'될 수 있는 인위적으로 구성된 핵산 분절이다. 폴리뉴클레오타이드 작제물은 하나 이상의 선택된 돌연변이 펩타이드를 암호화하는 뉴클레오타이드 서열을 함유하는 DNA 또는 RNA(예를 들어, mRNA) 삽입물을 포함한다. 항원 제시(예를 들어, MHC 분자에 의한 하나 이상의 선택된 돌연변이 펩타이드의 제시)를 증가시키기 위해, 폴리뉴클레오타이드 작제물은 개선된 항원 제시를 위해 개발된 변이를 추가로 포함할 수 있고, 따라서 하나 이상의 선택된 돌연변이 펩타이드에 대한 개선된 면역원성을 포함할 수 있다. 일부 예에서, 변형은 그 전문이 모든 목적으로 본 명세서에 참조로서 포함되는 국제 공개 W02005038030A1에 기재된 바와 같이 MHC 분자의 사슬의 막횡단 영역 및 세포질 영역을 폴리뉴클레오타이드 작제물로 통합하는 것이다.
- [0246] 증가된 안정성 및 번역 효율을 갖는 RNA 삽입물을 제공하기 위해, 폴리뉴클레오타이드 작제물은 개선된 안정성 및 번역을 위해 개발된 변이를 더 포함할 수 있고, 따라서 하나 이상의 선택된 돌연변이 펩타이드에 대한 면역원성을 개선할 수 있다. 일부 예에서, 변형은 그 전문이 모든 목적으로 본 명세서에 참조로서 포함되는 국제 공개 W02007036366A2에 기술된 바와 같이 인간 베타-글로빈 유전자의 3'-비번역 영역의 적어도 2의 복제수를 갖는 핵산 서열을 폴리뉴클레오타이드 작제물에 통합하는 것이다. 다른 예에서, 변형은 그 전문이 모든 목적으로 본 명세서에 참조로서 포함되는 국제 공개 W02017060314A3에 기술된 바와 같이 3'-비번역 영역, 가령, F1 3' UTR를 코딩하는 핵산 서열의 통합이다.
- [0247] 증가된 안정성 및 발현을 갖는 RNA 삽입물을 제공하기 위해, 폴리뉴클레오타이드 작제물은 개선된 안정성 및 발현을 위해 개발된 변형을 추가로 포함할 수 있고, 따라서 하나 이상의 선택된 돌연변이 펩타이드에 대한 면역원성을 개선할 수 있다. 일부 경우, 5'-캡 구조와 같은 RNA 말단에 캡을 통합하는 변형이 있다. 캡 구조는 그 전문이 모든 목적으로 본 명세서에 참조로서 포함되는 국제 공개 W02011015347A1에 기술된 바와 같이 베타-S-ARCA의 D1 부분입체이성질체일 수 있다.
- [0248] 높은 선택성을 갖는 폴리뉴클레오타이드 작제물을 항원 제시 세포에 전달하기 위해, 조성물은 폴리뉴클레오타이드 작제물의 개선된 흡수를 위한 양이온성 리포좀 또는 리포플렉스를 더 포함할 수 있고, 따라서 하나 이상의 선택된 돌연변이 펩타이드에 대한 개선된 면역원성을 가질 수 있다. 일부 경우에, 조성물은 폴리뉴클레오타이드 작제물을 포함하는 나노입자를 포함한다. 나노입자는 그 전문이 모든 목적으로 본 명세서에 참조로서 포함되는 국제 공보 W02013143683A1에 기술된 바와 같이 하나 이상의 지질, 가령, DOTMA 및 DOPE을 포함하는 리포플렉스일 수 있다.
- [0249] 일부 실시예는 하나 이상의 선택된 돌연변이 펩타이드를 포함하는 유효량의 조성물(예를 들어, 백신)을 개체에 투여함으로써 개체에서 의학적 병태(예를 들어, 종양) 또는 질병(예를 들어, 암)을 치료하는 것을 포함한다. 개체는 질병 샘플이 수집된 것과 동일한 개체일 수 있다. 일부 경우에는, 질병 샘플이 수집된 개체와 비교하여 상이한 개체에게 백신이 투여된다. 예를 들어, 상이한 개체는 질병 샘플이 수집된 개체와 관련이 있을 수 있고, 특정 유형의 암이 발병할 유전적 위험이 있고/있거나, 질병 샘플을 수집한 대상체의 하나 이상의 MHC 대립유전자와 동일한(또는 유사한) 서열에 대응하는 하나 이상 또는 모든 대립유전자를 갖는 MHC 분자를 가질 수 있다.
- [0250] 일부 실시예에서, (예를 들어, 대상체의 샘플에서 검출된) 돌연변이 펩타이드의 세트의 각각에 대해, 본 명세서에 개시된 하나 이상의 기법이 사용되어 돌연변이 펩타이드가 대상체의 MHC 분자에 결합할지 여부(또는 이러한 결합의 강도, 안정성 및/또는 출현율)를 예측 및/또는 대상체의 MHC 분자가 돌연변이 펩타이드를 제시할지 여부(및/또는 그러한 제시의 출현율)를 예측할 수 있다. 예측은 돌연변이 펩타이드의 불완전한 서브세트를 선택하기 위해 사용될 수 있다(예를 들어, 돌연변이 펩타이드의 MHC 제시가 있을 것으로 예측됨). 선택은 각각의 돌연변이 펩타이드에 대해 예측 메트릭에 대응하는 메트릭을 절대 임계값 및/또는 다른 돌연변이 펩타이드의 메트릭의 예측 메트릭과 비교하는 것(예를 들어, 상대적 비교를 수행하는 것)을 포함할 수 있다. 각각의 선택된 돌연변이 펩타이드는 다음을 갖는 것으로 식별될 수 있다: 종양 세포 표면에 제시될 높은 가능성, 종양 특이적 면역 반응을 유도할 높은 가능성, 전문 항원 제시 세포(가령, 수지상 세포)에 의해 미접촉 T 세포에 제시될 높은 가능성, 중추 또는 말초 내성을 통해 억제될 낮은 가능성 및/또는 대상체 내 정상 조직에 대한 자가면역 반응을 유도할 수 있을 낮은 가능성.
- [0251] 일부 실시예는 MHC 분자에 결합할 가능성이 있고 종양 세포의 표면에서 MHC 분자에 의해 제시될 하나 이상의 펩타이드(예를 들어, 돌연변이 펩타이드)를 식별하기 위한 모델을 생성 및/또는 사용하는 것을 포함한다. 보다 구체적으로, 훈련 데이터 세트는 데이터 요소의 세트를 포함할 수 있으며, 각각의 데이터 요소는 항원결정기(또는 펩타이드)의 서열(예를 들어, 잠재적으로 펩타이드의 N-플랭크 및 펩타이드의 C-플랭크의 서열), MHC 분자의 부

분서열, 펩타이드 및 MHC 분자에 관한 하나 이상의 실험 결과(가령, 결합 친화도 및/또는 용리된-리간드 제시 데이터)를 포함한다.

[0252] 어텐션-기반 머신 러닝 모델은 훈련 데이터 세트의 적어도 일부분을 사용하여 훈련될 수 있다. 훈련 데이터 세트는 여러 훈련 데이터 요소를 포함할 수 있다. 각각의 훈련 데이터 요소는 서열의 표현 및 결과(예를 들어, 서열에 대응하는 펩타이드의 적어도 일부가 MHC 분자에 의해 제시 및/또는 면역원성을 트리거하는지 여부를 나타냄)를 포함할 수 있다. 제시가 검출되지 않은 훈련 데이터 요소는 계산적으로 생성될 수 있다. 예를 들어, 양성 세트(양성 용리된 리간드 제시 데이터에 대응)의 각각의 기원 단백질에 대해, 가능한 균일한 확률로 각각의 길이에 대해 하나 이상의 또는 모든 가능한 펩타이드 단편(예를 들어, 지정 길이 범위, 예를 들어 8에서 11까지)이 생성될 수 있다. (예를 들어, 잠재적으로 최대 길이, 예컨대 10개 아미노산을 갖는) N-말단 및 C-말단 플랭킹 서열은 유지될 수 있다. 일부 경우에, 훈련 데이터에서 양성 사례로 표현된 각각의 대립유전자에 대해, 펩타이드 단편(예를 들어, 8:11의 하나, 그 이상 또는 모든 길이)이 생성될 수 있다. 생성 및/또는 후속 선택은 주어진 길이를 갖는 서열의 발생 확률이 길이에 걸쳐 균일하도록 수행될 수 있다. N-말단 및 C-말단 플랭킹 서열은 특정 최대 길이(예를 들어, 10개 아미노산의 최대 길이)로 유지될 수 있거나 유지되었을 수 있다.

[0253] 어텐션-기반 머신 러닝 모델은 1, 2, 3, 4, 5, 6, 7, 8개 이상의 변환기 인코더 망(예를 들어, 각각은 원-헤드 어텐션 및 피드포워드 망을 포함함)을 포함할 수 있다. 예를 들어 어텐션-기반 머신 러닝 모델은 펩타이드의 표현을 처리하도록 구성된 변환기 인코더, MHC 분자의 표현을 처리하도록 구성된 변환기 인코더, 잠재적으로 펩타이드 N-플랭크의 표현을 처리하도록 구성된 변환기 인코더, 및 펩타이드 C-플랭크의 표현을 처리하도록 구성된 변환기 인코더를 포함하는 다수의 제1 레벨 변환기 인코더를 포함할 수 있다. 어텐션-기반 머신 러닝 모델은 제1 레벨 변환기 인코더에 의해 생성된 통합(예를 들어, 연결된) 결과를 처리하도록 구성된 제2 레벨 변환기 인코더를 더 포함할 수 있다.

[0254] 어텐션-기반 머신 러닝 모델은 (가령, 드롭아웃이 적용된 후) 제5 변환기 인코더로부터의 결과를 처리하여 예측(가령, 실수) 결합 친화도 및/또는 예측 제시(가령 이진 예측)를 생성하도록 구성된 피드포워드 망(가령, 하나, 둘, 또는 그 이상의 은닉 층이 있는 완전 연결 피드포워드 망)을 더 포함할 수 있다. 어텐션-기반 머신 러닝 모델은 모델의 앙상블 내에서 (가령, 동일한 구성을 갖는) 하나 또는 다수의 모델일 수 있다. 훈련 데이터 세트는 앙상블 내에서 다양한 모델을 훈련하기 위해 랜덤하게 파싱, 셔플 및/또는 분할될 수 있다. 손실 함수는 오차 항(가령, 평균 제곱 오차 또는 중앙값 제곱 오차) 및/또는 엔트로피 항(가령, 교차 엔트로피 또는 이진 교차 엔트로피)을 사용할 수 있다. 두 가지 상이한 유형의 결과(가령, 결합 친화도 및 제시 발생) 각각을 예측하도록 모델이 동시에 훈련되도록 멀티태스킹 학습이 사용될 수 있다. 정적 또는 비-정적 학습물이 사용될 수 있다. 예를 들어 학습물 어닐링(가령, 단계적 어닐링 또는 코사인 어닐링 사용)이 사용되어 반복구간 동안 학습율을 감소시킬 수 있다. 검증-데이터 평가가 사용되어 잠재적으로 훈련을 조기에 종료할 수 있다(가령, 성과 목표가 충족되었다고 판단한 경우).

[0255] MHC는 생체내 다수의 대립유전자(예를 들어, 인간당 6개의 대립유전자)를 포함한다. 따라서, 이 단일 MHC 분자에 대해, 다수의 서열 입력이 생성될 수 있다(예를 들어, 각각은 다수의 대립유전자의 단일 대립유전자를 나타냄). 다수의 서열 입력의 각각은 하나 이상의 신경망(가령, 하나 이상의 변환기 인코더)을 이용해 개별적으로 처리되어, 각각의 대립유전자와 연관된 신생항원의 예측된 결합 또는 제시 값을 생성할 수 있다. 함수(예를 들어, 소프트맥스 함수)는 다수의 대립유전자 중 어떤 대립유전자가 가장 높은 제시 예측과 연관되어 있는지 식별할 수 있다. 훈련 동안, 이 특정 서열 입력에 대한 이 최대 제시 예측은 이진 손실 함수를 사용하여 실제 제시 값에 비교되어 파라미터를 튜닝하기 위한 오차를 생성할 수 있다.

[0256] 어떤 경우에는, 플랭크(예를 들어, N-플랭크)로부터 얼마나 많은 아미노산이 제시되는 펩타이드 코어로 긴 펩타이드를 트리밍할 때를 결정하기 위해 펩티다아제에 의해 사용되는지는 알려져 있지 않다. 훈련 데이터를 생성할 때 이러한 미지의 문제를 해결하기 위해, 기법(가령, 의사-랜덤 선택 기법)을 기반으로 선택된 길이, 가령, 지정 범위 내의 길이(가령, 1 내지 10개 아미노산)로 플랭크가 트리밍될 수 있다. 선택 기법은 분포(예를 들어, 균일 또는 가우시안 분포)를 사용하여 길이를 선택할 수 있다. 일부 경우, 임계 길이(가령, 10개 아미노산) 미만인 플랭크가 트리밍되지 않는다. 일부 경우, 플랭크 트리밍은 N-플랭크 상에서 C 측면을 보존하는 방식으로 정의된다.

[0257] 그런 다음 훈련된 모델은 하나 이상의 돌연변이-펩타이드 서열(예를 들어, N-플랭크 영역, 후보 항원결정기 영역 및/또는 C-플랭크 영역)의 표현(들) 및 (대상체와 연관된) MHC 분자의 부분서열을 포함하고 예측된 결합 친화도 및/또는 제시 예측을 생성하는 입력 세트를 수신할 수 있다. 돌연변이-펩타이드가 MHC 분자에 안정적으로

결합하여 제시될 것으로 예상되는 경우, 돌연변이-펩타이드는 대상체를 치료하기 위해 사용되는 조성물(예를 들어, 백신)에 포함되도록 선택될 수 있다.

[0258] II.D. 머신 러닝 모델을 위한 입력 데이터의 예시적인 식별

[0259] 본 명세서에 기재된 입력 데이터를 식별하기 위한 예시적인 방법 및 시스템은 예를 들어 도 1 및 3의 머신 러닝 모델(132) 및/또는 도 4a-4c에서 기재된 머신 러닝 모델(132)에 대한 입력 데이터를 식별하는 데 사용될 수 있다.

[0260] 주어진 대상체와 연관된 돌연변이 펩타이드의 세트의 각각은 돌연변이 펩타이드의 결합 친화도, 제시 확률 및/또는 면역원성에 대한 하나 이상의 예측을 생성하기 위해 어텐션-기반 머신 러닝 모델을 사용하여 분석될 수 있다. 이들 예측을 생성하기 위해, 머신 러닝 모델은 돌연변이 펩타이드에 대응하는 펩타이드(가령, 코딩) 서열 및 하나 이상의 다른 서열 또는 부분서열(가령, MHC-I 분자, MHC-II 분자, 또는 T-세포 수용체에 대응)을 수신하고 처리할 수 있다. 일부 경우에, 펩타이드 서열의 세트(예를 들어, 돌연변이 펩타이드의 세트에 대응하는 변이-코딩 서열의 세트) 각각에 대해 예측이 생성된다. 돌연변이 펩타이드의 세트는 대상체로부터 수집된 질병 샘플에 존재하지만 (예를 들어, 대상체 또는 다른 대상체로부터의) 하나 이상의 비질병 샘플에서는 관찰되지 않는 펩타이드에 대응할 수 있다.

[0261] 주어진 대상체와 연관된 돌연변이 펩타이드의 세트를 식별하기 위해 다양한 방법이 이용 가능하다. 돌연변이는 대상체의 질병 세포의 게놈, 전사, 단백질 또는 엑솜에 존재할 수 있지만, 비-질병은 샘플, 예를 들어, 상기 대상체 또는 다른 대상체로부터의 비-질병 샘플에는 존재하지 않을 수 있다. 돌연변이는 (1) 단백질에서 상이한 아미노산을 유도하는 비동의 돌연변이, (2) 종결 코돈이 변형되거나 결실되어 C-말단에서 새로운 종양 특이적 서열을 갖는 더 긴 단백질의 번역을 유도하는 관독-통과 돌연변이(read-through mutation), (3) 성숙한 mRNA에 인트론을 포함시켜 독특한 종양 특이적 단백질 서열을 만드는 스플라이스 부위 돌연변이, (4) 2개의 단백질의 접합부에서 종양 특이적 서열을 갖는 키메라 단백질을 발생시키는 염색체 재배열(즉, 유전자 융합), (5) 새로운 종양 특이적 단백질 서열을 가진 새로운 오픈 리딩 프레임으로 이어지는 프레임시프트 삽입 또는 결실을 포함하나 이에 한정되지는 않는다. 돌연변이는 또한 비프레임시프트 삽입결실, 미스센스 또는 넌센스 치환, 스플라이스 부위 변경, 게놈 재배열 또는 유전자 융합, 또는 neoORF를 발생시키는 임의의 게놈 또는 발현 변경을 포함할 수 있다.

[0262] 질병이 있는 세포에서 예를 들어, 스플라이스 부위, 프레임이동, 관독 또는 유전자 융합 돌연변이로부터 발생하는 돌연변이 또는 돌연변이된 폴리펩타이드를 갖는 펩타이드는 질병이 있는 샘플에서 DNA, RNA 또는 단백질을 시퀀싱하고 얻어진 서열을 비-질병 샘플로부터의 서열과 비교함으로써 확인할 수 있다.

[0263] 일부 구현예에서, 질병 샘플 및 비-질병 샘플로부터의 전장 게놈 시퀀싱(WGS) 또는 전장 엑솜 시퀀싱(WES) 데이터가 획득되고 비교될 수 있다. 비-질병 샘플과 질병 샘플 리드를 인간 참조 게놈에 정렬한 후, 단일 뉴클레오타이드 변이체(SNV), 유전자 융합 및 삽입 또는 결실 변이체(indel)를 포함하는 체세포 변이체가 변이-검출 알고리즘을 사용하여 검출될 수 있다. 하나 이상의 변이 검출기(variant caller)가 사용되어 상이한 체세포 변이 유형(즉, SNV, 유전자 융합 또는 삽입결실)을 검출할 수 있다(그 전문이 모든 목적으로 본 명세서에서 참조로서 포함되는, Xu et al., "A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data." *Comput. Struct. Biotechnol. J.* 16: 15-24 (2018)을 참조할 수 있다).

[0264] 일부 예에서, 돌연변이 펩타이드는 개체로부터의 질병 샘플의 전사체 서열에 기초하여 식별된다. 예를 들어, 전체 또는 부분 전사체 서열(예를 들어, RNA-Seq과 같은 방법에 의해)은 개체의 질병 조직으로부터 얻어질 수 있고 시퀀싱 분석을 받을 수 있다. 그런 다음 질병 조직 샘플로부터 획득된 서열이 참조 샘플로부터 획득된 서열에 비교될 수 있다. 선택적으로, 질병 조직 샘플은 전장 전사체 RNA-Seq에 적용된다. 선택적으로, 전사체 서열은 참조 샘플과 비교하기 전에 특정 서열에 대해 "강화"된다. 예를 들어, 시퀀싱 분석을 받기 전에 특정한 원하는 서열(가령, 질병 특이적 서열)을 강화하도록 특정 프로브가 설계될 수 있다. 전장-전사체 시퀀싱 및 표적화된 시퀀싱 방법이 해당 분야에 공지되어 있고, 가령, Tang, F. et al., "mRNA-Seq whole-transcriptome analysis of a single cell," *Nature Methods*, 2009, v.6, 377-382; Ozsolak, F., "RNA sequencing: advances, challenges and opportunities," *Nature Reviews*, 2011, v.12, 87-98; German, M.A et al., "Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends," *Nature Biotechnology*, 2008, v.26, 941-946; 및 Wang, Z. et al., "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews*, 2009, v.10, p.57-63에 보고되어 있다. 이들 참조문헌의 각각은 본 명세서에 그 전체가 모든 목적으로 참조로서 포함된다.

- [0265] 일부 구현예에서, 전사체 시퀀싱 기법의 비제한적 예를 들면 RNA 폴리(A) 라이브러리, 마이크로어레이 분석, 병렬 시퀀싱, 대규모 병렬 시퀀싱, PCR 및 RNA-Seq를 이 있다. RNA-Seq은 전사체의 일부 또는 실질적으로 전체를 시퀀싱하기 위한 고처리량 기법이다. 요컨대, 전사체 서열의 분리된 모집단은 어댑터가 한쪽 또는 양쪽 끝에 부착된 cDNA 단편 라이브러리로 변환된다. 증폭 여부에 관계없이 각 cDNA 분자가 분석되어 일반적으로 30-400 염기쌍의 짧은 서열 정보가 획득된다. 그런 다음 이들 서열 정보 조각이 참조 게놈, 참조 전사물에 정렬되거나 신규로 조립되어 전사체 구조(즉, 전사 경계) 및/또는 발현 수준을 나타낼 수 있다.
- [0266] 일단 획득되면, 질병 샘플의 서열은 참조 샘플의 대응하는 서열에 비교될 수 있다. 서열 비교는 질병 조직의 핵산 서열을 참조 샘플의 대응하는 서열과 정렬함으로써 핵산 수준에서 수행될 수 있다. 그런 다음 암호화된 아미노산에서 하나 이상의 변화를 일으키는 유전적 서열 변이가 식별된다. 대안으로, 서열 비교는 아미노산 수준에서 수행될 수 있는데, 즉 핵산 서열은 비교가 수행되기 전에 먼저 인실리코(in silico)로 아미노산 서열로 변환된다. 아미노산 기반 접근 방식 또는 핵산 기반 접근 방식이 사용되어 펩타이드에서 하나 이상의 돌연변이(가령, 하나 이상의 점 돌연변이)를 식별할 수 있다. 핵산 기반 접근법과 관련하여, 발견된 변이체는 (가령, 다중 코돈 변이체와 개별 펩타이드 돌연변이를 연관시키는 조희표를 통해) 주어진 관찰 가능한 돌연변이 단백질질을 일으킬 하나 이상의 핵산 서열(가령, DNA 서열, RNA 서열 또는 mRNA 서열)을 식별하는 데 사용될 수 있다.
- [0267] 일부 실시예에서, 질병 샘플의 서열과 참조 샘플의 서열의 비교는 수동 정렬, FAST-A11(FASTA) 및 기본 로컬 정렬 검색 도구(BLAST: Basic Local Alignment Search Tool)와 같은 해당분야 공지된 기술에 의해 완료될 수 있다. 일부 실시예에서, 짧은 판독 정렬기, 예를 들어 GSNAP, BWA 및 STAR를 사용하여 질병 샘플로부터의 서열과 참조 샘플의 서열의 비교를 완료할 수 있다.
- [0268] 일부 실시예에서, 참조 샘플은 매칭된 무질병 샘플이다. 본 명세서에서 사용될 때, "매칭된" 무질병 조직 샘플은 동일하거나 유사한 샘플, 예를 들어 질병 샘플과 동일하거나 유사한 조직 유형으로부터의 샘플로부터 선택되는 것이다. 일부 실시예에서, 매칭되는, 무질병 조직 및 질병 조직은 동일한 개체로부터 기원할 수 있다. 일부 실시예에서 본 명세서에 기재된 참조 샘플은 동일한 개체로부터의 무질병 샘플이다. 일부 실시예에서, 참조 샘플은 상이한 개체(예를 들어, 질병을 앓지 않는 개체)로부터의 무질병 샘플이다. 일부 실시예에서, 참조 샘플은 상이한 개체의 모집단으로부터 얻어진다. 일부 실시예에서, 참조 샘플은 유기체와 연관된 공지된 유전자의 데이터베이스이다. 일부 구현예에서 참조 샘플은 세포주로부터 유래할 수 있다. 일부 실시예에서, 참조 샘플은 유기체와 연관된 공지된 유전자 및 매칭된 무질병 샘플로부터의 게놈 정보의 조합일 수 있다. 일부 실시예에서, 변이-코딩 서열은 아미노산 서열 내 점 돌연변이를 포함할 수 있다. 일부 실시예에서, 변이-코딩 서열은 아미노산 결실 또는 삽입을 포함할 수 있다.
- [0269] 일부 실시예에서, 변이-코딩 서열의 세트는 먼저 게놈 및/또는 핵산 서열에 기초하여 식별된다. 이 초기 세트는 전사체 시퀀싱 데이터베이스 내 변이-코딩 서열의 존재를 기반으로 더 좁은 세트의 발현 변이-코딩 서열을 얻기 위해 추가로 필터링된다(따라서 "발현된" 것으로 간주됨). 일부 실시예에서, 변이-코딩 서열의 세트는 전사체 시퀀싱 데이터베이스를 통한 필터링에 의해 적어도 약 10, 20, 30, 40, 50배 이상 감소된다.
- [0270] 대안으로, 단백질 질량 분석법은 돌연변이 펩타이드, 예를 들어 종양 세포 상의 MHC 단백질에 결합된 돌연변이의 존재를 식별하거나 검증하기 위해 사용될 수 있다. 펩타이드는 예를 들어, 종양 세포와 같은 병든 세포로부터 또는 종양으로부터 면역침강된 HLA 분자로부터 산-용리될 수 있고, 이어서 질량 분석법을 사용하여 식별될 수 있다.
- [0271] 돌연변이 펩타이드는 예를 들어 5개 이상, 8개 이상, 11개 이상, 15개 이상, 20개 이상, 40개 이상, 80개 이상, 100개 이상, 120개 이하, 100개 이하, 80개 이하, 60개 이하, 50개 이하, 40개 이하, 30개 이하, 25개 이하, 20개 이하, 18개 이하, 15개 이하 또는 13개 이하의 아미노산을 가질 수 있다.
- [0272] 종양 특이적 T 세포 수용체 서열은 또한 예를 들어 단일 세포 T 세포 수용체 시퀀싱에 의해 식별될 수 있다. 예를 들어, De Simone et al. "Single Cell T Cell Receptor Sequencing: Techniques and Future Challenges," *Front. Immunol.* 9: 1638 (2018); Zong et al. "Very rapid cloning, expression and identifying specificity of T-cell receptors for T-cell engineering," *PloS ONE* 15(2):e0228112 (2020)를 참조할 수 있다(모든 목적을 위해 전체 내용이 본 명세서에 참조로 포함됨). 추가로 또는 대안으로 T 세포 레퍼토리의 높은 처리량 시퀀싱이 특정 질병에 대한 종양 특이적 시그니처를 식별하기 위해 수행될 수 있다. 예를 들어, Wang et al. "High-throughput sequence of CD4+T cell repertoire reveals disease-specific signatures in IgG4-related disease," *Arthritis Research & Therapy* 21: 295 (2019)를 참조할 수 있다(모든 목적을 위해 전체 내

용이 본 명세서에 참조로 포함됨).

- [0273] MHC-I 서열 및/또는 MHC-II 서열이, 예를 들어, HLA 유전자형분석 또는 질량분석을 통해 결정될 수 있다(Caron et al., "Analysis of Major Histocompatibility Complex (MHC) Immunopeptides Using Mass Spectroscopy," *Molecular and Cellular Proteomics* 14(12): 3105-3117 (2015)(모든 목적을 위해 전체 내용이 본 명세서에 참조로 포함됨).
- [0274] II.E. 머신 러닝 모델을 위한 훈련 데이터의 예시적인 식별
- [0275] 본 명세서에 기재된 훈련 데이터를 식별하기 위한 예시적인 방법 및 시스템은 예를 들어 도 1 및 3의 머신 러닝 모델(132) 및/또는 도 4a-4c에서 기재된 머신 러닝 모델(132)에 대한 훈련 데이터를 식별하는 데 사용될 수 있다. 예를 들어, 이들 방법 및 시스템은 도 1의 훈련 데이터(131)를 식별하는 데 사용될 수 있다.
- [0276] 훈련 세트는 다수의 다른 샘플(예를 들어, 잠재적으로 하나 이상의 다른 대상체와 연관됨)로부터 수집된 데이터를 사용하여 생성될 수 있다. 다수의 다른 샘플 각각은 예를 들어 조직(예를 들어, 생검), 단일 세포, 다수의 세포, 세포 단편 또는 체액의 분취량을 포함할 수 있다. 일부 경우, 훈련된 모델에 의해 처리될 입력 데이터와 연관된 대상체와 비교하여 상이한 유형의 대상체로부터 다수의 다른 샘플이 수집된다. 예를 들어, 머신 러닝 모델은 하나 이상의 세포주로부터의 샘플을 처리함으로써 수집된 훈련 데이터를 사용하여 훈련될 수 있고, 훈련된 머신 러닝 모델은 인간 대상체로부터의 하나 이상의 샘플을 처리함으로써 결정된 입력 데이터를 처리하는 데 사용될 수 있다.
- [0277] 훈련 데이터 세트는 다수의 훈련 요소를 포함할 수 있다. 다수의 훈련 요소의 각각은 대응하는 펩타이드의 임의의 변이를 각각 코딩 및/또는 표현하는 펩타이드 서열의 세트(야생형 또는 변이-코딩 서열의 세트를 포함함), 및 MHC 분자의 부분서열 또는 의사서열을 포함하는 입력 데이터를 포함할 수 있다. 입력 데이터는 본 명세서(예를 들어, 섹션 II.D)에 개시된 하나 이상의 기법에 따라 수집될 수 있다.
- [0278] 각각의 훈련 요소는 또한 하나 이상의 실험 기반 결과를 포함할 수 있다. 실험 기반 결과는 야생형 펩타이드 또는 돌연변이 펩타이드(훈련 요소의 변이-코딩 서열과 연관됨)와 MHC 분자(훈련 요소의 MHC 분자 서브서열과 연관됨) 사이의 하나 이상의 특정 유형의 상호작용이 발생하는지 여부 및/또는 정도를 나타낼 수 있다. 특정 유형의 상호작용은 예를 들어 MHC 분자에 대한 펩타이드의 결합 및/또는 세포(예를 들어, 종양 세포)의 표면 상의 MHC 분자에 의한 펩타이드의 제시를 포함할 수 있다.
- [0279] 결과는 펩타이드와 MHC 분자 사이의 결합 친화도를 포함할 수 있다. 결과는 주어진 펩타이드가 주어진 MHC 분자와 결합하는지 여부, 이러한 결합의 강도, 이러한 결합의 안정성 및/또는 이러한 결합의 발생 경향을 특징짓는 정성적 데이터 및/또는 정량적 데이터를 포함하거나 이를 기반으로 할 수 있다. 예를 들어, ELISA, 폴다운 어세이, 젤 이동 어세이, 바이오센서 기반 방법론, 예를 들어 표면 플라스몬 공명(Surface Plasmon Resonance), 등온 적정 비색법(Isothermal Titration Colorimetry), 바이오레이어 간섭계(BioLayer Interferometry), 또는 마이크로스케일 열영동(MicroScale Thermophoresis)을 사용하여 이진 결합 친화도 표시자 또는 정성적 이진 친화도 결과가 생성될 수 있다.
- [0280] 결과는 예를 들어 주어진 MHC 분자가 주어진 펩타이드를 제시하는지 여부 및/또는 제시 확률을 추가로 또는 대안으로 특징지을 수 있다. MHC 리간드는 샘플에서 면역침강될 수 있다. MHC 분자가 리간드를 제시하는지 여부를 결정하기 위해 후속 용리 및 질량분석법이 사용될 수 있다.
- [0281] III. 제약상 허용되는 조성물 및 제조
- [0282] 하나 이상의 변이-코딩 서열은 본 명세서에 기재된 하나 이상의 머신 러닝 모델로부터의 결과에 기초하여 변이-코딩 서열의 대상체 특이적 세트의 세트로부터 선택될 수 있다. 예를 들어, 선택은 예측된 결합 친화도가 500 nM 미만인 변이-코딩 서열의 대상체 특이적 세트의 각각을 식별하는 것을 포함할 수 있으며, 이에 대해 MHC 분자는 변이-코딩 서열에 의해 식별되는 돌연변이 펩타이드를 제시할 것이라고 예측 및/또는 돌연변이 펩타이드가 면역 반응을 트리거할 것으로 예측된다. 모델의 출력은 다른 척도를 가질 수 있음이 자명할 것이어서, 500nM은 예를 들어 [0,1] 척도 상의 다른 값(예를 들어, 0.42)에 대응할 수 있다.
- [0283] 제약상 허용되는 조성물은 선택된 변이-코딩 서열 중 하나 이상 또는 모두를 사용하여 개발 및/또는 제조될 수 있다. 상기 조성물은 선택된 단일 변이-코딩 서열에 대응하는 돌연변이 펩타이드를 포함할 수 있다. 조성물은 다수의 선택된 변이-코딩 서열에 대응하는 돌연변이 펩타이드 및/또는 돌연변이 펩타이드 전구체를 포함할 수 있다. 펩타이드 후보의 서브세트(예를 들어, 5, 10, 15, 20, 30 또는 그 사이의 임의의 수, 가장 높은 제시 예

측과 연관됨)는 추가 전구체 개발을 위해 사용될 수 있다.

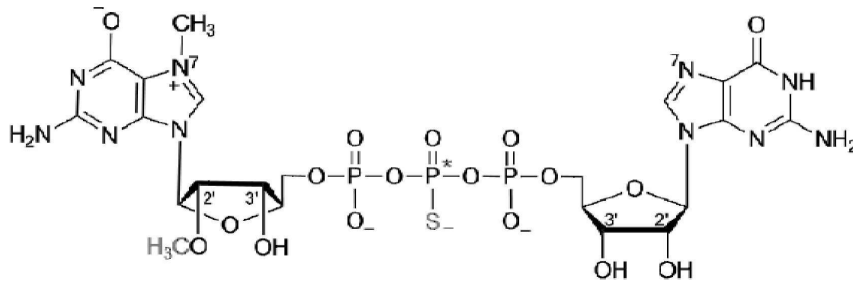
- [0284] 조성물 내의 하나 이상 또는 모든 돌연변이 펩타이드 각각은 예를 들어 약 7 내지 약 40개의 아미노산(예를 들어, 약 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 20, 22, 25, 30, 35, 40, 45, 50, 60 또는 70개 중 임의의 길이의 아미노산)의 길이를 가질 수 있다. 일부 실시예에서, 조성물 내의 하나 이상 또는 모든 돌연변이 펩타이드 각각의 길이는 지정된 범위 내에 있다(예를 들어, 8 내지 11개 아미노산, 8 내지 12개 아미노산 또는 8 내지 15개 아미노산). 일부 실시예에서, 조성물 내의 하나 이상 또는 모든 돌연변이 펩타이드 각각은 길이가 약 8 내지 10개 아미노산이다. 조성물 내의 하나 이상 또는 모든 돌연변이 펩타이드 각각은 분리된 형태일 수 있다. 조성물 내의 하나 이상 또는 모든 돌연변이 펩타이드 각각은 돌연변이 펩타이드의 말단(또는 각각의 말단)에 하나 이상의 펩타이드를 첨가함으로써 생성된 "긴 펩타이드"일 수 있다. 조성물 내의 하나 이상 또는 모든 돌연변이 펩타이드 각각은 태그될 수 있고, 융합 단백질일 수 있고/있거나 하이브리드 분자일 수 있다.
- [0285] 제약상 허용되는 조성물은 하나 이상 또는 모든 선택된 변이-코딩 서열 각각에 대해 변이-코딩 서열에서 식별된 아미노산을 포함하거나 구성되는 펩타이드를 암호화하는 하나 이상의 핵산을 포함하거나 이를 이용함으로써 개발 및/또는 제조될 수 있다. 핵산(들)은 DNA, RNA 및/또는 mRNA를 포함할 수 있다. 임의의 다수의 코돈이 주어진 아미노산을 암호화할 수 있다는 점을 고려하면, 코돈은 예를 들어 주어진 유형의 유기체에서 발현을 최적화하거나 촉진하기 위해 선택될 수 있다. 이러한 선택은 주어진 유형의 유기체에 의해 각각의 다수의 잠재적 코돈이 사용되는 빈도, 주어진 유형의 유기체에서 다수의 잠재적 코돈 각각의 번역 효율 및/또는 다수의 잠재적 코돈 각각을 향한 주어진 유형의 유기체의 편향 정도를 기반으로 할 수 있다.
- [0286] 일부 경우에, 조성물은 앞서 기재된 돌연변이 펩타이드(들) 또는 돌연변이 펩타이드(들)의 전구체를 암호화하는 핵산을 포함할 수 있다. 핵산은 돌연변이 펩타이드(또는 이의 전구체)를 코딩하는 서열의 측면에 있는 서열을 포함할 수 있다. 일부 경우에, 핵산은 둘 이상의 선택된 변이-코딩 서열에 대응하는 항원결정기를 포함한다. 일부 경우에, 핵산은 앞서 기재된 돌연변이 펩타이드 또는 전구체를 암호화하는 폴리뉴클레오타이드 서열을 갖는 DNA이다.
- [0287] 일부 경우, 핵산은 RNA이다. 일부 경우, RNA는 앞서 기재된 돌연변이 펩타이드 또는 전구체를 암호화하는 폴리뉴클레오타이드 서열을 갖는 DNA 주형으로부터 역전사된다. 일부 경우, RNA는 mRNA이다. 일부 경우, RNA는 벌거벗은(naked) mRNA이다. 일부 경우, RNA는 변형 mRNA(예를 들어, 프로타민을 사용하여 분해로부터 보호된 mRNA, 변형 5'CAP 구조를 함유하는 mRNA, 또는 변형 뉴클레오타이드를 함유하는 mRNA)이다. 일부 실시예에서, RNA는 단일 가닥 mRNA이다.
- [0288] 조성물은 앞서 기재된 돌연변이 펩타이드 및/또는 돌연변이 펩타이드를 암호화하는 핵산(들)을 포함하는 세포를 포함할 수 있다. 조성물은 돌연변이 펩타이드 및/또는 돌연변이 펩타이드를 암호화하는 핵산(들)에 대한 하나 이상의 적합한 벡터 및/또는 하나 이상의 전달 시스템을 추가로 포함할 수 있다. 일부 예에서, 돌연변이 펩타이드 및/또는 돌연변이 펩타이드를 암호화하는 핵산을 포함하는 세포는 비인간 세포, 예를 들어 박테리아 세포, 원생동물 세포, 진균 세포 또는 비인간 동물 세포이다. 일부 경우, 돌연변이 펩타이드 및/또는 돌연변이 펩타이드를 암호화하는 핵산을 포함하는 세포는 인간 세포이다. 일부 경우, 인간 세포는 면역 세포이다. 일부 경우, 면역 세포는 항원 제시 세포(APC)이다. 일부 경우, APC는 대식세포, 단핵구, 수지상 세포, B 세포 및 미세아교세포와 같은 전문 APC이다. 다른 경우, 전문 APC는 대식세포 또는 수지상 세포이다. 일부 경우, 돌연변이 펩타이드 및/또는 돌연변이 펩타이드를 암호화하는 핵산 서열(들)을 포함하는 APC가 세포 백신으로 사용되어 CD4+ 또는 CD8+ 면역 반응을 유도한다. 다른 경우에, 세포 백신으로 사용되는 조성물은 돌연변이 펩타이드 및/또는 돌연변이 펩타이드를 암호화하는 핵산 서열(들)을 포함하는 APC에 의해 프라임된 돌연변이 펩타이드 특이적 T 세포를 포함한다.
- [0289] 조성물은 제약상 허용되는 보조제(adjuvant) 및/또는 제약상 허용되는 부형제(excipient)를 포함할 수 있다. 보조제는 조성물에 혼합되어 돌연변이 펩타이드에 대한 면역 반응을 변형시키는 임의의 물질을 말한다. 보조제는 예를 들어 면역 자극제를 사용하여 접합(conjugate)될 수 있다. 부형제는 활동 또는 면역원성을 증가시키고, 안정성을 부여하고, 생물학적 활성을 증가 및/또는 혈청 반감기를 증가시키기 위해 특정 돌연변이 펩타이드의 분자량을 증가시킬 수 있다.
- [0290] 제약상 허용되는 조성물은 백신일 수 있으며, 이는 특정 대상체에 특이적인(예를 들어 잠재적으로 이를 위해 개발된) 개별화된 백신을 포함할 수 있다. 예를 들어, MHC 서열은 특정 대상체로부터의 샘플을 사용하여 식별되었을 수 있고, 그 조성물은 특정 대상체를 위해 개발 및/또는 사용될 수 있다.

- [0291] 백신은 핵산 백신일 수 있다. 핵산은 돌연변이 펩타이드 또는 돌연변이 펩타이드의 전구체를 암호화할 수 있다. 핵산 백신은 돌연변이 펩타이드(또는 이의 전구체)를 코딩하는 서열의 측면에 있는 서열을 포함할 수 있다. 일부 경우, 핵산 백신은 둘 이상의 선택된 변이-코딩 서열에 대응하는 항원결정기를 포함한다. 일부 경우, 핵산 백신은 DNA 기반 백신이다. 일부 경우, 핵산 백신은 RNA 기반 백신이다. 일부 경우, RNA 기반 백신은 mRNA를 포함한다. 일부 경우, RNA 기반 백신은 벌거벗은 mRNA를 포함한다. 일부 경우, RNA 기반 백신은 변형 mRNA(예를 들어, 프로타민을 사용하여 분해로부터 보호된 mRNA, 변형 5'CAP 구조를 함유하는 mRNA, 또는 변형 뉴클레오타이드를 함유하는 mRNA)를 포함한다. 일부 실시예에서, RNA 기반 백신은 단일 가닥 mRNA를 포함한다.
- [0292] 핵산 백신은 차세대 면역요법의 일부로 사용되는 특정 대상체를 위해 제조된 개별화된 신생항원 특이적 요법을 포함할 수 있다. 개별화된 백신은 먼저 특정 대상체의 샘플에서 돌연변이 펩타이드를 검출하고 후속적으로 각각의 검출된 돌연변이 펩타이드에 대해 펩타이드가 특정 대상체의 MHC에 결합할지, MHC에 의해 제시될지, 특정 대상체의 T-세포 수용체에 결합할지, 및/또는 면역학적 반응을 트리거할지 여부 및/또는 정도를 예측함으로써 설계되었을 수 있다. 이들 예측에 기초하여, 검출된 돌연변이 펩타이드의 서브세트(예를 들어, 적어도 1, 적어도 2, 적어도 3, 적어도 5, 적어도 8, 적어도 10, 적어도 12, 적어도 15개, 적어도 18개, 최대 40개, 최대 30개, 최대 25개, 최대 20개, 최대 18개, 최대 15개 및/또는 최대 10개의 돌연변이 펩타이드)가 선택될 수 있다. 각각의 선택된 돌연변이 펩타이드에 대해, 돌연변이 펩타이드를 코딩하는 합성 mRNA 서열이 식별될 수 있다. mRNA 백신은 mRNA-리포플렉스를 형성하기 위해 지질과 복합화된 mRNA(돌연변이 펩타이드의 일부 또는 전부를 암호화함)를 포함할 수 있다. mRNA-리포플렉스를 포함하는 백신을 투여하면 mRNA가 TLR7 및 TLR8을 자극하여 수지상 세포에 의한 T 세포 활성화를 유발할 수 있다. 또한, 투여로 인해 mRNA가 돌연변이 펩타이드로 번역되어 MHC 분자에 결합 및 제시될 수 있고 T 세포 반응을 유도할 수 있다.
- [0293] 조성물은 실질적으로 순수한 돌연변이 펩타이드, 실질적으로 순수한 이의 전구체, 및/또는 돌연변이 펩타이드 또는 이의 전구체를 암호화하는 실질적으로 순수한 핵산을 포함할 수 있다. 조성물은 돌연변이 펩타이드, 이의 전구체, 및/또는 돌연변이 펩타이드 또는 이의 전구체를 암호화하는 핵산을 함유하기 위해 하나 이상의 적합한 벡터 및/또는 하나 이상의 전달 시스템을 포함할 수 있다. 적합한 벡터 및 전달 시스템은 아데노바이러스, 백시니아 바이러스, 레트로바이러스, 헤르페스 바이러스, 아데노 관련 바이러스 또는 하나 이상의 바이러스 요소를 포함하는 하이브리드에 기초한 시스템과 같은 바이러스를 포함한다. 비-바이러스 전달 시스템은 양이온성 지질 및 양이온성 중합체(예를 들어, 양이온성 리포솜)를 포함한다. 일부 실시예에서, '유전자 총(gene-gun)'과 같은 물리적 전달이 사용될 수 있다.
- [0294] 특정 실시예에서, RNA 기반 백신은 5'→3' 방향으로 다음을 포함하는 RNA 분자를 포함한다: (1) 5' 캡, (2) 5' 비번역 영역(UTR), (3) 분비 신호 펩타이드를 암호화하는 폴리뉴클레오타이드 서열, (4) 중앙 표본에 존재하는 암 특이적 체세포 돌연변이로부터 생성된 하나 이상의 돌연변이 펩타이드를 암호화하는 폴리뉴클레오타이드 서열, (5) 주 조직적합성 복합체(MHC) 분자의 막횡단 및 세포질 도메인의 적어도 일부를 암호화하는 폴리뉴클레오타이드 서열, (6) (a) AES(Amino-Terminal Enhancer of Split) mRNA 또는 이의 단편의 3' 비번역 지역, 및 (b) 미토콘드리아로 암호화된 12S RNA의 비-코딩 RNA 또는 이의 단편을 포함하는 3' UTR, 및 (7) 폴리(A) 서열. 이 예시 RNA 분자는 아래 섹션 V와 관련하여 논의된 바와 같이 어텐션-기반 예측 모델의 예시적 구현을 평가하는 데에도 사용되었다.
- [0295] 특정 실시예에서, RNA 분자는 아미노산 링커를 암호화하는 폴리뉴클레오타이드 서열을 추가로 포함하고, 여기서 아미노산 링커 및 하나 이상의 돌연변이 펩타이드 중 첫번째를 암호화하는 폴리뉴클레오타이드 서열은 제1 링커-신생항원결정기 모듈을 형성하고, 제1 링커-신생항원결정기 모듈을 형성하는 폴리뉴클레오타이드 서열은 분비 신호 펩타이드를 암호화하는 폴리뉴클레오타이드 서열과 5'→3' 방향으로 MHC 분자의 막횡단 및 세포질 도메인의 적어도 일부분을 암호화하는 폴리뉴클레오타이드 서열 사이에 있다. 특정 실시예에서, 아미노산 링커는 서열 GGSGGGSGG를 포함한다. 특정 실시예에서, 아미노산 링커를 암호화하는 폴리뉴클레오타이드 서열은 서열 GCGGCUCUGGAGGAGGCGGCCGCGGAGGC를 포함한다.
- [0296] 특정 실시예에서, RNA 분자는 5'→3' 방향으로 다음을 더 포함한다: 적어도 제2 링커-항원결정기 모듈 — 적어도 제2 링커-항원결정기 모듈은 아미노산 링커를 암호화하는 폴리뉴클레오타이드 서열 및 신생항원결정기를 암호화하는 폴리뉴클레오타이드 서열을 포함하고, 여기서 제2 링커-신생항원결정기 모듈을 형성하는 폴리뉴클레오타이드 서열은 제1 링커-신생항원결정기 모듈의 신생항원결정기를 암호화하는 폴리뉴클레오타이드 서열과 5'→3' 방향에서 MHC 분자의 막횡단 및 세포질 도메인의 적어도 일부분을 암호화하는 폴리뉴클레오타이드 서열 사이에 있으며, 여기서 제1 링커-항원결정기 모듈의 신생항원결정기는 제2 링커-항원결정기 모듈의 신생항원결정기와 상이함 — . 특정 실시예에서, RNA 분자는 5개의 링커-항원결정기 모듈을 포함하고, 여기서 5개의 링커-항원

결정기 모듈은 각각 상이한 신생항원결정기를 암호화한다. 특정 실시예에서, RNA 분자는 10개의 링커-항원결정기 모듈을 포함하고, 여기서 10개의 링커-항원결정기 모듈은 각각 상이한 신생항원결정기를 암호화한다. 특정 실시예에서, RNA 분자는 20개의 링커-항원결정기 모듈을 포함하고, 여기서 20개의 링커-항원결정기 모듈은 각각 상이한 신생항원결정기를 암호화한다.

[0297] 특정 실시예에서, RNA 분자는 아미노산 링커를 암호화하는 제2 폴리뉴클레오타이드 서열을 더 포함하고, 여기서 아미노산 링커를 암호화하는 제2 폴리뉴클레오타이드 서열은 3' 방향에서 가장 말단에 있는 신생항원결정기를 암호화하는 폴리뉴클레오타이드 서열과 MHC 분자의 막횡단 및 세포질 도메인의 적어도 일부분을 암호화하는 폴리뉴클레오타이드 서열 사이에 있다.

[0298] 특정 실시예에서, 5' 캡은 하기 구조의 D1 부분입체이성질체를 포함한다:



[0299]

[0300] 특정 실시예에서, 5' UTR은 서열 UUCUUCUGGUCCCCACAGACUCAGAGAGAACCCGCCACC를 포함한다. 특정 실시예에서, 5' UTR은 서열 GCGAACUAGUAUUCUUCUGGUCCCCACAGACUCAGAGAGAACCCGCCACC를 포함한다.

[0301] 특정 실시예에서, 분비 신호 펩타이드는 아미노산 서열 MRVMAPRTLILLISGALALTETWAGS를 포함한다. 특정 실시예에서, 분비 신호 펩타이드를 암호화하는 폴리뉴클레오타이드 서열은 서열 AUGAGAGUGAUGGCCCCAGAACCCUGAUCCUGCUGCUGUCUGGCGCCUGGCCCUGACAGAGACAUGGGCCGGAAGC를 포함한다.

[0302] 특정 구현예에서, MHC 분자의 막횡단 및 세포질 도메인의 적어도 일부분은 아미노산 서열 IVGVAGLAVLAVVIGAVVATVMCRRKSSGGKGSYSQAASSDSAQGSVDVSLTA를 포함한다. 특정 실시예에서, MHC 분자의 막횡단 및 세포질 도메인의 적어도 일부분을 암호화하는 폴리뉴클레오타이드 서열은 서열 AUCGUGGGAUUGUGGCAGGACUGGCAGUGCUGGCCUGGUGGUGAUGCGAGCCGUGGUGGCUACCGUGAUGUGCAGACGGAAGUCCAGCGGAGGCAAGGGC GGCAGCUACAGCCAGGCCGCCAGCUCUGAUAGCGCCAGGGCAGCGACGUGUCACUGACAGCC를 포함한다.

[0303] 특정 실시예에서, AES mRNA의 3' 비번역 영역은 서열 CUGGUACUGCAUGCAGCAAUGCUAGCUGCCCCUUCUCCGUCUGGGUACCCGAGUCUCCCCGACCUGGGUCCCAGGUAGUCUCCACCUCACCUGCC CCACUACCACCUCUGCUAGUCCAGACACCUC를 포함한다. 특정 실시예에서, 미토콘드리아적으로 암호화된 12S RNA의 비암호화 RNA는 서열 CAAGCAGCAGCAAUGCAGCUCAAAACGCUUAGCCUAGCCACACCCCGAGGAAACAGCAGUGAUUAACCUUAGCAAUAAACGAAAGUUUAACUAAGCUA UACUAAACCCAGGGUUGGUCAAUUCUGGCCAGCCACACCG를 포함한다. 특정 실시예에서, 3' UTR은 서열 CUCGAGCUGGUACUGCAUGCAGCAAUGCUAGCUGCCCCUUCUCCGUCUGGGUACCCGAGUCUCCCCGACCUGGGUCCCAGGUAGUCUCCACCUCCA CCUGCCACUCACCACCUCUGCUAGUCCAGACACCUCUCCAGCAGCAAUGCAGCUCAAAACGCUUAGCCUAGCCACACCCCGAGGAAACAGCAG UGAUUAACCUUAGCAAUAAACGAAAGUUUAACUAAGCUAUAACUACCCAGGGUUGGUCAAUUCUGUGCCAGCCACCCGAGACCUGGUCCAGAGUCGCUA GCCGCGUCGCU를 포함한다.

[0304] 특정 실시예에서, 폴리(A) 서열은 120개의 아데닌 뉴클레오타이드를 포함한다.

[0305] 특정 실시예에서, RNA 기반 백신은 5'→3' 방향으로 다음을 포함하는 RNA 분자를 포함한다: 폴리뉴클레오타이드 서열 GCGAACUAGUAUUCUUCUGGUCCCCACAGACUCAGAGAGAACCCGCCACCAUGAGAGUGAUGGCCCCAGAACCCUGAUCCUGCUGCUGUCUGGCGCCUG GCCUGACAGAGACAUGGCGCGAAGC, 종양 표본에 존재하는 암 특이적 체세포 돌연변이로부터 도출된 하나 이상의 돌 연변이 펩타이드를 암호화하는 폴리뉴클레오타이드 서열, 및 폴리뉴클레오타이드 서열 AUCGUGGGAUUGUGGCAGGACUGGCAGUGCUGGCCUGGUGGUGAUGCGAGCCGUGGUGGCUACCGUGAUGUGCAGACGGAAGUCCAGCGGAGGCAAGGGC GGCAGCUACAGCCAGGCCGCCAGCUCUGAUAGCGCCAGGGCAGCGACGUGUCACUGACAGCCUAGUAACUCGAGCUGGUACUGCAUGCAGCAAUGCUAGC UGCCCCUUCUCCGUCUGGGUACCCGAGUCUCCCCGACCUGGGUCCAGGUAGUCUCCACCUCACCUGCCCCACUCACCACCUCUGCUAGUCCAGA CACCUCUCCAGCAGCAGCAAUGCAGCUCAAAACGCUUAGCCUAGCCACACCCCGAGGAAACAGCAGUGAUUAACCUUAGCAAUAAACGAAAGUUUAAC

UAAGCUAUACUAACCCCAGGGUUGGUCAAUUCGUGCCAGCCACACCGAGACCUUGGUCCAGAGUCGCUAGCCGCGUCGCU.

- [0306] 일부 실시예에서, 본 명세서에 기재된 돌연변이 펩타이드(가령, 본 명세서에 기재된 머신 러닝 기법으로부터의 결과에 기초하여 선택된 변이-코딩 서열에 의해 식별된 아미노산의 정렬된 세트를 포함하거나 이로 구성됨)는 돌연변이 펩타이드 특이적 치료제, 가령, 항체 치료제를 제조하는 데 사용될 수 있다. 예를 들어, 돌연변이 펩타이드는 돌연변이 펩타이드를 특이적으로 인식하는 항체를 생성 및/또는 식별하는 데 사용될 수 있다. 이 항체는 치료제로 사용될 수 있다. 합성 짧은 펩타이드는 단백질 반응성 항체를 생성하는 데 사용되었다. 합성 펩타이드로 면역화할 때의 장점은 무제한적인 양의 순수한 안정 항원을 사용할 수 있다는 것이다. 이 접근법은 짧은 펩타이드 서열을 합성하고, 큰 운반체 분자에 연결하며, 펩타이드-캐리어 분자로 대상체를 면역화하는 것을 포함한다. 항체의 특성은 기본 서열 정보에 따라 다르다. 일반적으로 원하는 펩타이드에 대한 좋은 반응은 서열과 커플링 방법을 신중하게 선택하여 생성될 수 있다. 대부분의 펩타이드는 좋은 반응을 이끌어낼 수 있다. 항펩타이드 항체의 장점은 돌연변이 펩타이드의 아미노산 서열을 결정한 직후에 준비될 수 있고 단백질의 특정 영역이 항체 생산을 위해 특이적으로 표적화할 수 있다는 것이다. 머신 러닝 모델이 면역원성을 예측한 돌연변이 펩타이드를 선택 및/또는 이를 선별제거하면 결과 항체가 중앙 환경에서 천연 단백질을 인식할 가능성이 높아질 수 있다. 돌연변이 펩타이드는 예를 들어 15개 이하, 18개 이하 또는 20개 이하, 25개 이하, 30개 이하, 35개 이하, 40개 이하, 50개 이하, 60개 이하, 70개 이하, 85개 이하, 100개 이하, 110개 이하의 잔기일 수 있다. 돌연변이 펩타이드는, 예를 들어 9개 이상, 10개 이상, 15개 이상, 20개 이상, 25개 이상, 30개 이상, 50개 이상 또는 70개 이상의 잔기일 수 있다. 짧은 펩타이드일수록 항체 생산을 향상시킬 수 있다.
- [0307] 펩타이드-운반체 단백질 커플링이 고역가 항체의 생산을 촉진하기 위해 사용될 수 있다. 커플링 방법은 예를 들어 위치 지정 커플링 및/또는 $-NH_2$, $-COOH$, $-SH$ 및 페놀성 $-OH$ 와 같은 아미노산의 반응성 작용기에 의존하는 기술을 포함할 수 있다. 항펩타이드 항체 생성에 사용되는 임의의 적합한 방법은 본 발명의 방법에 의해 식별된 돌연변이 펩타이드와 함께 이용될 수 있다. 이러한 두 가지 알려진 방법은 다중 항원 펩타이드 시스템(MAP) 및 지질 코어 펩타이드(LCP 방법)이다. MAP의 장점은 접합 방법(conjugation method)이 필요하지 않다는 것이다. 운반체 단백질이나 연결 결합은 면역화된 숙주에 도입되지 않는다. 한 가지 단점은 펩타이드의 순도를 제어하기가 더 어렵다는 것이다. 또한 MAP는 일부 숙주의 면역 반응 시스템을 우회할 수 있다. LCP 방법은 다른 항펩타이드 백신 시스템보다 더 높은 역가를 제공하는 것으로 알려져 있으므로 유리할 수 있다.
- [0308] 또한 본 명세서에 개시된 기술을 사용하여 식별된 하나 이상의 돌연변이 펩타이드를 포함하는 분리된 MHC/펩타이드 복합체가 본 명세서에 제공된다. 이러한 MHC/펩타이드 복합체는 예를 들어 항체, 가용성 TCR 또는 TCR 유사체를 식별하기 위해 사용될 수 있다. 이러한 항체의 한 유형은 특정 HLA 환경의 맥락에서 중앙 관련 항원으로 부터의 펩타이드에 결합하는 항체이기 때문에 TCR 모방체(TCR mimics)로 불리운다. 이러한 유형의 항체는 표면에 복합체를 발현하는 세포의 용해를 매개할 뿐만 아니라 복합체를 발현하는 이식된 암 세포주로부터 쥐를 보호하는 것으로 나타났다(예를 들어, Wittman et al., J. of Immunol. 177:4187-4195 (2006)를 참조). IgG mAbs로서의 TCR 모방체의 한 가지 장점은 친화도 성숙이 수행될 수 있고 분자가 현재 Fc 도메인을 통해 면역 이펙터 기능과 커플링된다는 것이다. 이러한 항체는 또한 독소, 사이토카인 또는 의약품과 같은 치료 분자를 중앙에 표적화하는 데 사용될 수 있다.
- [0309] 돌연변이 펩타이드를 이용해 개발된 또 다른 유형의 분자, 가령, 비-하이브리도마 기반 항체 생산 또는 박테리 오파지 상의 항펩타이드 Fab 분자와 같은 결합 적격 항체 단편의 생산을 사용하여 본 발명의 방법을 사용하여 선택된 것이 있다. 이들 단편은 또한 항펩타이드 MHC Fab-면역독소 접합체, 항펩타이드 MHC Fab-사이토카인 접합체 및 항펩타이드 MHC Fab-약물 접합체와 같은 중앙 전달을 위한 다른 치료 분자에 접합될 수 있다.
- [0310] IV. 면역원성 백신 또는 T 세포를 포함하는 치료 방법
- [0311] 일부 실시예는 면역원성 백신일 수 있는 백신을 포함하는 치료 방법을 제공한다. 일부 실시예에서, 질병(가령, 암) 치료 방법이 제공되며, 상기 방법은 본 명세서에 기재된 유효량의 조성물, 본 명세서에 개시된 기법을 사용하여 식별된 돌연변이 펩타이드, 이의 전구체 또는 본 명세서에 기재된 기법을 이용해 식별된 돌연변이 펩타이드(또는 전구체)를 암호화하는 핵산을 개체에게 투여하는 것을 포함할 수 있다.
- [0312] 일부 실시예에서, 질병(가령, 암)의 치료 방법이 제공된다. 방법은 대상체로부터 샘플(가령, 혈액 샘플)을 수집하는 단계를 포함할 수 있다. T 세포가 분리되고 자극될 수 있다. 분리는 예를 들어 밀도 구배 침강(가령, 그리고 원심분리), 면역자기 선택 및/또는 항체-복합체 필터링을 사용하여 수행될 수 있다. 자극은 예를 들어 미토겐(예를 들어, PHA 또는 Con A) 또는 항-CD3 항체(예를 들어, CD3에 결합하고 T-세포 수용체 복합체를 활성화하기 위한 것) 및 항 CD28 항체(가령, CD28에 결합하고 T 세포를 자극하기 위한 것)를 사용할 수 있는 항원 독립

적 자극을 포함할 수 있다. (본 명세서에 개시된 하나 이상의 기법에 따라, 돌연변이 펩타이드의 세트의 각각이 개체의 MHC 분자에 결합할지, 개체의 MHC 분자에 의해 제시될지, 및/또는 개체 내 면역 반응을 트리거할지 여부 및/또는 정도에 대한 예측에 대응하는 머신 러닝 모델에 의해 생성된 결과에 기초하여) 하나 이상의 돌연변이 펩타이드가 대상체의 치료에서 사용되기 위해 선택될 수 있다(또는 선택되었을 수 있다). 하나 이상의 돌연변이 펩타이드는 대상체와 연관된 하나 이상의 서열 표현(가령, MHC 서열, 변이-코딩 서열의 세트 및/또는 또는 T-세포 수용체 서열의 표현)을 식별하고 처리하는 것을 포함하는 본 명세서에 개시된 기법에 기초하여 선택되었을 수 있다. 하나 이상의 서열은 T 세포가 분리된 샘플 또는 상이한 샘플을 사용하여 검출되었을 수 있다.

[0313] 일부 예에서, 하나 이상의 돌연변이 펩타이드(또는 이의 전구체)는 돌연변이 펩타이드(예를 들어, 신생항원) 특이적 T 세포를 생성하기 위해 사용될 수 있다. 예를 들어, 말초 혈액 T 세포는 대상체로부터 분리되고 하나 이상의 돌연변이 펩타이드와 접촉하여 대상체에게 투여될 수 있는 돌연변이 펩타이드 특이적 T 세포 집단을 유도할 수 있다. 일부 예에서, 돌연변이 펩타이드-반응성 T 세포의 T 세포 수용체 서열이 시퀀싱될 수 있다. 시퀀싱이 정렬된 핵산 세트를 식별하는 경우, 핵산의 각각의 코돈은 아미노산으로 (가령, 조회(look-up) 기법을 통해) 번역될 수 있다. T 세포 수용체 서열(가령, 아미노산 T 세포 수용체 서열)이 얻어지면, 돌연변이 펩타이드를 특이적으로 인식하는 T 세포 수용체를 포함하도록 T 세포를 가공할 수 있다. 그런 다음 이러한 가공된 T 세포가 대상체에게 투여될 수 있다. 예를 들어, 그 전문이 모든 목적으로 본 명세서에 포함되는, Matsuda et al. "Induction of Neoantigen-Specific Cytotoxic T Cells and Construction of T-cell Receptor Engineered T Cells for Ovarian Cancer," Clin. Cancer Res. 1-11 (2018)를 참조할 수 있다. 본 명세서에 제공된 임의의 방법에서, T 세포는 대상체에게 투여되기 전에 생체내 및/또는 생체외에서 확장될 수 있다. 그런 다음 대상체는 확장된 T 세포 집단을 포함하는 조성물을 투여(예를 들어, 주입) 받을 수 있다.

[0314] 일부 예에서, 질병(가령, 암)에 대한 치료 방법이 제공되며, 상기 방법은 하나 이상의 돌연변이 펩타이드(또는 이의 하나 이상의 전구체)를 유효량으로, 가령, 생체내에서 T 세포를 프라이밍, 활성화, 확장하기에 유효한 양으로 포함하는 조성물을 개체에게 투여하는 것을 포함할 수 있다.

[0315] 일부 실시예에서, 질병(가령, 암)에 대한 치료 방법이 제공되며, 상기 방법은 본 명세서에 기재된 기법을 사용하여 선택된 돌연변이 펩타이드의 전구체를 포함하는 유효량의 조성물을 개체에게 투여하는 것을 포함할 수 있다. 일부 실시예에서, 면역원성 백신이 본 명세서에 기재된 기법을 사용하여 선택된 제약상 허용되는 돌연변이 펩타이드를 포함할 수 있다. 일부 실시예에서, 면역원성 백신은 본 명세서에 기재된 기법을 사용하여 선택된 돌연변이 펩타이드(예를 들어, 단백질, 펩타이드, DNA 및/또는 RNA)에 대한 제약상 허용되는 전구체를 포함할 수 있다. 일부 실시예에서, 질병(가령, 암)에 대한 치료 방법이 제공되며, 상기 방법은 본 명세서에 기재된 기법을 사용하여 선택된 돌연변이 펩타이드를 특이적으로 인식하는 유효량의 항체를 개체에게 투여하는 것을 포함할 수 있다. 일부 실시예에서, 질병(가령, 암)에 대한 치료 방법이 제공되며, 상기 방법은 본 명세서에 기재된 기법을 사용하여 선택된 돌연변이 펩타이드를 특이적으로 인식하는 유효량의 가용성 TCR 또는 TCR 유사체를 개체에게 투여하는 것을 포함할 수 있다.

[0316] 일부 실시예에서, 암은 다음 중 어느 하나이다: 암종, 림프종, 모세포종, 육종, 백혈병, 편평 세포암, 폐암(가령, 소세포 폐암, 비-소세포 폐암, 폐의 선암종, 및 폐의 편평 암종), 복막암, 간세포암, 위암 또는 위암(위장암 포함), 췌장암, 교모세포종, 자궁경부암, 난소암, 간암, 방광암, 간종, 유방암, 결장암, 흑색종, 자궁내막 또는 자궁 암종, 타액선 암종, 신장 또는 신세포암, 간암, 전립선암, 외음부암, 갑상선암, 간암종, 두경부암, 결장직장암, 직장암, 연조직 육종, 카포시 육종, B 세포 림프종(가령, 저등급/여포성 비호지킨 림프종(NHL), 소림프구성(SL) NHL, 중급/여포성 NHL, 중급 미만성 NHL, 고등급 면역모세포성 NHL, 고등급 림프구성 NHL, 고등급 작은 비절단 세포 NHL, 거대 NHL, 맨틀 세포 림프종, AIDS 관련 림프종, 및 발렌스트롬 마크로글로불린혈증), 만성 림프구성 백혈병(CLL), 급성 림프구성 백혈병(ALL), 골수종, 털 세포 백혈병, 만성 골수모세포 백혈병, 및 이식 후 림프증식성 장애(PTLD), 수정체종과 관련된 비정상 혈관증식, 부종(가령, 뇌종양과 연관된 것), 및 메이그 증후군.

[0317] 본 명세서에 개시된 실시예는 개별화-의료 전략의 일부 또는 전부를 식별 및/또는 구현하는 것을 포함할 수 있다. 예를 들어, 개체로부터의 샘플을 사용하여 MHC 서열 및/또는 변이-코딩 서열의 세트를 결정하고; 본 명세서에 개시된 머신 러닝 모델(예를 들어, 어텐션-기반 머신 러닝 모델)을 사용하여 MHC 서열 및 변이-코딩 서열의 표현을 처리함으로써, 하나 이상의 돌연변이 펩타이드가 백신에서 사용되도록 선택될 수 있다. 그런 다음 하나 이상의 돌연변이 펩타이드(및/또는 이의 전구체)가 동일 개체에 투여될 수 있다.

[0318] 일부 실시예에서, 개체의 질병(가령, 암)을 치료하는 방법이 제공되며, 상기 방법은: a) (본 명세서에 개시된

하나 이상의 기법에 따라, 돌연변이 펩타이드의 세트의 각각이 개체의 MHC 분자에 결합할지, 개체의 MHC 분자에 의해 제시될지, 및/또는 개체 내 면역 반응을 트리거할지 여부 및/또는 정도에 대한 예측에 대응하는 머신 러닝 모델에 의해 생성된 결과에 기초하여) 개체의 하나 이상의 돌연변이 펩타이드를 식별하는 단계, b) 식별된 돌연변이 펩타이드(들) 또는 돌연변이 펩타이드(들)의 전구체 또는 식별된 펩타이드(들) 또는 펩타이드 전구체(들)를 암호화하는 핵산(들)(가령, 폴리뉴클레오타이드 가령, DNA 또는 RNA)을 합성하는 단계, 및 c) 돌연변이 펩타이드(들), 돌연변이-펩타이드 전구체(들) 또는 핵산(들)을 개체에게 투여하는 단계를 포함한다.

[0319] 일부 실시예에서, 개체의 질병(가령, 암)을 치료하는 방법이 제공되며, 상기 방법은: a) (본 명세서에 개시된 하나 이상의 기법에 따라, 돌연변이 펩타이드의 세트의 각각이 개체의 MHC 분자에 결합할지, 개체의 MHC 분자에 의해 제시될지, 및/또는 개체 내 면역 반응을 트리거할지 여부 및/또는 정도에 대한 예측에 대응하는 머신 러닝 모델에 의해 생성된 결과에 기초하여) 개체의 하나 이상의 돌연변이 펩타이드를 식별하는 단계, b) 식별된 돌연변이 펩타이드(들) 또는 돌연변이 펩타이드(들)의 하나 이상의 전구체를 암호화하는 핵산(가령, 폴리뉴클레오타이드, 가령, DNA 또는 RNA)의 세트를 식별하는 단계, c) 핵산의 세트를 개체에게 투여하는 단계를 포함한다.

[0320] 일부 실시예에서, 개체의 질병(가령, 암)을 치료하는 방법이 제공되며, 상기 방법은: a) (본 명세서에 개시된 하나 이상의 기법에 따라, 돌연변이 펩타이드의 세트의 각각이 개체의 MHC 분자에 결합할지, 개체의 MHC 분자에 의해 제시될지, 및/또는 개체 내 면역 반응을 트리거할지 여부 및/또는 정도에 대한 예측에 대응하는 머신 러닝 모델에 의해 생성된 결과에 기초하여) 개체의 하나 이상의 돌연변이 펩타이드를 식별하는 단계, b) 돌연변이 펩타이드를 특이적으로 인식하는 항체를 생성하는 단계, 및 c) 펩타이드를 개체에게 투여하는 단계를 포함한다.

[0321] 본 명세서에 제공된 방법은 암으로 진단되었거나 암이 있는 것으로 의심되는 개체(예를 들어, 인간)를 치료하는 데 사용될 수 있다. 일부 실시예에서, 개체는 인간일 수 있다. 일부 실시예에서, 개체는 18, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 또는 85세 중 어느 하나일 수 있다. 일부 실시예에서, 개체는 남성일 수 있다. 일부 실시예에서, 개체는 여성일 수 있다. 일부 실시예에서, 개체는 수술을 거부했을 수 있다. 일부 실시예에서 개체는 의학적으로 수술이 불가능할 수 있다. 일부 실시예에서, 개체는 Ta, Tis, T1, T2, T3a, T3b, 또는 T4의 임상 단계에 있을 수 있다. 일부 실시예에서, 암은 재발성일 수 있다. 일부 실시예에서, 개체는 암과 연관된 하나 이상의 증상을 나타내는 인간일 수 있다. 일부 실시예에서, 개체는 유전적으로 또는 그 밖의 다른 식으로 암 발생에 소인(예를 들어, 위험 인자를 가짐)을 가질 수 있다.

[0322] 본 명세서에 제공된 방법은 수술후 보조 환경(adjuvant setting)에서 실시될 수 있다. 일부 실시예에서, 상기 방법은 선행보조 환경(neoadjuvant setting)에서 실시되는데, 즉 상기 방법은 1차/최종 요법 전에 수행될 수 있다. 일부 실시예에서, 상기 방법은 이전에 치료를 받은 개체를 치료하기 위해 사용된다. 본 명세서에 제공된 임의의 치료 방법은 이전에 치료를 받은 적이 없는 개체를 치료하는 데 사용될 수 있다. 일부 실시예에서, 상기 방법은 1차 요법으로 사용된다. 일부 실시예에서, 상기 방법은 2차 요법으로 사용된다.

[0323] 일부 실시예에서, 개체에서 기존의 암 종양 전이(예컨대, 폐 전이 또는 림프절로의 전이)의 발생률 또는 부담을 감소시키는 방법이 제공되며, 상기 방법은 본 명세서에 개시된 유효량의 조성물을 개체에게 투여하는 단계를 포함한다. 일부 실시예에서, 개체의 암의 질병 진행까지의 시간을 연장시키는 방법이 제공되며, 상기 방법은 본 명세서에 개시된 유효량의 조성물을 개체에게 투여하는 단계를 포함한다. 일부 실시예에서, 암을 앓는 개체의 생존을 연장시키는 방법이 제공되며, 상기 방법은 암을 앓는 개체에게 본 명세서에 개시된 유효량의 조성물을 투여하는 단계를 포함한다.

[0324] 일부 실시예에서, 적어도 하나 이상의 화학요법제가 본 명세서에 개시된 조성물에 더하여 투여될 수 있다. 일부 실시예에서, 하나 이상의 화학요법제는 (반드시 그런 것은 아니지만) 상이한 부류의 화학요법제에 속할 수 있다.

[0325] 일부 실시예에서, 개체의 질병(가령, 암)을 치료하는 방법이 제공되며, 상기 방법은 a) 본 명세서에 개시된 백신(예를 들어, 본 명세서에 개시된 머신 러닝 기법에 기초하여 선택된 돌연변이 펩타이드 또는 이의 전구체를 포함함), 및 b) 면역조절제를 투여하는 단계를 포함한다. 일부 실시예에서, 개체의 질병(가령, 암)을 치료하는 방법이 제공되며, 상기 방법은 a) 본 명세서에 개시된 백신(예를 들어, 본 명세서에 개시된 머신 러닝 기법에 기초하여 선택된 돌연변이 펩타이드 또는 이의 전구체를 포함함), 및 b) 관문 단백질의 길항제를 투여하는 단계를 포함한다. 일부 실시예에서, 개체의 질병(가령, 암)을 치료하는 방법이 제공되며, 상기 방법은 a) 본 명세서에 개시된 백신(예를 들어, 본 명세서에 개시된 머신 러닝 기법에 기초하여 선택된 돌연변이 펩타이드 또는 이의 전구체를 포함함), 및 b) 세포예정사 1(PD-1)의 길항제, 가령, 항-PD-1를 투여하는 단계를 포함한다. 일부 실시예에서, 개체의 질병(가령, 암)을 치료하는 방법이 제공되며, 상기 방법은 a) 본 명세서에 개시된 백신(예

를 들어, 본 명세서에 개시된 머신 러닝 기법에 기초하여 선택된 돌연변이 펩타이드 또는 이의 전구체를 포함함), 및 b) 리간드예정사 1(PD-L1)의 길항제, 가령, 항-PD-L1를 투여하는 단계를 포함한다. 일부 실시예에서, 개체의 질병(가령, 암)을 치료하는 방법이 제공되며, 상기 방법은 a) 본 명세서에 개시된 백신(예를 들어, 본 명세서에 개시된 머신 러닝 기법에 기초하여 선택된 돌연변이 펩타이드 또는 이의 전구체를 포함함), 및 b) 세포독성 T-림프구 관련 단백질 4(CTLA-4)의 길항제, 가령, 항-CTLA-4를 투여하는 단계를 포함한다.

[0326] 다양한 개시가 아미노산 서열의 사용을 언급하고 있음을 이해할 것이다. 핵산 서열이 추가로 또는 대안으로 사용될 수 있다. 예를 들어, 질병 특이적 샘플은 대응하는 질병 특이적이지 않은 샘플(예를 들어, 동일한 대상체 또는 상이한 대상체)에 존재하지 않는 핵산 서열의 세트를 식별하기 위해 시퀀싱될 수 있다. 마찬가지로, MHC 분자 및/또는 T-세포 수용체의 핵산 서열이 추가로 식별될 수 있다. 핵산 질병 특이적 핵산 서열 및 MHC 분자 (또는 T-세포 수용체)의 각각의 표현은 본 명세서에 기재된 어텐션-기반 모델에 의해 처리될 수 있다(예를 들어, 그리고 핵산 서열의 표현을 이용해 훈련되었을 수 있다).

[0327] V. 예시

[0328] V.A. 개요

[0329] 예시적인 펩타이드-MHC(MHC 클래스 I) 어텐션-기반 머신 러닝 모델(이하 "P-MHC-I 모델") 및 예시적인 펩타이드-MHC(MHC 클래스 II) 어텐션-기반 머신 러닝 모델(이하 "P-MHC-II 모델")(총칭하여 그리고 개별적으로 본 명세서에서 P-MHC 모델로 지칭됨)이 개발되었다. 이들 모델은 도 1의 머신 러닝 모델(132)에 대한 구현예이다. P-MHC-I 모델 아키텍처 및 P-MHC-II 모델 아키텍처 모두는 도 3 및 도 4a에 도시된 아키텍처에 대응하여 구현되었다.

[0330] P-MHC 모델은 개별화된 암 백신 개발에서 신생항원 제시를 예측하기 위한 예시적인 어텐션-기반 딥 러닝 모델이다. P-MHC 모델은 N-플랭크 서열, 펩타이드 서열, 및 MHC 서열(MHC 의사서열)을 입력으로서 수신하고 제시 또는 용리된 리간드(EL) 점수를 출력한다. 자연 발생 아미노산의 공간에 걸쳐 있는 어휘가 구축되어 아미노산 서열을 나타내기 위해 토큰화되었다. 입력된 아미노산 서열은 문자로 토큰화되었으며, 각각은 고유한 문자로 표현된다. 모델은 입력 N-플랭크 서열과 펩타이드 서열을 6개의 MHC 대립유전자 중 하나와 쌍을 만들고 6 쌍의 상호작용이 P-MHC-I 모델로 피드포워드되었으며, 특정 결합 MHC 대립유전자를 선택하기 위해, 12개의 MHC 동종이형 중 하나와 12 쌍의 상호작용이 P-MHC-II로 피드포워드되었다.

[0331] 따라서, P-MHC 모델은 내부적으로 다중 대립형질 데이터의 디콘볼루션을 수행한다. 펩타이드-MHC 상호작용 출력을 용리할 가능성이 가장 높은 출력은 0과 1 사이의 값으로 정규화되고 이진 교차-엔트로피 손실 함수(binary cross-entropy loss function)를 사용하여 실제 제시 값과 비교하여 모델 파라미터 조정을 위한 오차를 생성한다. 과적합을 방지하고 모델 강건성을 높이기 위해 P-MHC 모델은 모델 훈련에서 앙상블 방법을 사용한다.

[0332] 이전에 사용 가능한 다른 모델(가령, NetMHCpan-4.0(이하 "모델 A"), P-MHC-I 모델에 대한 Immune Epitope Database and Analysis Resource (IEDB) v2.13(이하, "모델 B") 및 P-MHC-II 모델에 대한 NetMHCIIpan-4.1(이하 "모델 C"))와 비교되는 P-MHC-I 모델 및 P-MHC-II 모델의 훈련 및 성능에 대응하는 예시적인 결과 및 통계. P-MHC-I 및 P-MHC-II 모델은 펩타이드 제시에 대해 다른 모델보다 일관되게 성능이 우수했고 P-MHC-I 모델은 CD8 T 세포 반응 예측에 대해 다른 모델보다 성능이 우수했다. P-MHC 모델은 적어도 다중대립유전자 데이터에서 펩타이드-MHC 쌍의 디콘볼루션을 수행하고 단일대립유전자 및 다중대립유전자 형식 모두에서 증강된 훈련 데이터에 대해 쉽게 훈련될 수 있기 때문에 성능이 더 우수하다.

[0333] V.B. 물질 및 방법

[0334] V.B.1. P-MHC 모델 훈련 — 번역펩티도믹스 데이터

[0335] P-MHC 모델을 훈련하기 위한 번역펩티도믹스 데이터 세트를 구축하기 위해 질량 분석 실험으로부터의 펩타이드 용리 데이터가 사용되었다. 이 데이터에는 세포주, 조직 샘플 및 PBMC 기증자의 다중 대립유전자 데이터 및 단일대립유전자 펩타이드 용리 데이터를 포함하는 비공개 및 공개 데이터 세트가 혼합되어 있다.

[0336] V.B.1.a. 제시-라벨링된 데이터

[0337] **양성 세트(EL=1).** 각각의 배치(batch)에 대해, 양성 펩타이드-MHC(가령, 펩타이드-HLA) 쌍을 다음과 같은 방식으로 처리되었다:

[0338] 1) 펩타이드가 인간 단백질체에 정렬되었다.

- [0339] 2) 각각의 펩타이드에 대해 최대 10개 아미노산 길이의 플랭킹 서열이 N-말단 및 C-말단 위치에 유지되었다.
- [0340] 3) 다수의 유전자에 매핑된 펩타이드가 다운스트림 분석에서 제거되었다. 이러한 펩타이드는 EL=1 세트에 참여하지 않았다. (EL=0 펩타이드는 EL=1 펩타이드의 증거가 있는 단백질로부터만 생성되었기 때문에 이러한 제한은 EL=0에 부과되지 않았다.). 48,329개의 클래스 I 펩타이드가 이 기준에 의해 걸러졌다. 이는 큰 숫자이지만 음성 세트에 대한 신뢰도를 높인다.
- [0341] 4) 동일한 유전자에 매핑되지만 플랭킹 서열이 다른 펩타이드도 다운스트림 분석에서 제거되었다. 이로써 11,443개의 클래스 I 펩타이드가 추가로 제거되었다.
- [0342] 5) 번역 후 변형(PTM)을 포함하는 펩타이드도 다운스트림 분석에서 제거되었다. 7,080개의 Class I 펩타이드가 제거되었다.
- [0343] **음성 세트(EL=0).** 음성 펩타이드-MHC(가령, 펩타이드-HLA) 쌍은 컴퓨터를 통해 생성되었다. 각각의 대립유전자에 대해, 양성 세트(EL=1)의 각각의 기원 단백질에 대해, MHC 클래스 I에 대해 길이 8-11, MHC 클래스 II에 대해 8-30의 모든 가능한 펩타이드 단편이 생성되었는데 각각의 길이에 대해 균일한 확률로 생성되었다. N-말단 및 C-말단 플랭킹 서열도 최대 10개 아미노산 길이로 유지되었다. EL=1 데이터에 나타난 모든 펩타이드-유전자형 쌍이 EL=0 데이터에서 제거되었다. 추가로, MHC 클래스 II에 대해 구축된 데이터세트의 경우, EL=1 펩타이드(동일한 유전자형과 쌍을 이룸)에서 찾을 수 있는 임의의 길이 9의 부분서열을 갖는 펩타이드-유전자형 쌍이 제거된다.
- [0344] V.B.1.b. 벤치마크 데이터 세트
- [0345] 섹션 VI.B.1.a에서 논의된 위의 EL 데이터를 훈련, 검증 및 테스트 세트로 분할하여 벤치마크 데이터 세트가 생성됐다. 훈련 및 검증 세트는 P-MHC 모델 훈련에 사용된 반면, 테스트 세트는 훈련에 명시적으로 사용되지 않았으며 모델의 성능을 정량화하는 데만 사용되었다. MHC 클래스 I 데이터의 경우, 단일 대립유전자 데이터가 사용되어 각각의 대립유전자에 대한 단일대립유전자 데이터에서 10%의 펩타이드를 유지하여 테스트 데이터 세트를 생성했다. MHC 클래스 II 데이터의 경우, 모든 데이터, 다중 대립유전자 및 단일 대립유전자가 테스트/검증 데이터 세트를 생성하는 데 사용된다.
- [0346] 데이터세트의 특징은 다음과 같다: 모든 펩타이드 길이는 클래스 I의 경우 [8,14] 아미노산 범위로, 클래스 II의 경우 [8,30] 아미노산 범위로 제한되었다. 모든 펩타이드는 주 서열(즉, 항원결정기) 및 플랭킹 서열에서 표준 아미노산을 포함하도록 제한되었다. 모든 대립유전자 이름은 MHC 단백질 내의 다음의 아미노산 위치에 의해 정의된 34개의 아미노산 부분서열로 대체되었다: (7, 9, 24, 45, 59, 62, 63, 66, 67, 69, 70, 73, 74, 76, 77, 80, 81, 84, 95, 97, 99, 114, 116, 118, 143, 147, 150, 152, 156, 158, 159, 163, 167, 171), 또는 알파 및 베타 MHCII 단백질 내 위치:
- [0347] 알파: 9,11,22,24,31,52,53,58,59,61,65,66,68,72,73; 및
- [0348] 베타: 9,11,13,26,28,30,47,57,67,70,71,74,77,78,81,85,86,89,90.
- [0349] 이들 위치는 MHC-I/II 단백질이 펩타이드와 접촉하는 결합 포켓 내 위치로 이전에 기재된 바 있다. 데이터 포인트에 대한 고유한 부분서열의 세트가 이하에서 '의사유전자형(pseudoGenotype)'으로 지칭될 수 있다. 일부 경우, 다수의 대립유전자 이름이 동일한 34개 아미노산 부분서열을 특징으로 할 수 있다. 이들 대립유전자는 어텐션-기반 P-MHC 모델을 훈련하기 위해 동일한 것으로 간주되었다. 모든 빈 플랭킹 서열(단백질의 끝에 매핑되는 펩타이드)에는 특별한 아미노산 문자 "\$"가 할당되었다. NA를 "적용할 수 없음"으로 해석하는 특정 프로그래밍 언어로 인해 플랭킹 서열이 아미노산 알파벳에서 'NA'로 읽히는 6개의 데이터 포인트가 고려 대상에서 제외되었다.
- [0350] 훈련/검증/테스트 분할은 다음과 같은 방식으로 수행되었다:
- [0351] **EL = 1인 경우:** 각각의 처리 배치(각 배치는 데이터 세트의 원래 소스를 기반으로 함)에 대해, 단일대립유전자 데이터가 70/20/10의 비율로 훈련/검증/테스트 그룹 간에 랜덤하게 분할되었다. MHC 클래스 II의 경우, 정확한 유전자형이 매핑되는 펩타이드에 대한 훈련/검증/테스트 데이터 세트 간에 펩타이드 서열의 어떠한 길이 9 부분 서열도 겹치지 않음이 보장된다. 단일대립유전자 데이터는 전체 데이터세트에 걸쳐 각각 111(39)개의 고유한 MHC 클래스 I(MHC 클래스 II) 대립유전자를 나타내는 105(41)개의 고유한 부분서열로 구성된다. 모든 다중 대립유전자 데이터는 클래스 I 데이터 세트에 대한 훈련에 전적으로 사용되었다. 다중 대립유전자 데이터는 전체 데이터 세트에 걸쳐 126(76)개의 고유한 MHC 클래스 I(MHC 클래스 II) 유전자형으로 구성된다. 처리 배치 전체의

데이터가 결합되고 중복되는 {펩타이드, nFlank, cFlank, mhc0, mhc1, mhc2, mhc3, mhc4, mhc5}(MHC 클래스 I) 및 {펩타이드, nFlank, cFlank, mhc_dq1_1, mhc_dq1_2, mhc_dq1_3, mhc_dq1_4, mhc_dp1_1, mhc_dp1_2, mhc_dp1_3, mhc_dp1_4, mhc_dr1_1, mhc_dr1_2, mhc_dr3_1, mhc_dr3_2, mhc_dr4_1, mhc_dr4_2, mhc_dr5_1, mhc_dr5_2}(MHC 클래스 II) 튜플이 제거되었다.

[0352] **EL = 0에 대해:** EL = 0에 대해: 각각의 처리 배치에 대해, 각각의 {펩타이드, 의사유전자형} 쌍에 대해, 음성 펩타이드 데이터가 훈련 및 검증 그룹에서 EL=1 데이터와 함께 1:1 비로 샘플링되었다. 테스트 그룹에서, MHC 클래스 I에 대해 1:99 비, MHC 클래스 II는 1:9 비로 샘플링되었다. 처리 배치 전체의 데이터가 결합되었고 중복 관찰이 제거되었다. 이는 결과적으로 테스트 데이터의 1.71%가 MHC 클래스 I에 대해 양성으로(1% 대신), MHC 클래스 II에 대해 11.15%(10% 대신)가 되었다.

[0353] '의사유전자형(pseudoGenotype)'의 다중 부분서열, 즉 다중 대립유전자형 데이터를 사용한 관찰의 경우, 각각의 대립유전자에 대한 양성 펩타이드를 제거하여 음성 펩타이드가 생성된 다음 소스 단백질에서 랜덤 펩타이드가 선택되었다.

[0354] V.B.1.c. 벤치마크 QC

[0355] 이하의 다운스트림 QC 절차를 따라 데이터에 중복이 없음이 확인되었다: 1) 펩타이드, nFlank 및 cFlank 서열에서 표준 아미노산만 허용되며, 2) {nFlank, peptide, cFlank, 의사유전자형} 튜플의 각각의 세트는 고유하고, 3) EL=1 및 EL=0 세트에서 {nFlank, 펩타이드, cFlank, 의사유전자형} 튜플의 중첩이 없다. MHC 클래스 II의 경우, 동일한 의사유전자형을 갖는 펩타이드에 대해 EL=1과 EL=0 사이의 펩타이드 서열 내의 길이 9 부분서열 사이에 중복이 없음이 추가로 보장된다.

[0356] MHC(HLA) 의사유전자형의 수는 대립유전자의 수와 다를 수 있는데, 그 이유는 대립유전자 이름이 (2-필드 분해능, 즉 4자리 분해능에서의) 다른 일부 대립유전자가 동일한 의사유전자형을 가질 수 있기 때문이다.

[0357] V.B.2. P-MHC-I 모델 성능을 평가하기 위한 면역원성 데이터세트

[0358] P-MHC-I 모델의 성능을 평가하기 위해, 2개의 상이한 데이터세트가 사용되었다. 제1 테스트 면역원성 데이터세트의 경우, 종양학 대상체가 자신의 DNA를 시퀀싱하고 표준 P-MHC 결합으로부터 IEDPv.2.13BA를 사용하여 예측이 수행되어 MHC에 의해 제시 및 결합되는 신생항원을 예측할 수 있다. 따라서 예측된 신생항원은 그들의 발현, 변이 대립유전자 빈도, 및 종양 조직에서의 클론성을 사용하여 추가로 우선순위화되었다. 대상체는 이후에 위에서 소개한 바와 같이 RNA 백신을 투여받았다. RNA 백신에 도입된 신생항원에 대한 T 세포 반응은 멀티머(multimer) 및 ELISPOT 어세이를 사용하여 투약된 대상체에서 모니터링되었다. 이들 어세이에서 여러 대조군을 사용하여 기술적인 인공물로 생각되는 T 세포 반응이 제거되었다. 제2 테스트 면역원성 데이터세트에서 TESLA(Tumor Neoantigen Selection Alliance) 컨소시엄에 의해 식별된 관문 차단 요법(RNA 백신 요법이 아님)을 받는 종양학 대상체로부터 시퀀싱 데이터가 획득되었다. P-MHC 결합 예측은 NetMHCcons 1.0을 사용하여 MHC에 의해 제시되거나 결합된 신생항원을 예측하기 위해 수행되었다. 면역원성 어세이는 P-MHC-I 모델에 의해 예측된 신생항원에 대해 실행되었으며 P-MHC-I 모델의 성능을 평가하는 데 사용되었습니다.

[0359] V.B.2.a. 투약 대상체 멀티머 어세이

[0360] 제1 테스트 면역원성 데이터세트의 경우, 펩타이드-MHC 멀티머에 의한 CD8 T 세포의 검출에 대한 양성 또는 음성 결과에 대해 멀티머 어세이 데이터가 평가되었다. 보수적 기준이 긍정적인 결과를 선언하는 데 사용되었다: 구체적으로 이중 사랑체 양성 CD8 T 세포 수가 0.05%보다 큰지 여부. 일부 신생항원결정기는 신생항원결정기 특이적 CD8 T 세포가 0.05% 미만임에도 불구하고 T 세포 표현형 검사가 T 세포 반응을 강력하게 시사하는 경우 양성으로 불렸다. 멀티머 어세이 데이터로부터, 1318개의 신생항원결정기가 음성으로 판정되었으며, 보수적 기준에 따라 이들 중 일부는 위음성일 것으로 예상된다. 27개의 신생항원결정기-HLA 쌍은 백신접종 후 양성으로 선언되었고(신생 반응이라고 함) 20 쌍은 기존의 CD8 T 세포 반응으로 선언되었다.

[0361] V.B.2.b. 투약 대상체 ELISpot 어세이

[0362] 또한 제1 테스트 면역원성 데이터세트에 대해 ELISpot 데이터가 수집됐다. 펩타이드 재자극이 없는 음성 대조군의 스팟 카운트 및 펩타이드 재자극이 있는 테스트 사례에 대해 통계적 평가가 수행되어 양성 콜을 선언하고(순열 접근 방식 사용) 수동으로 추가 검증하여 주어진 대상체 방문에 대한 신생항원의 면역원성에 대한 양성 또는 음성 결과를 할당했다. 신생항원은 치료전 또는 치료후 여부에 상관없이 대상체 방문 중 임의의 대상체 방문에서 양성 결과를 나타내면 ELISpot 분석에서 양성으로 선언되었다. 신생항원은 다음 기준에 따라 신항원을 추가

로 필터링됐다: (a) 심사관이 어세이 결과 값은 'NA'가 아니라고 결정했음, (b) 평가된 P-MHC-1 스코어링 방법 (P-MHC-I, 모델 A, 모델 B) 중 어느 것도 신생항원에 'NA' 값을 할당하지 않았음, 및 (c) 풀링된 신생항원 중 재자극에 사용된 것은 고려 대상에서 제외됨.

[0363] 모든 필터링 단계 후, ELISpot 어세이에서 평가된 각각의 세포 유형에 대한 양성(면역원성) 및 음성(비면역원성) 신생항원의 분포가 아래에 나와 있다. Assay.value_binary=TRUE는 면역원성 신생항원을 의미하며, 비면역원성 결과는 Assay.value_binary=FALSE로 라벨링되었다.

Assay.t_cell_type	Assay.value_binary	
	FALSE	TRUE
CD4	144	17
CD8	207	59
PBMC	522	62

[0364]

[0365] 양성 어세이는 ELISpot 어세이의 스팟 카운트를 기반으로 두 세트로 추가 분류되었다. 각각의 ELISpot 어세이에는 복제 실험이 있었고 평균 스팟 수는 복제된 것들에 걸쳐 특정되었다. 양성 신생항원의 경우, 모든 방문에 걸친 평균 스팟 카운트의 최댓값이 고려되었고, 양성 신생항원이 두 세트로 분할되었는데, 하나는 이 스팟 카운트 값이 <50이고, 다른 하나는 이 스팟 카운트 값이 >=50이다. 후자의 세트는 더 광범위한 T 세포 반응을 유도한 신생항원을 나타내며, 더 적은 스팟 카운트를 가진 세트에 비해 ELISpot 결과의 위양성 해석을 포함할 가능성이 적다. 50개의 스팟의 선택은 ELISpot 양성 판정에 사용된 원래 임계값(스팟 카운트 > 15)보다 상당히 높았기 때문에 임의의 결정이었다.

[0366] V.B.2.c. TESLA 멀티머 분석

[0367] 제2 테스트 면역원성 데이터세트의 경우, TESLA 컨소시엄은 신생항원 예측을 검증했다. TESLA의 대상체 식별자에서 대상체 1, 2, 3, 4, 10, 12 및 16에 대한 분석 데이터가 사용 가능했다. 어세이 결과는 네 가지 상이한 어세이: TCR_FLOW_I, TCR_FLOW_II, 나노입자 어세이 및 TCR 반응성 어세이에 기초하여 TESLA에 의해 제공되었다. 이 예시에서는 TCR_FLOW_I 어세이 결과가 사용됐다. 다른 어세이는 다음과 같은 이유로 무시되었다: (a) 나노입자 어세이는 매우 민감하도록 설계된 단일 세포 분석이므로 위양성률이 높을 것으로 예상됨, (b) TCR_FLOW_II는 TCR_FLOW_I와 대부분 중복되며 둘 다 다른 실험실에서 수행되고 TCR_FLOW_II는 데이터 포인트가 더 적음, TCR 반응성 어세이는 7일 동안 IL-2 및 짧은 펩타이드로 T 세포를 사전자극한 후 짧은 펩타이드로 재자극한 후의 세포내 IFN γ /TNF α 염색 어세이이다. TESLA 팀은 펩타이드-MHC 제시 예측을 평가하기 위해 이 어세이를 사용하는 것을 승인하지 않았다. 선택된 어세이는 16개의 양성 결과와 196개의 음성 결과를 보였다.

[0368] V.B.3. 비교 모델 - NetMHCpan 및 IEDB 점수

[0369] P-MHC-I 모델에 대한 성능 비교를 위해, 모델 A 및 모델 B가 사용되어 BA 및 EL 값을 펩타이드-HLA 쌍에 할당하였다. P-MHC-II 모델에 대한 성능 비교를 위해 모델 C가 사용되어 EL 값을 펩타이드-MHC(HLA) 쌍에 할당했다. 이들 방법에 의해 백분위 점수로 출력되는 BA 및 EL 값은 (이 예시에서) BA 또는 EL로 지칭된다. 이들 백분위수 값은 더 낮은 값일수록 더 높은 선호도 또는 제시 가능성을 의미하도록 동작한다. 이러한 값의 역수를 취해 더 높은 값이 더 강한 친화도 또는 제시 가능성을 나타내도록 거동하는 점수(가령, MHC-I의 경우 모델 A의 결합 친화도 점수, 모델 A에 대한 용리 점수, 모델 B에 대한 결합 친화도 점수, MHC-II의 경우, 모델 C에 대한 결합 친화도 점수)를 얻음으로써 변환된 점수 체계가 사용됐다. 신생항원결정기-HLA 쌍의 경우, 이러한 단일 점수가 획득된다. 신생항원의 경우, 모든 신생항원결정기-HLA 쌍이 돌연변이를 포함하는 8-14mer 길이의 신생항원 후보에 대해 고려되었고, 신생항원 점수를 나타내기 위해 가장 높은 점수를 가진 쌍이 선택됐다.

[0370] V.C. 결과

[0371] V.C.1. 제시 데이터에 대한 P-MHC-I 모델 성능

[0372] 도 14a-c는 하나 이상의 실시예에 따른 예시적인 정밀도-재현율(PR: precision-recall) 곡선을 갖는 플롯이다. 도 14a-c는 이전에 사용된 접근법과 비교하여 P-MHC-I 모델의 성능을 도시한다. 용리된 리간드(EL) 테스트 데이터세트가 사용되어 P-MHC-I 모델의 EL 출력, 모델 A의 EL 출력 및 모델 C의 결합 친화도(BA) 출력 간의 제시 예측 성능이 평가되었다.

[0373] 도 14a는 P-MHC-I 모델의 성능을 나타내는 플롯(1400)을 포함한다. 도 14b는 모델 A의 용리 출력에 대한 성능을 나타내는 플롯(1402)을 포함한다. 도 14c는 모델 B의 결합 친화도 출력에 대한 성능을 나타내는 플롯(1404)을 포함한다. 각각의 플롯(1400, 1402 및 1404)의 곡선 상의 점은 점수의 상위 1.71% 분위수에 대한 점수 임계값에

대응한다(금 표준 테스트 데이터의 1.71%가 양성이기 때문에 선택됨). 평균 정밀도(AP)는 임계값에 독립적인 성능을 나타낸다. F1 점수, 정밀도 및 재현율 값은 1.71% 임계값을 기반으로 한다.

[0374] 모델 A 및 모델 B 값은 이들 방법으로부터의 백분위수 순위 출력이었다. P-MHC-I 모델 값은 P-MHC-I 모델의 (최종 노드의) 출력에서 가져왔다. 이들 PR 곡선을 기반으로 도 14a-c는 P-MHC-I 모델이 모델 A 및 모델 C 둘 다에 비해 개선된 성능을 보임을 나타낸다(AP 값 0.85 대 모델 A의 경우 0.78 및 모델 B의 경우 0.57). 방법의 AP 값은 대립 유전자별로 비교되었다.

[0375] 도 15는 하나 이상의 실시예에 따른 테스트 데이터 세트에서 각각의 대립유전자에 대한 모델 A 및 P-MHC-I 모델의 용리-리간드 출력의 예시적인 평균 정밀도 값을 비교하는 플롯(1500)이다. 단일 대립유전자인 테스트 데이터 세트에는 적어도 1000개의 데이터 포인트가 포함되었으며 67개의 대립 유전자가 기준을 충족했다. 플롯(1500)에 나타난 바와 같이, 모델 A보다 P-MHC-I 모델이 더 높은 성능을 보였다. 플롯(1500)에서의 마커의 패턴은 대립유전자가 HLA-A, B 또는 C 유전자에서 유래했는지 여부를 나타낸다. 마커의 크기는 해당 대립유전자에 대해 P-MHC-I 모델을 교육하는 데 사용되는 단일 대립유전자 데이터의 양을 나타내며, 이는 각 대립유전자에 대한 테스트 데이터의 양과도 상관관계가 있다.

[0376] 도 16a 및 16b는 하나 이상의 실시예에 따르는 쥐 데이터세트에 대한 P-MHC-I 모델의 성능에 인간 데이터세트에 대한 P-MHC-I 모델의 성능을 비교하는 플롯(1600) 및 플롯(1602)이다. 이 플롯에서 볼 수 있듯이 P-MHC-I 모델은 P-MHC-I 모델의 평균 정밀도가 인간 및 쥐 데이터 세트 모두에 대해 유사하여 두 데이터 세트 모두에 대해 잘 수행되었다. 이러한 결과는 P-MHC-I 모델이 다양한 종에 걸쳐 바람직한 성능으로 사용될 수 있는 범종 모델일 수 있음을 보여준다.

[0377] V.C.2. 제시 데이터에 대한 P-MHC-II 모델 성능

[0378] 도 17a 및 17b는 하나 이상의 실시예에 따르는 제시 데이터에 대한 Model C에 P-MHC-II 모델의 성능을 각각 비교하는 플롯(1700) 및 플롯(1702)이다. 모델 C 값은 백분위 순위 출력이었다. P-MHC-II 모델 값은 P-MHC-II 모델의 출력(최종 노드)에서 가져왔다. PR 곡선의 평균 정밀도를 사용하면 도 17a 및 17b는 AP가 .69인 P-MHC-II 모델이 AP가 .31인 모델 C보다 향상된 성능을 나타냄을 나타낸다. 이 두 가지 방법의 AP 값은 대립유전자별로 비교되었다.

[0379] 도 18a 및 18b는 하나 이상의 실시예에 따르는 홀드아웃 데이터세트에 대한 Model C에 P-MHC-II 모델의 성능을 비교하는 각각의 플롯(1800) 및 플롯(1802)이다. 다시, AP가 .84인 P-MHC-II 모델은 AP가 .46인 모델 C보다 향상된 성능을 보여준다.

[0380] 도 19는 하나 이상의 실시예에 따르는 테스트 데이터세트에 대한 모델 C에 P-MHC-II 모델에 대한 평균 정밀도의 유전자형별 비교를 보여주는 플롯(1900)이다. 유전자형별로 P-MHC-II 모델은 모델 C보다 성능이 향상되었다.

[0381] V.C.3. 제1 및 제2 테스트 면역원성 데이터 세트에 대한 성능

[0382] 제1 및 제2 테스트 면역원성 데이터세트가 사용되어 T 세포 반응 데이터에 대한 P-MHC 제시 예측의 성능을 평가했다. 이들 평가에서 면역원성 데이터에 대한 훈련이 수행되지 않았으며 P-MHC 제시 점수를 계산하기 위해 신생항원 및 MHC 단백질의 아미노산 서열만 사용되었다. 다른 특징, 예를 들어 유전자 발현 또는 돌연변이 대립유전자가 환원주의적 방식으로 CD8 T 세포 반응을 예측하는 데 P-MHC 제시 예측의 기여도를 평가하는 데 사용되지 않았다.

[0383] V.C.3.a. 투약 대상체 멀티머 어세이

[0384] 도 20은 하나 이상의 실시예에 따르는 CD 8 멀티머 어세이 데이터(제1 테스트 면역원성 데이터세트)에 대한 P-MHC-I 모델(EL 출력), 모델 A(EL 출력) 및 모델 B(BA 출력)의 성능을 예시하는 수신기 작동 특성(ROC) 곡선의 플롯(2000)이다. 멀티머 어세이로부터 양성 신생항원결정기를 예측하는 능력과 관련하여 성능을 평가하였다. 모델 A 및 모델 B의 경우, 값이 역변환되어 각각 EL 및 BA 점수를 얻었고, 값이 높을수록 더 강한 결합 친화도 또는 제시 가능성을 나타냈다. AUC(area-under-curve)는 계단 함수를 기반으로 계산되었다. ROC 곡선을 그리는 단계 함수는 진양성률(tp)과 위양성률(fpr)을 나타내는 점을 가로 방향과 세로 방향으로 연결했다. 진양성률(tp) 및 위양성률(fpr) 값은 R 패키지 ROCR을 사용하여 계산되었다.

[0385] V.C.3.b. 투약 대상체 ELISpot 어세이

[0386] 도 21a-d는 하나 이상의 실시예에 따르는 ELISpot 어세이(제1 테스트 면역원성 데이터세트)에 대한 P-MHC-I 모

델(EL 출력), 모델 A(EL 출력) 및 모델 B(BA 출력)의 성능을 예시하는 플롯(2102, 2104, 2106, 및 2108)이다. 도시된 바와 같이 P-MHC-I 모델은 강력한 예측력으로 잘 수행되었다. 플롯은 PBMC ELISpots(도 21a, PBMC 패널) 및 CD8 ELISpot(도 21b, CD8 패널)에 대해 나타난 개별 서브플롯과 함께 예시적인 ROC 곡선을 보여준다. 양성 CD8 ELISpot 데이터는 더 강한 T 세포 반응(도 21c, CD8, 스팟 ≥ 50) 및 상대적으로 약한 T 세포 반응(도 D, CD8, 스팟 < 50)에 대해 두 개의 세트에 더 분할되었고, ROC 곡선이 만들어졌다. 이 두 세트에 대한 ROC 곡선을 만들기 위해 동일한 음성 신생항원 세트가 사용됐다.

[0387] V.C.3.c. TESLA 멀티머 분석

[0388] 도 22a-d는 하나 이상의 실시예에 따르는 모델 A(BA 출력), 모델 A(EL 출력), 모델 C(BA 출력) 및 P-MHC-I 모델(EL 출력)의 성능을 예시하는 플롯(2202, 2204, 2206, 및 2208)이다. TESLA 면역원성 데이터(제2 테스트 면역원성 데이터 세트)에서 성능이 평가됐으며, 멀티머 어세이로부터의 결과가 사용됐다. 이들 플롯은 TESLA 연구로부터 멀티머 어세이에 의해 평가된 예시적인 신생항원결정기-HLA 쌍에 대응하는 산점도이다. 반응은 TESLA에 의해 특정된 어세이에서 양성 적중의 경우 참(TRUE)이고 비면역원성 신생항원결정기의 경우 거짓(FALSE)이다. 윌콕슨 순위 합 테스트(Wilcoxon rank sum test)가 사용되어 양측 대립 가설에 대한 p-값을 계산했다. Y축은 더 높은 값일수록 더 강한 펩타이드-MHC 결합 또는 제시에 대응하도록 변환된 점수를 나타낸다.

[0389] 도 23은 하나 이상의 실시예에 따르는 TESLA 멀티머 어세이 데이터를 이용한 모델 A(EL 출력) 및 모델 B(BA 출력), 및 P-MHC-I 모델(EL 출력)에 대한 ROC 곡선을 비교하는 플롯(2300)의 예시이다. 멀티머 어세이는 TCR_FLOW_I 어세이였다. 곡선 아래 면적은 P-MHC-I 모델에서 가장 높았다.

[0390] V.D. 결론

[0391] 따라서, P-MHC 제시 예측 방법은 두 가지 유형의 평가 데이터 세트, 즉 면역펩티도믹스 실험으로부터의 P-MHC 제시 데이터 및 다양한 면역원성 어세이로부터의 T 세포 반응 데이터에 대해 평가되었다. 면역펩티도믹스 데이터에 대해 훈련된 제시 예측자는 이들 데이터 세트 다수에서 현재 생산 방법(IEDBV2.13BA 출력)에 비해 성능이 더 우수하다. P-MHC 모델은 데이터 세트 다수에서 향상된 성능 값을 보여주었다. 따라서 면역펩티도믹스 데이터에 대해 훈련된 어텐션-기반 기술을 사용하는 것이 생체외 결합 친화도 데이터에 기반한 모델보다 우수할 수 있다.

[0392] VI. 컴퓨터 구현 시스템

[0393] 도 26은 다양한 실시예에 따른 컴퓨터 시스템의 블록도이다. 컴퓨터 시스템(2600)은 도 1에서 전술한 컴퓨팅 플랫폼(102)에 대한 하나의 구현의 예일 수 있다.

[0394] 하나 이상의 예에서, 컴퓨터 시스템(2600)은 정보를 통신하기 위한 버스(2602) 또는 다른 통신 메커니즘, 및 정보를 처리하기 위해 버스(2602)와 결합된 프로세서(2604)를 포함할 수 있다. 다양한 실시예에서, 컴퓨터 시스템(2600)은 프로세서(2604)에 의해 실행될 명령을 결정하기 위해 버스(2602)에 결합된 RAM(random-access memory)(2606) 또는 다른 동적 저장 장치일 수 있는 메모리를 또한 포함할 수 있다. 메모리는 프로세서(2604)에 의해 실행될 명령의 실행 동안 임시 변수 또는 다른 중간 정보를 저장하기 위해서도 사용될 수 있다. 다양한 실시예에서, 컴퓨터 시스템(2600)은 정적 정보 및 프로세서(2604)를 위한 명령을 저장하기 위해 버스(2602)에 결합된 ROM(read only memory)(2608) 또는 다른 정적 저장 장치를 더 포함할 수 있다. 저장 장치(2610), 가령, 자기 디스크 또는 광 디스크가 제공될 수 있고 정보 및 명령을 저장하기 위해 버스(2602)에 연결될 수 있다.

[0395] 다양한 실시예에서, 컴퓨터 시스템(2600)은 컴퓨터 사용자에게 정보를 표시하기 위해 버스(2602)를 통해 음극선관(CRT) 또는 액정 디스플레이(LCD)와 같은 디스플레이(2612)에 연결될 수 있다. 영숫자 및 기타 키를 포함하는 입력 장치(2614)는 프로세서(2604)에 정보 및 명령 선택을 통신하기 위해 버스(2602)에 연결될 수 있다. 다른 유형의 사용자 입력 장치는 프로세서(2604)로 방향 정보 및 명령어 선택을 통신하고 디스플레이(2612) 상의 커서 이동을 제어하기 위한 커서 제어부(2616), 가령, 마우스, 조이스틱, 트랙볼, 제스처 입력 장치, 시선-기반 입력 장치, 또는 커서 방향 키이다. 이 입력 장치(2614)는 두 개의 축, 즉, 제1 축(가령, x) 및 제2 축(가령, y)에서 2 자유도를 가짐으로써 장치가 평면에서 위치를 지정할 수 있게 하는 것이 일반적이다. 그러나, 본 명세서에서 3차원(예를 들어, x, y 및 z) 커서 이동을 허용하는 입력 장치(2614)도 고려된다는 것을 이해해야 한다.

[0396] 본 교시의 특정한 구현에 따라, 결과는 프로세서(2604)가 RAM(2606)에 포함된 하나 이상의 명령의 하나 이상의 서열을 실행하는 것에 응답하여 컴퓨터 시스템(2600)에 의해 제공될 수 있다. 이러한 명령은 다른 컴퓨터 판독형 매체 또는 컴퓨터 판독형 저장 매체, 가령, 저장 장치(2610)로부터 RAM(2606)으로 판독될 수 있다.

RAM(2606)에 포함된 명령의 시퀀스의 실행은 프로세서(2604)로 하여금 본 명세서에 설명된 프로세스를 수행하게 할 수 있다. 대안으로, 본 교시를 구현하기 위해 소프트웨어 명령 대신에 또는 소프트웨어 명령과 조합하여 유선 회로가 사용될 수 있다. 따라서, 본 교시의 구현은 하드웨어 회로와 소프트웨어의 특정 조합으로 제한되지 않는다.

[0397] 본 명세서에서 사용될 때 용어 "컴퓨터 판독형 매체"(예를 들어, 데이터 저장소, 데이터 스토리지, 저장 장치, 데이터 저장 장치 등) 또는 "컴퓨터 판독형 저장 매체"는 프로세서(2604)에게 실행되도록 명령을 제공하는 데 참여하는 임의의 매체를 지칭한다. 이러한 매체는 여러 형태, 비제한적 예를 들면, 비휘발성 매체, 휘발성 매체 및 전송 매체를 취할 수 있다. 비휘발성 매체의 비제한적 예는 광학, 솔리드 스테이트, 자기 디스크, 가령, 저장 장치(2610)를 포함할 수 있다. 휘발성 매체의 예는 RAM(2606)과 같은 동적 메모리를 포함할 수 있지만 이에 제한되지 않는다. 전송 매체의 예는 버스(2602)를 구성하는 전선을 포함하여 동축 케이블, 구리 와이어 및 광섬유를 포함할 수 있지만 이에 제한되지 않는다.

[0398] 컴퓨터 판독형 매체의 일반적인 형태는 예를 들어 플로피 디스크, 플렉시블 디스크, 하드 디스크, 자기 테이프 또는 그 밖의 다른 임의의 자기 매체, CD-ROM, 그 밖의 다른 임의의 광학 매체, 펀치 카드, 종이 테이프, 구멍의 패턴을 갖는 그 밖의 다른 임의의 물리적 매체, RAM, PROM 및 EPROM, FLASH-EPROM, 그 밖의 다른 임의의 메모리 칩 또는 카트리지를, 또는 컴퓨터가 읽을 수 있는 그 밖의 다른 임의의 유형의 매체를 포함한다.

[0399] 컴퓨터 판독형 매체에 더하여, 명령 또는 데이터는 실행을 위해 컴퓨터 시스템(2600)의 프로세서(2604)에 하나 이상의 명령의 시퀀스를 제공하기 위해 통신 장치 또는 시스템에 포함된 전송 매체 상의 신호로서 제공될 수 있다. 예를 들어, 통신 장치는 명령 및 데이터를 나타내는 신호를 갖는 송수신기를 포함할 수 있다. 명령 및 데이터는 하나 이상의 프로세서로 하여금 본 명세서에 개시된 기능을 구현하게 하도록 구성된다. 데이터 통신 전송 연결의 대표적인 예는 전화 모뎀 연결, WAN(Wide Area Network), LAN(Local Area Network), 적외선 데이터 연결, NFC 연결, 광통신 연결 등을 포함할 수 있지만 이에 제한되지 않는다.

[0400] 본 명세서에 설명된 방법론, 흐름도, 도면 및 첨부된 개시 내용은 컴퓨터 시스템(2600)을 독립형 장치로 사용하거나 클라우드 컴퓨팅 네트워크와 같은 공유 컴퓨터 처리 리소스의 분산 네트워크에서 구현될 수 있음을 이해해야 한다.

[0401] 본 명세서에 설명된 방법론은 적용 분야에 따라 다양한 수단으로 구현될 수 있다. 예를 들어, 이러한 방법론은 하드웨어, 펌웨어, 소프트웨어 또는 이들의 조합으로 구현될 수 있다. 하드웨어 구현의 경우, 처리 유닛은 하나 이상의 ASIC(application specific integrated circuit), DSP(digital signal processor), DSPD(digital signal processing device), PLD(programmable logic device), FPGA(field programmable gate array), 프로세서, 제어기, 마이크로제어기, 마이크로프로세서, 전자 장치, 본 명세서에 설명된 기능을 수행하도록 설계된 그 밖의 다른 전자 장치 또는 이들의 조합 내에서 구현될 수 있다.

[0402] 다양한 실시예에서, 본 교시의 방법은 C, C++, Python 등과 같은 종래의 프로그래밍 언어로 작성된 펌웨어 및/또는 소프트웨어 프로그램 및 애플리케이션으로 구현될 수 있다. 펌웨어 및/또는 소프트웨어로 구현되는 경우, 본 명세서에 기술된 실시예는 컴퓨터로 하여금 전술한 방법을 수행하게 하기 위한 프로그램이 저장된 비일시적 컴퓨터 판독형 매체 상에서 구현될 수 있다. 본 명세서에 설명된 다양한 엔진은 컴퓨터 시스템, 가령, 컴퓨터 시스템(2600) 상에 제공될 수 있으며, 이에 따라 프로세서(2604)는 이들 엔진에 의해 제공되는 분석 및 결정을 실행할 수 있으며, 메모리 구성요소 RAM(2606), ROM(2608) 또는 저장 장치(2610) 중 임의의 하나 또는 조합에 의해 제공되는 명령 및 입력 장치(2614)를 통해 제공되는 사용자 입력에 종속된다.

[0403] VII. 용어의 예시적 설명

[0404] 본 명세서에서 사용될 때, 용어 "펩타이드", "폴리펩타이드" 및 "단백질"은 아미노산 잔기의 중합체를 지칭하기 위해 상호교환적으로 사용된다. 이 용어는 공유 펩타이드 결합에 의해 연결된 아미노산 잔기가 있는 전체 길이 단백질을 포함하여 모든 길이의 아미노산 사슬을 포함한다.

[0405] 본 명세서에서 사용될 때 "돌연변이 펩타이드"는 개별 대상체의 정상 조직(예를 들어, 정상 조직의 야생형 아미노산 서열)에 존재하지 않는 펩타이드를 지칭할 수 있다. 돌연변이 펩타이드는 적어도 하나의 돌연변이가 아미노산을 포함하고 질병 조직(예를 들어, 특정 대상체로부터 수집된 것)에 존재할 수 있지만 정상 조직(예를 들어, 상기 특정 대상체로부터 수집된 것, 상이한 대상체로부터 수집된 것, 및/또는 또는 데이터베이스에서 정상 조직에 대응하는 것으로 식별된 것)에는 존재하지 않을 수 있다. 돌연변이 펩타이드는 항원결정기(epitope)를 포함할 수 있다. 항원결정기는 MHC 분자 또는 T 세포 수용체(TCR)가 결합하는 돌연변이 펩타이드의 일부이다. 따라

서, 돌연변이 펩타이드의 항원결정기와 MHC 분자 또는 TCR 사이의 이러한 결합은 (돌연변이 펩타이드가 대상체 "자체"와 연관되지 않은 결과로서) 면역 반응을 유도할 수 있다. 돌연변이 펩타이드는 신생항원을 포함하거나 신생항원일 수 있다. 돌연변이 펩타이드는 예를 들어, 단백질 내 상이한 아미노산으로 이어지는 비동의 돌연변이(가령, 점 돌연변이), 종결 코돈이 변형되거나 결실되어 C-말단에 새로운 종양 특이적 서열을 갖는 더 긴 단백질의 번역을 야기하는 관독-통과 돌연변이(read-through mutation), 독특한 종양 특이적 단백질 서열을 야기하는 스플라이스 부위 돌연변이, 두 단백질의 접합부에서 종양 특이적 서열을 갖는 키메라 단백질을 생성하는 염색체 재배열(즉, 유전자 융합) 및/또는 종양 특이적 단백질 서열과 함께 새로운 개방 관독 프레임을 생성하는 프레임이동 삽입 또는 결실로부터 발생할 수 있다. 돌연변이 펩타이드는 폴리펩타이드(폴리펩타이드 서열에 의해 특징지어짐)를 포함 및/또는 뉴클레오타이드 서열에 의해 암호화될 수 있다.

[0406] 본 명세서에서 사용될 때, 펩타이드의 "C-플랭크"는 모체 단백질로부터 펩타이드의 C-말단의 업스트림에 있는 하나 이상의 아미노산을 지칭한다. 선택적으로, 펩타이드의 C-플랭크는 펩타이드의 C-말단의 업스트림에 1개, 2개, 3개, 4개, 5개 또는 그 이상의 아미노산 잔기를 포함한다.

[0407] 본 명세서에서 사용될 때, 펩타이드의 "N-플랭크"는 모체 단백질로부터 펩타이드의 N-말단 다운스트림의 하나 이상의 아미노산을 지칭한다. 선택적으로, 펩타이드의 N-플랭크는 펩타이드의 N-말단 다운스트림에 1개, 2개, 3개, 4개, 5개 또는 그 이상의 아미노산 잔기를 포함한다.

[0408] 본 명세서에서 사용될 때, 펩타이드의 "항원결정기"는 C-플랭크와 N-플랭크 사이의 펩타이드 영역을 의미할 수 있고 TCR에 의해 인식될 수 있다. 펩타이드의 항원결정기는 T 세포 상의 TCR과 항원 제시 세포의 MHC I에 의해 인식되는 펩타이드의 일부이다. 예를 들어, 항원결정기는 TCR이 결합하는 펩타이드, 예를 들어 펩타이드가 항원 제시 세포 상의 MHC I에 결합될 때 TCR이 결합하는 펩타이드일 수 있다.

[0409] 본 명세서에서 사용될 때, "리간드"는 용리 실험으로부터 세포 표면에서 MHC 분자에 의해 제시되거나 생체의 어세이에서 MHC에 결합되는 것으로 밝혀진 펩타이드다.

[0410] 본 명세서에서 사용될 때, "서열"은 아미노산 식별자의 정렬된 세트를 포함하는 아미노산 서열을 지칭한다.

[0411] 본 명세서에서 사용될 때 "펩타이드 서열"은 펩타이드의 적어도 일부분의 아미노산을 식별하는 서열을 지칭한다. 일부 경우에, 펩타이드 서열은 대응하는 참조 서열에서 관찰되지 않는 변이를 포함하는 변이-코딩 서열을 포함한다.

[0412] 펩타이드가 돌연변이 펩타이드를 포함하는 경우, 변이-코딩 서열은 돌연변이 또는 변이체의 아미노산을 식별한다. 그러나 펩타이드가 돌연변이 또는 변이를 포함하지 않는 경우, 변이-코딩 서열은 돌연변이 또는 변이의 아미노산을 식별하지 않는다(이 경우 참조 서열과 동일함). 변이-코딩 서열은 질병 및/또는 종양 샘플(예를 들어, 종양 세포 포함)을 수집하고 샘플에서 질병 및/또는 종양 세포에 대응하는 하나 이상의 서열을 식별하기 위해 시퀀싱 분석을 수행함으로써 결정될 수 있다. 일부 경우, 시퀀싱 분석이 아미노산 서열을 출력한다. 일부 경우, 시퀀싱 분석은 핵산 서열을 출력하며, 이는 후속적으로 처리되어 코돈을 아미노산 식별자로 변환하여 아미노산 서열을 생성할 수 있다. 변이-코딩 서열은 신생항원의 서열을 포함할 수 있다. 변이-코딩 서열은 펩타이드의 하나 이상의 말단(예를 들어, C-말단 및/또는 N-말단)을 포함할 수 있지만 반드시 그럴 필요는 없다. 변이-코딩 서열은 펩타이드의 항원결정기를 포함할 수 있다. 변이-코딩 서열은 대응하는 참조 서열에 비해 하나 이상의 변이체(예를 들어, 하나 이상의 아미노산 구별)를 갖는 펩타이드 내의 아미노산을 식별할 수 있다. 일부 경우, 변이-코딩 서열은 아미노산의 정렬된 세트를 포함한다. 일부 경우, 변이-코딩 서열은 (예를 들어, 유전자 참조 서열을, 가령, 유전자, 시작 위치 및/또는 끝 위치에 의해 식별함으로써, 또는 유전자, 시작 위치 및/또는 길이에 의해) 참조 펩타이드를 식별하고 참조 펩타이드에 대해 하나 이상의 점 돌연변이를 식별한다.

[0413] 본 명세서에 사용될 때, "참조 서열"은 비-돌연변이 펩타이드 또는 야생형 펩타이드(예를 들어, 야생형, 모 서열)의 적어도 일부 내의 아미노산을 확인하는 서열을 지칭할 수 있다. 비-돌연변이 또는 야생형 펩타이드는 변이체를 포함하지 않거나 돌연변이 펩타이드에 포함된 것보다 적은 수의 변이체를 포함할 수 있다. 참조 서열은 대응하는 변이-코딩 서열을 포함하는 유전자에 대해 동일한 유전자 내의 유전자 서열에 의해 암호화되는 아미노산 서열을 포함할 수 있다. 참조 서열은 대응하는 변이-코딩 서열과 연관된 유전자 서열과 연관된 유전자내 위치에 대해 유전자 내에서 동일한 시작부분과 정지부분에 뻗어 있는 유전자 서열에 의해 암호화된 아미노산 서열을 포함할 수 있다. 참조 서열은 하나 이상의 대상체(변이-코딩 서열을 결정하기 위해 질병 샘플이 수집된 대상체를 포함할 수 있지만 반드시 그럴 필요는 없음)로부터 비-질병 및/또는 비-종양 샘플을 수집하고, 샘플을 이용해 시퀀싱 분석을 수행함으로써 식별될 수 있다.

- [0414] 본 명세서에서 사용될 때, MHC 분자의 "의사서열"은 펩타이드와 접촉하는 MHC 분자의 아미노산의 정렬된 세트를 지칭할 수 있다.
- [0415] 본 명세서에서 사용될 때, 서열의 "표현"은 서열 내의 아미노산을 나타내거나 식별하는 값의 세트 및/또는 서열을 암호화하는 핵산을 나타내거나 식별하는 값의 세트를 포함할 수 있다. 예를 들어, 각각의 아미노산은 이진 문자열 및/또는 서로 구별되는 값의 벡터 및/또는 서로 아미노산을 나타내는 벡터로 나타낼 수 있다. 표현은 예를 들어 원-핫 암호화 또는 BLOSUM(BLOSUM) 행렬을 사용하여 생성될 수 있다. 예를 들어, 다차원(가령, 20차원 또는 21차원) 배열이 초기화(가령, 랜덤 또는 의사랜덤으로 초기화)된다. 초기화된 배열은 각각의 아미노산에 대해 해당 아미노산에 대응하는 고유한 벡터를 포함할 수 있다. 이러한 고유한 벡터의 사용이 대응하는 아미노산을 나타내는 것으로 추정될 수 있도록 값이 고정될 수 있다. 다중 코돈 중 임의의 것이 단일 아미노산을 암호화할 수 있는 경우, 주어진 서열의 다수의 가능한 핵산 표현이 있을 수 있다.
- [0416] 본 명세서에서 사용될 때, 펩타이드의 "제시"는 펩타이드의 적어도 일부가 특정 방식으로 MHC 분자에 결합됨으로써 세포 표면에 제시되는 것을 지칭한다. 제시된 펩타이드는 다른 세포, 가령, 근처의 T 세포에 접근할 수 있다.
- [0417] 본 명세서에서 사용될 때, "샘플"은 조직(예를 들어, 생검), 단일 세포, 다중 세포, 세포 단편 또는 체액의 분취량을 포함할 수 있다. 샘플은 비제한적 예를 들어 정맥 천자, 배설, 사정, 마사지, 생검, 바늘 흡인, 세척 샘플, 긁기, 수술 절개, 개입, 또 다른 유형의 샘플 수집 수단, 또는 이들의 조합에 의해 대상체로부터 획득될 수 있다.
- [0418] 본 명세서에서 사용될 때 "대상체"는 하나 이상의 세포, 조직 또는 유기체를 포함한다. 대상체는 생체내(in vivo), 탈체(ex vivo), 또는 생체외(in vitro), 남성 또는 여성 여부에 관계없이 인간 또는 비인간일 수 있다. 대상체는 포유동물, 가령, 인간일 수 있다.
- [0419] 본 명세서에 사용될 때, "결합 친화도"는 펩타이드(예를 들어, 특정 항원의 것)와 MHC(예를 들어, MHC 분자 및/또는 MHC 대립유전자) 사이의 결합의 친화도를 지칭한다. 결합 친화도는 펩타이드와 MHC 분자 사이의 결합의 안정성, 경향 및/또는 강도를 특징지을 수 있다.
- [0420] 본 명세서에 사용될 때, "면역원성"은 (예를 들어, T 세포 및/또는 B 세포를 통해) 면역 반응을 유도하는 능력을 지칭할 수 있다. "면역원성"인 펩타이드는 면역 반응을 유도할 수 있는 펩타이드일 수 있다.
- [0421] 본 명세서에 사용될 때 "MHC"는 주 조직 적합성 복합체(major histocompatibility complex)를 지칭한다. 인간 MHC는 인간 백혈구 항원(HLA) 복합체라고도 한다.
- [0422] **VIII. 실시예**
- [0423] 실시예 1. 방법이 제공된다. 상기 방법은 펩타이드의 세트를 특징짓는 펩타이드 서열의 세트를 액세스하는 단계를 포함하고, 펩타이드 서열의 세트의 각각의 펩타이드 서열은 대상체로부터의 질병 샘플을 처리함으로써 식별되었다. 상기 방법은 대상체의 면역단백질 복합체(IPC: immunoprotein complex)에 대해 식별된 면역단백질 복합체(IPC) 서열을 액세스하는 단계를 포함한다. 상기 방법은 어텐션-기반 머신 러닝 모델의 초기 어텐션 서브시스템 내 제1 어텐션 블록을 이용해 펩타이드 서열의 세트를 나타내는 펩타이드 표현의 세트 및 초기 어텐션 서브시스템 내 제2 어텐션 블록을 이용해 IPC 서열을 나타내는 면역단백질 복합체(IPC) 표현을 처리하여 출력을 생성하는 단계를 포함하며, 상기 출력은 대응하는 펩타이드-IPC 결합체에 대한 상호작용 예측, 상호작용 친화도 예측, 또는 면역원성 예측 중 적어도 하나를 포함한다. 상기 방법은 상기 출력에 기초하여 리포트를 생성하는 단계를 포함한다.
- [0424] 실시예 2. 실시예 1의 방법에서, 펩타이드 서열의 세트의 적어도 하나의 펩타이드 서열은 대응하는 참조 서열에 대한 변형을 포함하는 변이-코딩 서열을 포함한다.
- [0425] 실시예 3. 실시예 1 또는 실시예 2의 방법에서, 처리하는 것이 펩타이드 서열의 세트의 대응하는 펩타이드 서열에 대한 펩타이드 표현의 세트의 펩타이드 표현을 수신하는 것, 및 제1 어텐션 블록을 통해 펩타이드 표현을 변환 펩타이드 표현으로 변환하는 것 — 상기 제1 어텐션 블록은 어텐션 서브-블록의 세트의 각각의 어텐션 서브-블록이 셀프-어텐션 층을 포함하는 어텐션 서브-블록의 세트를 포함함 — 을 포함한다.
- [0426] 실시예 4. 실시예 1-3 중 어느 한 방법에서, 처리하는 것이 IPC 표현을 수신하는 것, 및 제2 어텐션 블록을 통해 IPC 표현을 변환 IPC 표현으로 변환하는 것 — 상기 제2 어텐션 블록은 어텐션 서브-블록의 세트의 각각의

어텐션 서브-블록이 셀프-어텐션 층을 포함하는 어텐션 서브-블록의 세트를 포함함 — 을 포함한다.

- [0427] 실시예 5. 실시예 1-4 중 어느 한 방법에서, 펩타이드 표현의 적어도 일부분이 펩타이드 서열 내 단량체에 대응하고 IPC 표현의 적어도 일부분이 IPC 서열 내 단량체에 대응하며, 처리하는 것은: 제1 어텐션 블록 및 제1 가중치 세트를 이용해 펩타이드 표현에 기초하여 변환 펩타이드 표현을 생성하는 것, 제2 어텐션 블록 및 제2 가중치 세트를 이용해 IPC 표현에 기초하여 변환 IPC 표현을 생성하는 것, 및 변환 펩타이드 표현 및 변환 MHC 표현을 이용해 복합 표현을 생성하는 것을 포함한다.
- [0428] 실시예 6. 실시예 1-5 중 어느 한 방법은, 상기 펩타이드 서열의 세트의 펩타이드 서열을 임베딩하여 상기 펩타이드 서열에 대한 임베딩된 펩타이드 표현을 생성하는 단계, 및 상기 펩타이드 서열에 대한 임베딩된 펩타이드 서열을 위치적으로 암호화하여 펩타이드 서열을 나타내는 펩타이드 표현의 세트의 펩타이드 표현을 생성하는 단계를 더 포함한다.
- [0429] 실시예 7. 실시예 1-6 중 어느 한 방법에서, 제1 어텐션 블록은 어텐션 서브-블록의 세트를 포함하고, 어텐션 서브-블록의 세트의 각각의 어텐션 서브-블록은 적어도 하나의 셀프-어텐션 층을 포함하는 신경망을 포함한다.
- [0430] 실시예 8. 실시예 1-7 중 어느 한 방법에서, 제2 어텐션 블록은 어텐션 서브-블록의 세트를 포함하고, 어텐션 서브-블록의 세트의 각각의 어텐션 서브-블록은 적어도 하나의 셀프-어텐션 층을 포함하는 신경망을 포함한다.
- [0431] 실시예 9. 실시예 1-8 중 어느 한 방법에서, 제1 어텐션 블록은 제1 복수의 어텐션 서브-블록을 포함하고, 제2 어텐션 블록은 제1 복수의 어텐션 서브-블록을 포함하며, 어텐션 서브-블록의 제1 세트 및 어텐션 서브-블록의 제2 세트의 각각의 어텐션 서브-블록은 적어도 하나의 셀프-어텐션 층을 포함하는 신경망을 포함한다.
- [0432] 실시예 10. 실시예 1-9 중 어느 한 방법에서, 상기 펩타이드 표현의 세트의 펩타이드 표현이 제1 어텐션 블록을 이용해 처리되는 통합 표현의 제1 부분을 형성하고, 통합 표현의 제2 부분이 N-플랭크 서열 또는 C-플랭크 서열 중 적어도 하나를 나타낸다.
- [0433] 실시예 11. 실시예 1-10 중 어느 한 방법에서, 상기 펩타이드 서열의 세트의 펩타이드 서열은 통합 서열의 제1 부분을 형성하고, 통합 서열의 제2 부분이 N-플랭크 서열 또는 C-플랭크 서열 중 적어도 하나를 포함하며, 어텐션-기반 머신 러닝 모델은 통합 서열을 수신 및 처리하여 상기 펩타이드 표현의 세트 중 펩타이드 서열에 대응하는 펩타이드 표현을 포함하는 통합 표현을 형성하는 표현 블록을 포함하고, 상기 통합 표현은 제1 어텐션 블록에 의해 처리된다.
- [0434] 실시예 12. 실시예 1-11 중 어느 한 방법은, IPC 서열을 임베딩하여 IPC 서열의 임베딩된 IPC 표현을 생성하는 단계, 및 IPC 서열의 임베딩된 IPC 표현을 위치적으로 암호화하여 IPC 표현을 생성하는 단계를 더 포함한다.
- [0435] 실시예 13. 실시예 1-12 중 어느 한 방법에서, 어텐션-기반 머신 러닝 모델은 복수의 셀프-어텐션 층을 포함하고 복수의 셀프-어텐션 층의 각각에 대해 대응하는 다운스트림 피드포워드 신경망을 포함한다.
- [0436] 실시예 14. 실시예 1-13 중 어느 한 방법에서, 제1 어텐션 블록은 펩타이드 표현의 세트의 펩타이드 표현을 수신 및 처리하여 변환 펩타이드 표현을 생성하도록 구성된 제1 신경망을 포함하고, 제2 어텐션 블록은 IPC 표현을 수신 및 처리하여 변환 IPC 표현을 생성하도록 구성된 제2 신경망을 포함하며, 제1 신경망 및 제2 신경망 각각은 적어도 하나의 셀프-어텐션 층을 포함하고, 어텐션-기반 머신-러닝 모델은 변환 펩타이드 표현 및 변환 IPC 표현을 이용해 복합 표현을 생성하도록 구성된다.
- [0437] 실시예 15. 실시예 1-14 중 어느 한 방법에서, 어텐션-기반 머신 러닝 모델은: 복합 표현을 수신 및 처리하도록 구성된 신경망을 포함하는 복합 어텐션 블록 — 신경망은 셀프-어텐션 층을 포함함 — 을 더 포함한다.
- [0438] 실시예 16. 실시예 1-15 중 어느 한 방법에서, 어텐션-기반 머신 러닝 모델은: 어텐션 서브-블록의 세트를 포함하는 복합 어텐션 블록 — 어텐션 서브-블록의 세트의 각각의 어텐션 서브-블록은 적어도 하나의 셀프-어텐션 층을 포함하는 신경망을 포함함 — 을 더 포함한다.
- [0439] 실시예 17. 실시예 1-16 중 어느 한 방법에서, IPC는 주 조직적합성 복합체(MHC: major histocompatibility complex)를 포함하고 대응하는 펩타이드-IPC 결합체는 상기 펩타이드의 세트의 펩타이드 및 MHC를 포함하며, 대응하는 펩타이드-IPC 결합체에 대한 상호작용 친화도 예측은 펩타이드와 MHC 간 결합 친화도를 예측하고, 대응하는 펩타이드-IPC 결합체에 대한 상호작용 예측은 MHC가 세포 표면에서 펩타이드를 제시할 것인지 여부를 예측한다.
- [0440] 실시예 18. 실시예 1-17 중 어느 한 방법에서, 어텐션-기반 머신 러닝 모델은 복수의 훈련 펩타이드 서열 및 훈

런 MHC 서열의 세트에 대한 실험적 상호작용 친화도 데이터 또는 실험적 상호작용 데이터 중 적어도 하나를 포함하는 훈련 데이터 세트를 이용해 훈련된다.

- [0441] 실시예 19. 실시예 1-18 중 어느 한 방법에서, IPC는 T 세포 수용체(TCR)이고 대응하는 펩타이드-IPC 쌍은 상기 펩타이드의 세트의 펩타이드와 TCR 또는 TCR과 주 조직적합성 복합체(MHC)를 포함하며, 대응하는 펩타이드-IPC 결합체에 대한 면역원성 예측은 TCR에 대한 펩타이드의 면역원성을 예측하고, 어텐션-기반 머신 러닝 모델은 복수의 훈련 펩타이드 서열 및 훈련 TCR 서열의 세트에 대한 실험적 면역원성 데이터를 포함하는 훈련 데이터를 이용해 훈련된다.
- [0442] 실시예 20. 실시예 1-19 중 어느 한 방법에서, 훈련 데이터 세트는 복수의 훈련 데이터 요소를 포함하며, 복수의 훈련 데이터 요소의 적어도 하나의 훈련 데이터 요소는: 상기 펩타이드의 세트에 포함되지 않는 훈련 펩타이드를 특징짓는 훈련 펩타이드 서열, IPC와 상이한 훈련 IPC를 특징짓는 훈련 IPC 서열, 및 훈련 펩타이드와 훈련 IPC 간 상호작용 친화도 표시자를 식별하는 실험-기반 결과 - 상기 상호작용 친화도 표시자는 검정 또는 바이오센서-기반 방법을 이용해 검출되었음 - 중 적어도 하나를 포함한다.
- [0443] 실시예 21. 실시예 1-20 중 어느 한 방법에서, 훈련 데이터 세트는 복수의 훈련 데이터 요소를 포함하며, 복수의 훈련 데이터 요소의 적어도 하나의 훈련 데이터 요소는: 상기 펩타이드의 세트에 포함되지 않는 훈련 펩타이드를 특징짓는 훈련 펩타이드 서열, IPC와 상이한 훈련 MHC를 특징짓는 훈련 MHC 서열, 및 훈련 펩타이드가 훈련 MHC에 의해 세포 표면에 제시되었는지 여부를 식별하는 상호작용 표시자를 포함하는 실험-기반 결과 - 상기 상호작용 표시자를 결정하기 위해 면역침강법 또는 질량 분석법 중 적어도 하나가 사용되었음 - 중 적어도 하나를 포함한다.
- [0444] 실시예 22. 실시예 1-21 중 어느 한 방법에서, 처리 단계 전에, 복수의 펩타이드-IPC 결합체에 대한 결합 친화도, 상호작용 표시자, 또는 면역원성 표시자 중 적어도 하나를 포함하는 훈련 데이터 세트를 이용해 어텐션-기반 머신 러닝 모델을 훈련하는 단계를 더 포함하며, 상기 훈련 데이터 세트는 복수의 훈련 펩타이드 서열 및 복수의 훈련 주 조직적합성 복합체(MHC) 서열 또는 복수의 훈련 T 세포 수용체(TCR) 서열 중 적어도 하나를 포함한다.
- [0445] 실시예 23. 실시예 1-22 중 어느 한 방법에서, 처리하는 것은: 제1 어텐션 블록을 이용해 펩타이드 표현의 세트를 처리하고 제2 어텐션 블록을 이용해 IPC 표현을 처리하여 펩타이드-IPC 결합체의 세트에 대한 복합 표현의 세트를 생성하는 것, 상기 복합 표현의 세트를 처리하여 결과의 세트를 생성하는 것, 펩타이드-IPC 결합체의 세트 중 서브세트를 선택하는 것 - 선택된 상호작용의 세트가 펩타이드-IPC 결합체의 세트의 나머지 서브세트에 비교해서 상기 서브세트의 각각의 펩타이드-IPC 결합체에서 발생할 가능성이 높음 - 을 포함하고, 상기 리포트는 상기 서브세트 내 각각의 펩타이드를 식별한다.
- [0446] 실시예 24. 실시예 1-23 중 어느 한 방법에서, 펩타이드의 세트의 각각의 펩타이드가 사용되어 펩타이드-IPC 결합체의 세트를 형성하며, 어텐션-기반 머신 러닝 모델은 펩타이드-IPC 결합체의 세트의 각각의 펩타이드-IPC 결합체에 대한 면역원성 예측을 생성하도록 구성되며, 펩타이드-IPC 결합체의 세트의 펩타이드-IPC 결합체에 대한 면역원성 예측은 펩타이드-IPC 결합체 내 펩타이드의 중앙-특이적 면역원성의 예측이다.
- [0447] 실시예 25. 실시예 1-24 중 어느 한 방법에서, 상기 리포트는 상기 펩타이드의 세트로부터 펩타이드의 세트의 나머지 부분에 비해 증가된 중앙-특이적 면역원성을 갖는 펩타이드의 서브세트를 식별함을 포함한다.
- [0448] 실시예 26. 실시예 1-25 중 어느 한 방법에서, IPC는 주 조직적합성 복합체(MHC)이며, 펩타이드의 세트의 각각의 펩타이드가 사용되어 펩타이드-MHC 결합체의 세트를 형성하고, 어텐션-기반 머신 러닝 모델은 펩타이드-MHC 결합체의 세트의 각각의 펩타이드-MHC 결합체에 대한 상호작용 예측을 생성하도록 구성되며, 펩타이드-MHC 결합체의 세트의 펩타이드-MHC 결합체에 대한 상호작용 예측은 펩타이드-MHC 결합체 내 펩타이드가 세포 표면에서 MHC에 의해 제시되는지 여부의 예측이다.
- [0449] 실시예 27. 실시예 26의 방법에서, 상기 리포트는 상기 펩타이드의 세트로부터, 상기 펩타이드의 세트의 나머지 부분에 비해 MHC에 의한 증가된 제시 가능성을 갖는 펩타이드의 서브세트를 식별한다.
- [0450] 실시예 28. 실시예 1-27 중 어느 한 방법에서, 상기 펩타이드 서열의 세트의 펩타이드 서열은 돌연변이 펩타이드를 특징짓는 변이-코딩 서열이고, 상기 변이-코딩 서열은: 돌연변이 펩타이드의 N-말단에서의 서열을 식별하는 제1 부분, 및 돌연변이 펩타이드의 항원결정기의 서열을 식별하는 제2 부분을 포함하고, 상기 처리하는 것은: 초기 어텐션 서브시스템의 제1 셀프-어텐션 층을 이용해 변이-코딩 서열의 제1 부분의 제1 표현을 처리하는 것, 및 초기 어텐션 서브시스템의 제2 셀프-어텐션 층을 이용해 변이-코딩 서열의 제2 부분의 제2 표현을 처

리하는 것을 포함한다.

- [0451] 실시예 29. 실시예 28의 방법에서, 제1 표현 및 제2 표현은 제1 어텐션 블록 내에서 처리된다.
- [0452] 실시예 30. 실시예 1-29 중 어느 한 방법에서, 어텐션-기반 머신 러닝 모델은 하나 이상의 변환기 인코더(transformer encoder)를 포함하고, 상기 하나 이상의 변환기 인코더의 각각은 셀프-어텐션 층을 포함한다.
- [0453] 실시예 31. 실시예 1-30 중 어느 한 방법에서, IPC 서열 및 펩타이드 서열의 세트의 각각은 아미노산 식별자의 정렬된 세트를 포함한다.
- [0454] 실시예 32. 실시예 1-31 중 어느 한 방법에서, IPC 서열은 상기 질병 샘플을 이용해 식별된다.
- [0455] 실시예 33. 실시예 1-32 중 어느 한 방법에서, IPC 서열은 상기 대상체로부터의 생물학적 샘플을 이용해 식별된다.
- [0456] 실시예 34. 실시예 1-33 중 어느 한 방법에서, 상기 질병 샘플은 암 세포를 포함한다.
- [0457] 실시예 35. 실시예 1-34 중 어느 한 방법에서, 상기 대상체의 IPC는 주 조직적합성 복합체(MHC)를 포함하고, IPC 서열은 MHC 서열을 포함하며, IPC 표현은 MHC 표현을 포함한다.
- [0458] 실시예 36. 실시예 35의 방법에서, MHC는 MHC 클래스-I 분자를 포함한다.
- [0459] 실시예 37. 실시예 35의 방법에서, MHC는 MHC 클래스-II 분자를 포함한다.
- [0460] 실시예 38. 실시예 1-35 중 어느 한 방법에서, 대상체의 IPC는 T 세포 수용체(TCR)를 포함하고, IPC 서열은 TCR 서열을 포함하며, IPC 표현은 TCR 표현을 포함한다.
- [0461] 실시예 39. 실시예 1-38 중 어느 한 방법에서, 상기 질병 샘플은 조직을 포함한다.
- [0462] 실시예 40. 실시예 1-39 중 어느 한 방법에서, 상기 펩타이드의 세트의 적어도 하나의 펩타이드는 신생항원이다.
- [0463] 실시예 41. 실시예 1-40 중 어느 한 방법에서, 상기 펩타이드 서열의 세트의 적어도 하나의 펩타이드 서열은 상기 질병 샘플로부터 도출된 유전자 서열이다.
- [0464] 실시예 42. 실시예 1-41 중 어느 한 방법에서, 변이-코딩 서열의 세트의 각각의 적어도 하나는 상기 질병 샘플의 RNA 서열에 기초한다.
- [0465] 실시예 43. 실시예 1-42 중 어느 한 방법에서, 대응하는 펩타이드-IPC 결합체는 펩타이드의 세트로부터의 펩타이드 및 IPC를 포함하고, IPC는 주 조직적합성 복합체(MHC)이며, 상호작용 친화도 예측은 펩타이드와 MHC 간 결합에 대한 결합 친화도의 예측이고, 상호작용 예측은 MHC에 의한 세포 표면에서의 펩타이드의 제시의 예측이다.
- [0466] 실시예 44. 실시예 1-43 중 어느 한 방법은, 사용자에게 의해 입력된 입력 데이터를 수신하는 단계 - 상기 입력 데이터는 상기 대상체에 대응함 - 를 더 포함하고, 상기 입력 데이터의 수신에 응답하여, 데이터 저장소로부터의 검색을 통해 펩타이드 서열의 세트 및 IPC 서열이 액세스되며, 상기 리포트는 상기 펩타이드의 세트로부터 상기 대상체의 병태를 치료하기 위한 개별화된 백신에 포함될 펩타이드의 서브세트를 식별한다.
- [0467] 실시예 45. 실시예 44의 방법은, 대상체에 대한 개별화된 백신을 포함하는 권장 치료제를 생성하는 단계를 더 포함한다.
- [0468] 실시예 46. 실시예 1-45 중 어느 한 방법은, 사용자에게 의해 입력된 입력 데이터를 수신하는 단계 - 상기 입력 데이터는 상기 대상체에 대응하고, 상기 입력 데이터의 수신에 응답하여, 데이터 저장소로부터의 검색을 통해 펩타이드 서열의 세트 및 IPC 서열이 액세스됨 - , 및 상기 리포트에 기초하여 개별화된 백신에 포함되도록 치료 펩타이드의 세트를 결정하는 단계, 및 치료 펩타이드의 세트를 포함하는 개별화된 백신의 제조를 용이하게 하는 동작을 개시하는 단계를 더 포함한다.
- [0469] 실시예 47. 실시예 46에서, 상기 동작을 개시하는 단계는: 개별화된 백신의 제조에 포함되는 컴퓨터화된 프로세스를 트리거하는 알람을 생성하는 단계를 포함한다.
- [0470] 실시예 48. 실시예 1-47 중 어느 한 방법에서, 처리하는 것은: 어텐션-기반 머신 러닝 모델 내 임베딩 블록으로부터, 복수의 요소를 포함하는 표현을 수신하는 것 - 표현은 펩타이드 서열의 세트 내 펩타이드 서열을 나타내는 펩타이드 표현의 세트의 펩타이드 표현 또는 IPC 서열을 나타내는 IPC 표현이고, 다중-요소 데이터 세트의

각각의 요소가 펩타이드 서열 또는 IPC 서열 내 단량체에 대응함 — , 복수의 요소의 각각의 요소에 대해, 어텐션-기반 머신 러닝 모델의 셀프-어텐션 층과 연관된 키 가중치의 세트, 값 가중치의 세트, 및 쿼리 가중치의 세트에 각각 기반하여 키 벡터, 값 벡터, 및 쿼리 벡터를 결정하는 것, 복수의 요소의 변환(transformation)을 수행하여 복수의 수정 요소를 형성하는 것 — 상기 변환은 복수의 요소에 대해 생성된 어텐션 점수 및 복수의 요소의 각각에 대해 결정된 값 벡터를 이용해 수행됨 — , 및 복수의 수정 요소에 기초하여 출력을 생성하는 것을 포함한다.

- [0471] 실시예 49. 실시예 48의 방법에서, 상기 복수의 요소의 선택된 요소에 대해 변환을 수행하는 것은: 요소의 키 벡터 및 쿼리 벡터를 이용해 상기 선택된 요소의 어텐션 점수를 결정하는 것 — 상기 복수의 요소 중 상기 선택된 요소가 아닌 나머지 부분이 나머지 요소의 세트를 형성함 — , 상기 나머지 요소의 키 벡터 및 상기 선택된 요소의 쿼리 벡터를 이용해 상기 나머지 요소의 세트의 각각의 나머지 요소에 대해 추가 어텐션 점수를 결정하여 추가 어텐션 점수의 세트를 형성하는 것, 및 어텐션 점수, 추가 어텐션 점수의 세트, 복수의 요소의 각각의 요소의 값 벡터를 이용해 수정 요소를 생성하는 것을 포함한다.
- [0472] 실시예 50. 실시예 1-49 중 어느 한 방법은, 디스플레이 시스템 상의 그래픽 사용자 인터페이스 상에 상기 리포트를 디스플레이하는 단계를 더 포함한다.
- [0473] 실시예 51. 실시예 1-50 중 어느 한 방법에서, 처리하는 것은 제1 컴퓨팅 플랫폼 상에서 수행되며: 유선 통신 링크 또는 무선 통신 링크 중 적어도 하나를 포함하는 통신 링크의 세트를 통해 상기 리포트를 제2 컴퓨팅 플랫폼으로 전송하는 것을 더 포함한다.
- [0474] 실시예 52. 실시예 1-51 중 어느 한 방법은, 상기 펩타이드의 세트의 적어도 하나의 펩타이드를 상기 리포트에 기초하여 면역요법을 위한 표적으로서 포함하도록 결정하는 단계를 더 포함한다.
- [0475] 실시예 53. 실시예 52의 방법에서, 상기 면역요법은 T 세포 요법, 개별화된 암 요법, 항원-특이적 면역요법, 항원-의존적 면역요법, 백신 및 자연 살해(NK) 세포 요법으로 구성된 군 중에서 선택된다.
- [0476] 실시예 54. 실시예 1-53 중 어느 한 방법은, 상기 펩타이드의 세트의 적어도 하나의 펩타이드를 상기 리포트에 기초하여 면역요법을 위한 표적으로서 배제하도록 결정하는 단계를 더 포함한다.
- [0477] 실시예 55. 실시예 54의 방법에서, 상기 면역요법은 T 세포 요법, 개별화된 암 요법, 항원-특이적 면역요법, 항원-의존적 면역요법, 백신 및 자연 살해(NK) 세포 요법으로 구성된 군 중에서 선택된다.
- [0478] 실시예 56. 실시예 1-55 중 어느 한 방법에서, IPC는 인간 백혈구 항원(HLA: human leukocyte antigen) 분자이다.
- [0479] 실시예 57. 실시예 1-56 중 어느 한 방법은, 상기 대상체로부터의 질병 샘플을 시퀀싱하는 단계, 상기 대상체로부터의 질병 샘플의 시퀀싱에 기초하여 펩타이드 서열의 세트를 정의하는 단계, 상기 리포트에 기초하여, 펩타이드 서열의 세트의 서브세트를 식별하는 단계, 상기 펩타이드의 세트의 상기 서브세트에 포함되는 적어도 하나의 펩타이드에 대해 코딩하는 mRNA를 합성하는 단계, mRNA-리포플렉스 치료제를 생성하기 위해 mRNA를 지질과 복합화하는 단계, 및 mRNA-리포플렉스 치료제를 상기 대상체에게 투여하는 단계를 더 포함한다.
- [0480] 실시예 58. 백신이 하나 이상의 펩타이드, 하나 이상의 펩타이드를 암호화하는 복수의 핵산, 또는 하나 이상의 펩타이드를 발현하는 복수의 세포를 포함하며, 상기 하나 이상의 펩타이드는 실시예 1-49 중 어느 한 방법에 의해 생성된 리포트에 기초하여 펩타이드의 세트 중에서 선택되며, 상기 하나 이상의 펩타이드는 상기 펩타이드의 세트의 불완전한 서브세트이다.
- [0481] 실시예 59. 실시예 58의 백신에서, 상기 백신은 복수의 핵산을 포함하는 DNA 또는 복수의 핵산을 포함하는 RNA를 포함한다.
- [0482] 실시예 60. 실시예 58 또는 실시예 59의 백신에서, 상기 백신은 복수의 핵산을 포함하는 mRNA를 포함한다.
- [0483] 실시예 61. 실시예 58-60 중 어느 한 백신에서, 상기 백신은 종양 백신이다.
- [0484] 실시예 62. 백신을 제조하는 방법이, 백신을 생성하는 단계 — 상기 백신은: 하나 이상의 펩타이드, 하나 이상의 펩타이드를 암호화하는 복수의 핵산, 또는 하나 이상의 펩타이드를 발현하는 복수의 세포를 포함함 — 를 포함하며, 상기 하나 이상의 펩타이드는 실시예 1-49 중 어느 한 방법에 의해 생성된 리포트에 기초하여 펩타이드의 세트 중에서 선택되며, 상기 하나 이상의 펩타이드는 상기 펩타이드의 세트의 불완전한 서브세트이다.

- [0485] 실시예 63. 실시예 62의 방법에서, 상기 백신은 복수의 핵산을 포함하는 DNA, 복수의 핵산을 포함하는 RNA, 또는 복수의 핵산을 포함하는 mRNA를 포함한다.
- [0486] 실시예 64. 실시예 62 또는 실시예 63의 방법은, 하나 이상의 펩타이드 내 아미노산에 기초하여, 하나 이상의 펩타이드를 암호화하는 복수의 핵산을 식별하는 단계 - 상기 백신은 복수의 핵산을 포함함 - 를 더 포함한다.
- [0487] 실시예 65. 실시예 62-64 중 어느 한 방법에서, 상기 백신은 종양 백신이다.
- [0488] 실시예 66. 실시예 65의 방법에서, 상기 하나 이상의 펩타이드의 각각의 펩타이드에 대해, 상기 종양 백신은, 각각의 펩타이드를 암호화하는 뉴클레오타이드 서열, 각각의 펩타이드에 대응하는 아미노산 서열, 각각의 펩타이드에 대응하는 RNA, 각각의 펩타이드에 대응하는 DNA, 각각의 펩타이드에 대응하는 세포, 각각의 펩타이드에 대응하는 플라스미드, 또는 각각의 펩타이드에 대응하는 벡터 중 적어도 하나를 포함한다.
- [0489] 실시예 67. 실시예 62-66 중 어느 한 방법에서, 상기 백신은 부형제 또는 보조제 중 적어도 하나를 더 포함한다.
- [0490] 실시예 68. 실시예 62-67 중 어느 하나의 방법에서, 상기 백신은 RNA 분자를 포함하며, 상기 RNA 분자는 5'→3' 방향으로:
- [0491] 5' 캡,
- [0492] 5' 비번역 지역(UTR),
- [0493] 분비 신호 펩타이드를 암호화하는 폴리뉴클레오타이드 서열,
- [0494] 하나 이상의 펩타이드를 암호화하는 폴리뉴클레오타이드 서열,
- [0495] 주 조직적합성 복합체(MHC) 분자의 막횡단 및 세포질 도메인의 적어도 일부분을 암호화하는 폴리뉴클레오타이드 서열, AES(Amino-Terminal Enhancer of Split) mRNA 또는 이의 단편의 3' 비번역 지역, 및 미토콘드리아로 암호화된 12S RNA의 비-코딩 RNA 또는 이의 단편
- [0496] 을 포함하는 3' UTR, 및
- [0497] 폴리(A) 서열을 포함함을 포함한다.
- [0498] 실시예 69. 약학 조성물이 실시예 1-49 중 어느 한 방법에 의해 생성된 리포트에 기초하여 펩타이드의 세트 중에서 선택된 하나 이상의 펩타이드를 포함하며, 상기 하나 이상의 펩타이드는 상기 펩타이드의 세트의 불완전한 서브세트이다.
- [0499] 실시예 70. 약학 조성물이 실시예 1-49 중 어느 한 방법에 의해 생성된 리포트에 기초하여 펩타이드의 세트 중에서 선택된 하나 이상의 펩타이드를 암호화하는 핵산 서열을 포함하며 상기 하나 이상의 펩타이드는 상기 펩타이드의 세트의 불완전한 서브세트이다.
- [0500] 실시예 71. 면역원성 펩타이드가 실시예 1-49의 어느 하나의 실시예의 방법에 의해 생성된 리포트에 기초하여 식별된다.
- [0501] 실시예 72. 핵산 서열이 실시예 1-49 중 어느 하나의 방법에 의해 생성된 리포트에 기초하여 식별된다.
- [0502] 실시예 73. 실시예 72의 핵산 서열은, 상기 핵산 서열은 DNA 서열을 포함함을 포함한다.
- [0503] 실시예 74. 실시예 72 또는 실시예 73의 핵산 서열은, 상기 핵산 서열은 RNA 서열을 포함함을 포함한다.
- [0504] 실시예 75. 실시예 72-74 중 어느 한 핵산 서열은, 상기 핵산 서열은 mRNA 서열을 포함함을 포함한다.
- [0505] 실시예 76. 대상체를 치료하는 방법이 실시예 1-49 중 어느 한 방법에 의해 생성된 리포트에 기초하여 식별된 하나 이상의 펩타이드, 하나 이상의 약학 조성물, 또는 하나 이상의 핵산 서열 중 적어도 하나를 투여하는 단계를 포함한다.
- [0506] 실시예 77. 방법이, 대상체로부터 획득된 생물학적 샘플의 세트를 처리하여 펩타이드의 세트를 특징짓는 펩타이드 서열의 세트를 생성하는 단계, 상기 대상체로부터 획득된 생물학적 샘플의 세트를 처리하여 대상체의 면역단백질 복합체(IPC)에 대해 식별된 면역단백질 복합체(IPC) 서열을 생성하는 단계, 어텐션-기반 머신 러닝 모델의 초기 어텐션 서브시스템 내 제1 어텐션 블록을 이용해 펩타이드 서열의 세트를 나타내는 펩타이드 표현의 세트를 생성하는 단계, 상기 초기 어텐션 서브시스템 내 제2 어텐션 블록을 이용해 IPC 서열을 나타내는 면역단백질

복합체(IPC) 표현을 생성하는 단계, 펩타이드 표현의 세트 및 IPC 표현을 처리하여 출력을 생성하는 단계 — 상기 출력은 대응하는 펩타이드-IPC 결합체에 대한 상호작용 예측, 상호작용 친화도 예측, 면역원성 예측 중 적어도 하나를 포함하고, 대응하는 펩타이드-IPC 결합체는 상기 펩타이드의 세트의 펩타이드를 포함함 — 를 포함한다.

[0507] 실시예 78. 실시예 77의 방법에서, 상기 대상체로부터 획득된 생물학적 샘플의 세트를 처리하여 펩타이드 서열의 세트를 생성하는 것은: 상기 대상체로부터 획득된 생물학적 샘플의 세트 내 질병 샘플을 처리하여 펩타이드 서열의 세트를 생성하는 것을 포함한다.

[0508] 실시예 79. 실시예 77 또는 실시예 78의 방법은, 상기 대상체로부터 생물학적 샘플의 세트를 획득하는 단계 — 상기 생물학적 샘플의 세트는 질병 샘플을 포함함 — 를 더 포함한다.

[0509] 실시예 80. 실시예 77-79 중 어느 한 방법은, 상기 출력에 기초하여 리포트를 생성하는 단계를 더 포함한다.

[0510] 실시예 81. 방법이, 사용자 장치에서, 대상체에 대한 개별화된 백신을 설계하기 위한 요청을 수신하는 단계, 상기 사용자 장치로부터, 원격 시스템으로 통신을 전송하는 단계 — 상기 통신은 상기 대상체의 식별자를 포함하고, 상기 원격 시스템은: 펩타이드의 세트를 특징짓는 펩타이드 서열의 세트를 액세스하며 — 펩타이드 서열의 세트의 각각의 펩타이드 서열은 대상체로부터의 질병 샘플을 처리함으로써 식별되었음 —, 대상체의 면역단백질 복합체(IPC: immunoprotein complex)에 대해 식별된 면역단백질 복합체(IPC) 서열을 액세스하고, 어텐션-기반 머신 러닝 모델의 초기 어텐션 서브시스템 내 제1 어텐션 블록을 이용해 펩타이드 서열의 세트를 나타내는 펩타이드 표현의 세트 및 초기 어텐션 서브시스템 내 제2 어텐션 블록을 이용해 IPC 서열을 나타내는 면역단백질 복합체(IPC) 표현을 처리하여 출력을 생성하며 — 상기 출력은 대응하는 펩타이드-IPC 결합체에 대한 상호작용 예측, 상호작용 친화도 예측, 또는 면역원성 예측 중 적어도 하나를 포함함 —, 및 상기 출력에 기초하여 리포트를 생성하고, 상기 사용자 장치로 상기 리포트를 전송하며, 상기 사용자 장치에서 상기 리포트를 수신하도록 구성됨 — 을 포함한다.

[0511] 실시예 82. 실시예 81의 방법은, 상기 대상체로부터의 질병 샘플을 수집하는 단계, 크로마토그래피 또는 질량 분석법 중 적어도 하나를 이용하여 상기 질병 샘플 내 MHC 분자로부터의 펩타이드의 세트를 포함하는 다중 펩타이드를 용리하는 단계, 펩타이드의 세트를 시퀀싱하여 초기 서열의 세트를 생성하는 단계, 상기 초기 서열의 세트의 각각의 초기 서열을 참조 서열에 비교하는 단계, 및 비교에 기초하여 펩타이드 서열의 세트를 정의하는 단계 — 상기 펩타이드 서열의 세트 내 각각의 펩타이드 서열은 상기 참조 서열에 대한 변이를 포함하는 변이-코딩 서열임 — 를 더 포함한다.

[0512] 실시예 83. 대상체에 대한 치료제를 제조하기 위한 방법이 제공된다. 상기 방법은 컴퓨팅 장치로부터 리포트를 수신하는 단계 — 상기 컴퓨팅 장치는: 펩타이드의 세트를 특징짓는 펩타이드 서열의 세트를 액세스하며 — 펩타이드 서열의 세트의 각각의 펩타이드 서열은 대상체로부터의 질병 샘플을 처리함으로써 식별되었음 —, 대상체의 면역단백질 복합체(IPC: immunoprotein complex)에 대해 식별된 면역단백질 복합체(IPC) 서열을 액세스하고, 어텐션-기반 머신 러닝 모델의 초기 어텐션 서브시스템 내 제1 어텐션 블록을 이용해 펩타이드 서열의 세트를 나타내는 펩타이드 표현의 세트 및 초기 어텐션 서브시스템 내 제2 어텐션 블록을 이용해 IPC 서열을 나타내는 면역단백질 복합체(IPC) 표현을 처리하여 출력을 생성하며 — 상기 출력은 대응하는 펩타이드-IPC 결합체에 대한 상호작용 예측, 상호작용 친화도 예측, 또는 면역원성 예측 중 적어도 하나를 포함함 —, 및 상기 출력에 기초하여 리포트를 생성하고, 상기 리포트에 기초하여 치료제를 제조하기 위한 치료제 제조 계획을 생성하도록 구성됨 — 을 포함한다.

[0513] 실시예 84. 실시예 83의 방법은, 상기 치료제 제조 계획에 기초하여 치료제를 제조하는 단계를 더 포함한다.

[0514] 실시예 85. 방법이, 복수의 돌연변이 펩타이드를 특징짓는 복수의 변이-코딩 서열을 어텐션-기반 머신 러닝 모델로 입력하는 단계 — 복수의 변이-코딩 서열의 각각의 변이-코딩 서열은 대상체로부터의 질병 샘플을 처리함으로써 식별되었음 —, 상기 대상체의 면역단백질 복합체(IPC)에 대해 식별된 면역단백질 복합체(IPC) 서열을 어텐션-기반 머신 러닝 모델에 입력하는 단계 —상기 어텐션-기반 머신 러닝 모델은 어텐션-기반 머신 러닝 모델의 초기 어텐션 서브시스템 내 제1 어텐션 블록을 이용해 복수의 변이-코딩 서열을 나타내는 복수의 변이 표현을 처리하고 초기 어텐션 서브시스템 내 제2 어텐션 블록을 이용해 IPC 서열을 나타내는 면역단백질(IPC) 표현을 처리하여 출력을 생성하도록 구성되고, 상기 출력은 상호작용 예측, 상호작용 친화도 예측, 또는 대응하는 돌연변이 펩타이드-IPC 결합체에 대한 면역원성 예측 중 적어도 하나를 포함함 —, 및 상기 출력에 기초하여 생성된 리포트를 수신하는 단계, 및 상기 리포트에 기초하여, 복수의 돌연변이 펩타이드 중 상기 대상체에 대한

치료제에서 사용되기 위한 서브세트를 선택하는 단계를 포함한다.

- [0515] 실시예 86. 방법이, 돌연변이 펩타이드를 특징짓는 펩타이드 서열을 수신하는 단계 — 상기 펩타이드 서열은 대응하는 참조 서열에 대한 변이를 포함함 — , 주 조직적합성 복합체(MHC)에 대해 식별된 MHC 서열을 수신하는 단계, 어텐션-기반 머신 러닝 모델 내 상이한 처리 경로를 이용해 펩타이드 서열 및 MHC 서열을 처리하여 출력을 생성하는 단계 — 상기 출력은 돌연변이 펩타이드와 MHC 모두와 관련된 면역학적 활동에 대한 정보를 제공함 — , 및 상기 출력에 기초하여 리포트를 생성하는 단계를 포함한다.
- [0516] 실시예 87. 실시예 86의 방법에서, 처리하는 것은 상기 어텐션-기반 머신 러닝 모델 내 펩타이드 처리 경로를 통해 펩타이드 서열을 처리하는 것 — 상기 펩타이드 처리 경로는 제1 임베딩 블록 및 적어도 하나의 셀프-어텐션 층을 포함하는 제1 어텐션 블록을 포함함 — ,
- [0517] 상기 어텐션-기반 머신 러닝 모델 내 MHC 처리 경로를 통해 MHC 서열을 처리하는 것 — 상기 MHC 처리 경로는 제2 임베딩 블록 및 적어도 하나의 셀프-어텐션 층을 포함하는 제2 어텐션 블록을 포함함 — 을 포함한다.
- [0518] 실시예 88. 실시예 87의 방법에서, T 세포 수용체(TCR)에 대해 식별된 TCR 서열을 수신하는 단계를 더 포함하며, 상기 처리하는 것은:상기 어텐션-기반 머신 러닝 모델 내 TCR 처리 경로를 통해 TCR 서열을 처리하는 것 — 상기 TCR 처리 경로는 제3 임베딩 블록 및 적어도 하나의 셀프-어텐션 층을 포함하는 제3 어텐션 블록을 포함함 — 을 더 포함한다.
- [0519] 실시예 89. 실시예 86-88 중 어느 한 방법에서, 상기 면역학적 활동은 면역 반응을 포함하고 상기 정보는 상기 면역 반응을 유발할 수 있는 돌연변이 펩타이드의 능력에 대한 예측을 포함한다.
- [0520] 실시예 90. 실시예 86-89 중 어느 한 방법에서, 상기 처리하는 것은: 펩타이드 처리 경로를 통해 펩타이드 서열의 변환된 펩타이드 표현을 생성하는 것, MHC 처리 경로를 통해 MHC 서열의 변환된 MHC 표현을 생성하는 것, 변환된 펩타이드 표현 및 변환된 MHC 표현을 이용해 복합 표현을 생성하는 것, 상기 복합 표현을 처리하여 출력을 생성하는 것을 포함한다.
- [0521] 실시예 91. 실시예 86-90 중 어느 한 방법에서, 상기 면역학적 활동은 돌연변이 펩타이드의 MHC로의 결합을 포함하고 상기 출력은 상기 돌연변이 펩타이드가 MHC에 결합하는지 여부에 대응하는 제1 예측 또는 결합과 연관된 친화도에 대응하는 제2 예측 중 적어도 하나를 포함한다.
- [0522] 실시예 92. 실시예 86-91 중 어느 한 방법은, 상기 리포트에 기초하여 면역요법을 위한 표적으로서 상기 돌연변이 펩타이드를 포함하도록 결정하는 단계를 더 포함한다.
- [0523] 실시예 93. 실시예 92의 방법에서, 상기 면역요법은 T 세포 요법, 개별화된 암 요법, 항원-특이적 면역요법, 항원-의존적 면역요법, 백신 및 자연 살해(NK) 세포 요법으로 구성된 군 중에서 선택된다.
- [0524] 실시예 94. 실시예 86-93 중 어느 한 방법은, 상기 리포트에 기초하여 면역요법을 위한 표적으로서 상기 돌연변이 펩타이드를 배제하도록 결정하는 것 중 적어도 하나를 더 포함한다.
- [0525] 실시예 95. 실시예 94의 방법에서, 상기 면역요법은 T 세포 요법, 개별화된 암 요법, 항원-특이적 면역요법, 항원-의존적 면역요법, 백신 및 자연 살해(NK) 세포 요법으로 구성된 군 중에서 선택된다.
- [0526] 실시예 96. 실시예 86-95 중 어느 한 방법은, 상기 리포트에 기초하여, 상기 돌연변이 펩타이드, 상기 돌연변이 펩타이드의 전구체, 상기 돌연변이 펩타이드를 암호화하는 핵산, 또는 상기 돌연변이 펩타이드를 발현하는 복수의 세포 중 적어도 하나를 치료제에 포함하도록 결정하는 단계, 및 상기 치료제를 제조하는 단계를 더 포함한다.
- [0527] 실시예 97. 실시예 96의 방법은, 상기 치료제로 대상체를 치료하는 단계를 더 포함한다.
- [0528] 실시예 98. 실시예 86-97 중 어느 한 방법에서, 돌연변이 펩타이드를 특징짓는 펩타이드 서열은 대상체로부터의 질병 샘플을 시퀀싱함으로써 식별되었고, 상기 펩타이드 서열은 대응하는 참조 서열에 비해 적어도 하나의 서열 변이를 가지며, 치료제는 상기 리포트에 기초하여 상기 대상체에 대해 설계된다.
- [0529] 실시예 99. 방법이 돌연변이 펩타이드를 특징짓는 펩타이드 서열을 수신하는 단계 — 상기 펩타이드 서열은 대응하는 참조 서열에 대한 변이를 포함함 — , T 세포 수용체(TCR)에 대해 식별된 TCR 서열을 수신하는 단계, 어텐션-기반 머신 러닝 모델 내 상이한 처리 경로를 이용해 펩타이드 서열 및 TCR 서열을 처리하여 출력을 생성하는 단계 — 상기 출력은 돌연변이 펩타이드와 TCR 모두와 관련된 면역학적 활동에 대한 정보를 제공함 — , 및 상

기 출력에 기초하여 리포트를 생성하는 단계를 포함한다.

- [0530] 실시예 100. 실시예 99의 방법에서, 상기 처리하는 것은: 상기 어텐션-기반 머신 러닝 모델 내 펩타이드 처리 경로를 통해 펩타이드 서열을 처리하는 것 — 상기 펩타이드 처리 경로는 제1 임베딩 블록 및 제1 어텐션 블록을 포함함 —, 및 상기 어텐션-기반 머신 러닝 모델 내 TCR 처리 경로를 통해 TCR 서열을 처리하는 것 — 상기 TCR 처리 경로는 제2 임베딩 블록 및 제2 어텐션 블록을 포함함 — 을 더 포함한다.
- [0531] 실시예 101. 실시예 100의 방법은, 주 조직적합성 복합체(MHC)에 대해 식별된 MHC 서열을 수신하는 단계를 더 포함하며, 상기 처리하는 것은: 상기 어텐션-기반 머신 러닝 모델 내 MHC 처리 경로를 통해 MHC 서열을 처리하는 것 — 상기 MHC 처리 경로는 제3 임베딩 블록 및 MHC 제3 블록을 포함함 — 을 더 포함한다.
- [0532] 실시예 102. 실시예 99-101 중 어느 한 방법은, 상기 면역학적 활동은 면역 반응을 포함하고 상기 정보는 상기 면역 반응을 유발할 수 있는 돌연변이 펩타이드의 능력에 대한 예측을 포함함을 포함한다.
- [0533] 실시예 103. 실시예 99-102에서, 상기 처리하는 것은: 펩타이드 처리 경로를 통해 펩타이드 서열의 변환된 펩타이드 표현을 생성하는 것, TCR 처리 경로를 통해 TCR 서열의 변환된 TCR 표현을 생성하는 것, 변환 펩타이드 표현 및 변환 TCR 표현을 이용해 복합 표현을 생성하는 것, 상기 복합 표현을 처리하여 출력을 생성하는 것을 포함한다.
- [0534] 실시예 104. 실시예 99-103 중 어느 한 방법에서, 상기 면역학적 활동은 돌연변이 펩타이드의 MHC로의 결합을 포함하고 상기 출력은 상기 돌연변이 펩타이드가 MHC에 결합하는지 여부에 대응하는 제1 예측 또는 결합과 연관된 친화도에 대응하는 제2 예측 중 적어도 하나를 포함한다.
- [0535] 실시예 105. 실시예 99-104 중 어느 한 방법은, 상기 리포트에 기초하여 면역요법을 위한 표적으로서 상기 돌연변이 펩타이드를 포함하도록 결정하는 단계를 더 포함한다.
- [0536] 실시예 106. 실시예 105의 방법에서, 상기 면역요법은 T 세포 요법, 개별화된 암 요법, 항원-특이적 면역요법, 항원-의존적 면역요법, 백신 및 자연 살해(NK) 세포 요법으로 구성된 군 중에서 선택된다.
- [0537] 실시예 107. 실시예 99-106 중 어느 한 방법은, 상기 리포트에 기초하여 면역요법을 위한 표적으로서 상기 돌연변이 펩타이드를 배제하도록 결정하는 것 중 적어도 하나를 더 포함한다.
- [0538] 실시예 108. 실시예 107의 방법에서, 상기 면역요법은 T 세포 요법, 개별화된 암 요법, 항원-특이적 면역요법, 항원-의존적 면역요법, 백신 및 자연 살해(NK) 세포 요법으로 구성된 군 중에서 선택된다.
- [0539] 실시예 109. 실시예 99-108 중 어느 한 방법은, 상기 리포트에 기초하여, 상기 돌연변이 펩타이드, 상기 돌연변이 펩타이드의 전구체, 상기 돌연변이 펩타이드를 암호화하는 핵산, 또는 상기 돌연변이 펩타이드를 발현하는 복수의 세포 중 적어도 하나를 치료제에 포함하도록 결정하는 단계, 및 상기 치료제를 제조하는 단계를 더 포함한다.
- [0540] 실시예 110. 실시예 109의 방법은, 상기 치료제로 대상체를 치료하는 단계를 더 포함한다.
- [0541] 실시예 111. 실시예 99-110 중 어느 한 방법에서, 돌연변이 펩타이드를 특징짓는 펩타이드 서열은 대상체로부터의 질병 샘플을 시퀀싱함으로써 식별되었고, 상기 펩타이드 서열은 대응하는 참조 서열에 비해 적어도 하나의 서열 변이를 가지며, 치료제는 상기 리포트에 기초하여 상기 대상체에 대해 설계된다.
- [0542] 실시예 112. 시스템이, 하나 이상의 데이터 프로세서, 및하나 이상의 데이터 프로세서 상에서 실행될 때, 상기 하나 이상의 데이터 프로세서로 하여금 실시예 1-49, 77-81, 83, 85-95, 및 99-108 중 어느 하나를 수행하게 하는 명령을 포함하는 비-일시적 컴퓨터 판독형 저장 매체를 포함한다.
- [0543] 실시예 113. 하나 이상의 데이터 프로세서로 하여금 실시예 1-49, 77-81, 83, 85-95, 및 99-108 중 어느 하나를 수행하게 하도록 구성된 명령을 포함하며, 비일시적 기계 판독형 저장 매체에서 유형적으로 구현되는 컴퓨터 프로그램 프로덕트.
- [0544] IX. 추가 고려 사항
- [0545] 본 개시내용의 일부 실시예는 하나 이상의 데이터 프로세서를 포함하는 시스템을 포함한다. 일부 실시예에서, 시스템은 하나 이상의 데이터 프로세서 상에서 실행될 때 하나 이상의 데이터 프로세서로 하여금 여기서 개시된 하나 이상의 방법 중 일부 또는 전부 및/또는 하나 이상의 프로세스 중 일부 또는 전부를 수행하게 하는 명령을 포함하는 비일시적 컴퓨터 판독형 저장 매체를 포함한다. 본 개시내용의 일부 실시예는 하나 이상의 데이터 프

로세서 상에서 실행될 때 하나 이상의 데이터 프로세서로 하여금 여기서 개시된 하나 이상의 방법 중 일부 또는 전부 및/또는 여기서 개시된 하나 이상의 프로세스의 일부 또는 전부를 수행하게 하는 명령을 포함하는 비일시적 기계 판독형 저장 매체로 유형으로 구현되는 컴퓨터-프로그램 프로덕트를 포함한다.

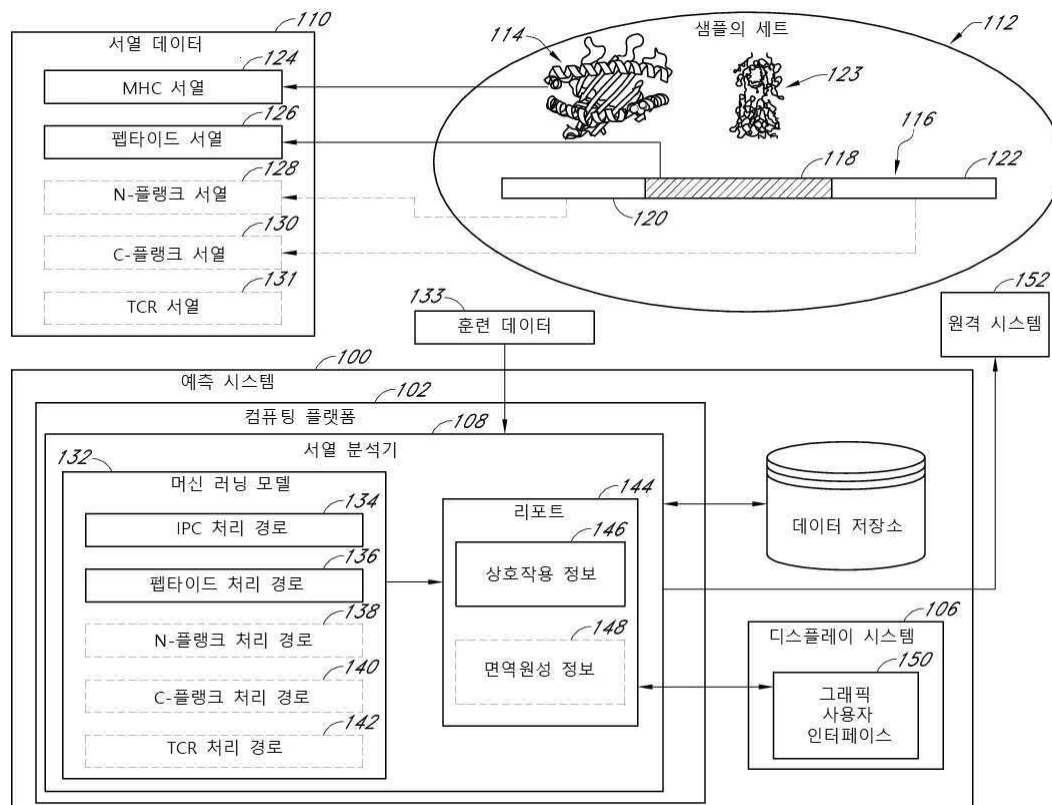
[0546] 채용된 용어 및 표현은 한정이 아닌 설명 측면에서 사용되며, 이러한 용어 및 표현의 사용에 도시되고 기재된 특징부 또는 이의 일부분의 임의의 균등물을 배제하는 어떠한 의도도 없고, 청구된 발명의 범위 내에서 다양한 수정이 가능함이 자명하다. 따라서, 본 발명이 실시예 및 선택적 특징으로 특정하게 개시되었지만 본 명세서에 개시된 개념의 수정, 및 변형이 해당 분야의 통상의 기술자에 의해 이용될 수 있고, 이러한 수정 및 변형이 첨부된 청구항에 의해 정의된 본 발명의 범위 내에 있는 것으로 간주됨이 이해될 것이다.

[0547] 설명은 단지 바람직한 예시적인 실시 형태를 제공하며, 본 개시내용의 범위, 적용 가능성 또는 구성을 제한하도록 의도되지 않는다. 오히려, 바람직한 예시적인 실시예의 본 설명은 다양한 실시예를 구현하기 위한 가능한 설명을 통상의 기술자에게 제공할 것이다. 첨부된 특허청구범위에 기재된 정신 및 범위를 벗어나지 않고 요소의 기능 및 배열에 다양한 변경이 이루어질 수 있음을 이해해야 한다.

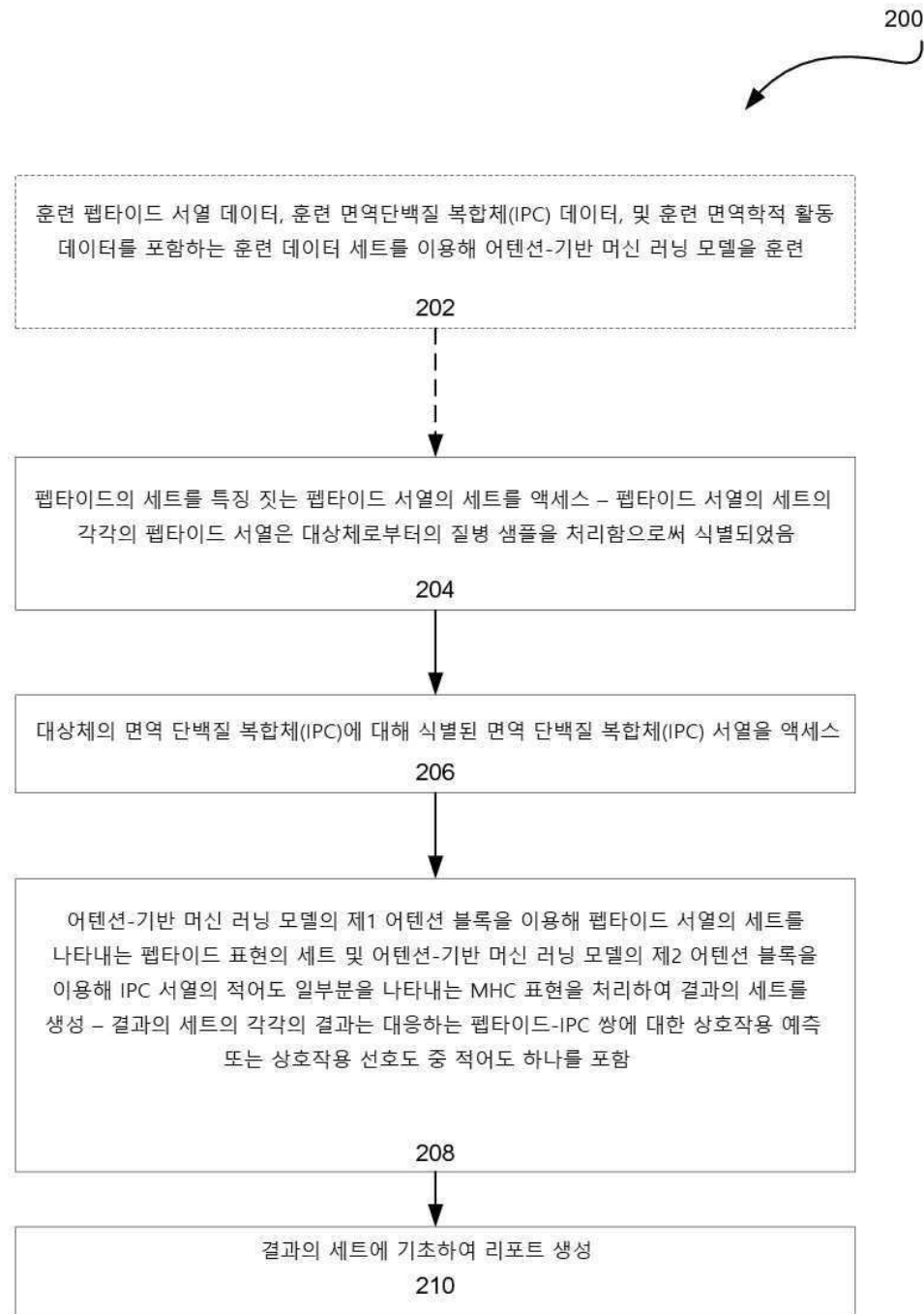
[0548] 실시예의 완전한 이해를 제공하기 위해 다음 설명에서 특정 세부사항이 제공된다. 그러나, 이러한 특정 세부사항 없이 실시예가 실시될 수 있음이 이해될 것이다. 예를 들어, 회로, 시스템, 네트워크, 프로세스 및 그 밖의 다른 구성요소는 실시예를 불필요한 세부 사항으로 모호하게 하지 않기 위해 블록도 형태의 구성요소로 표시될 수 있다. 또 다른 예에서, 잘 알려진 회로, 프로세스, 알고리즘, 구조 및 기술은 실시예를 모호하게 하는 것을 피하기 위해 불필요한 세부 사항 없이 표시될 수 있다.

도면

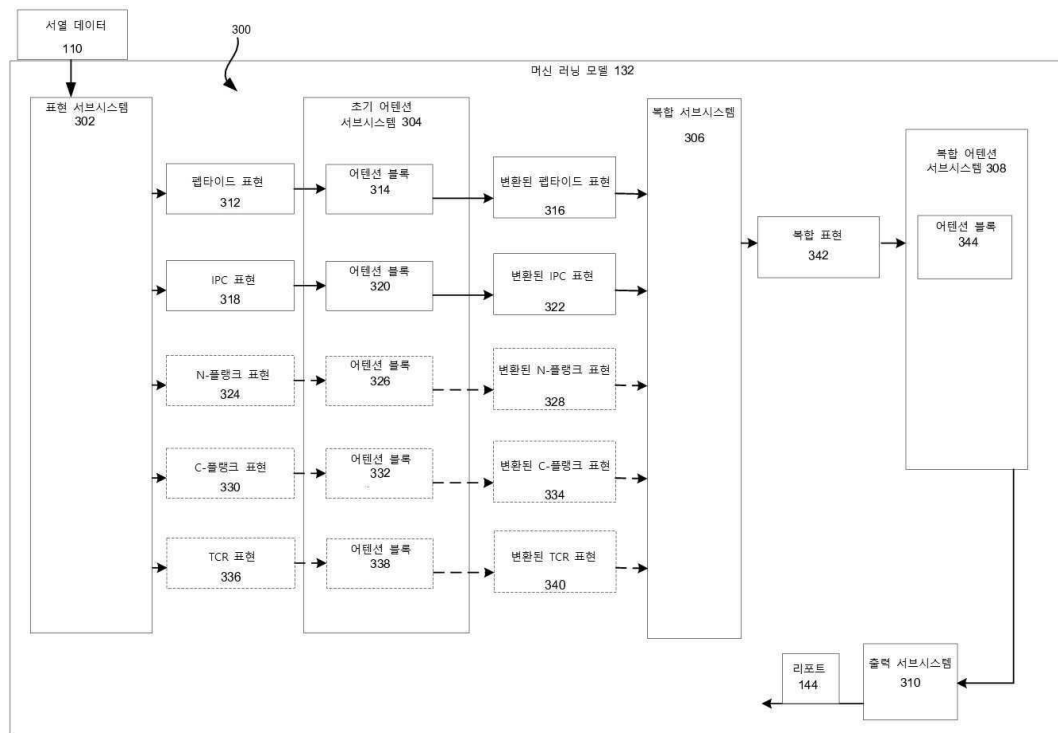
도면1



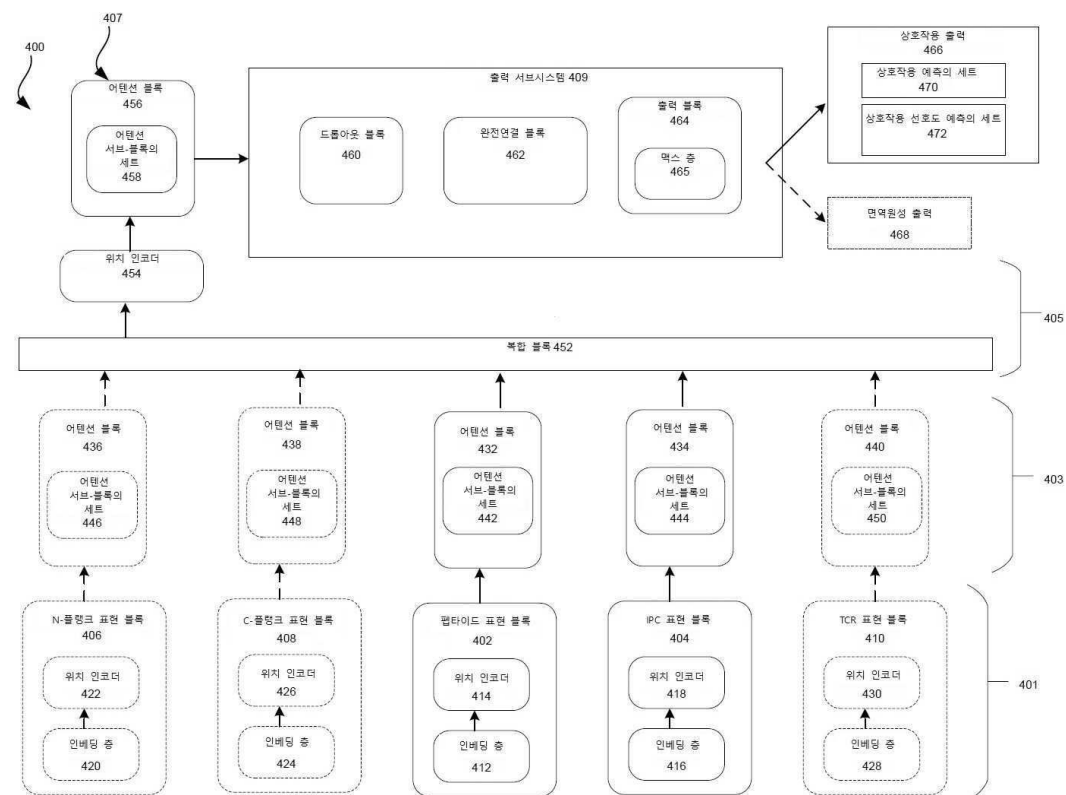
도면2



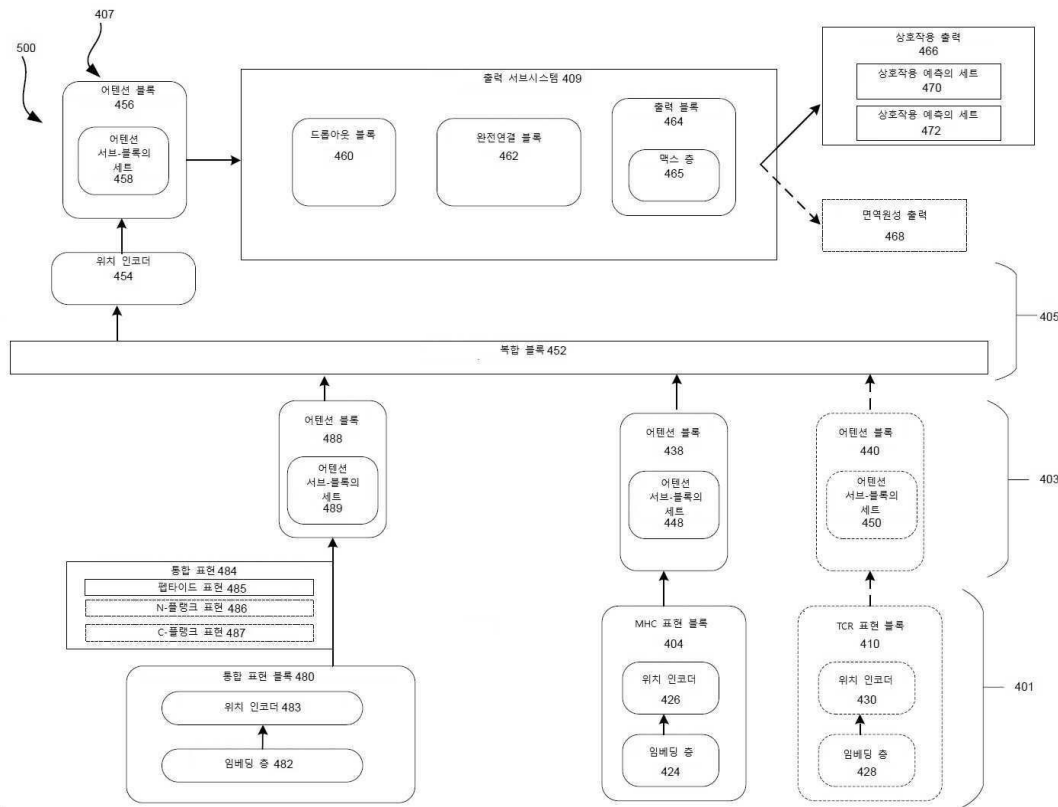
도면3



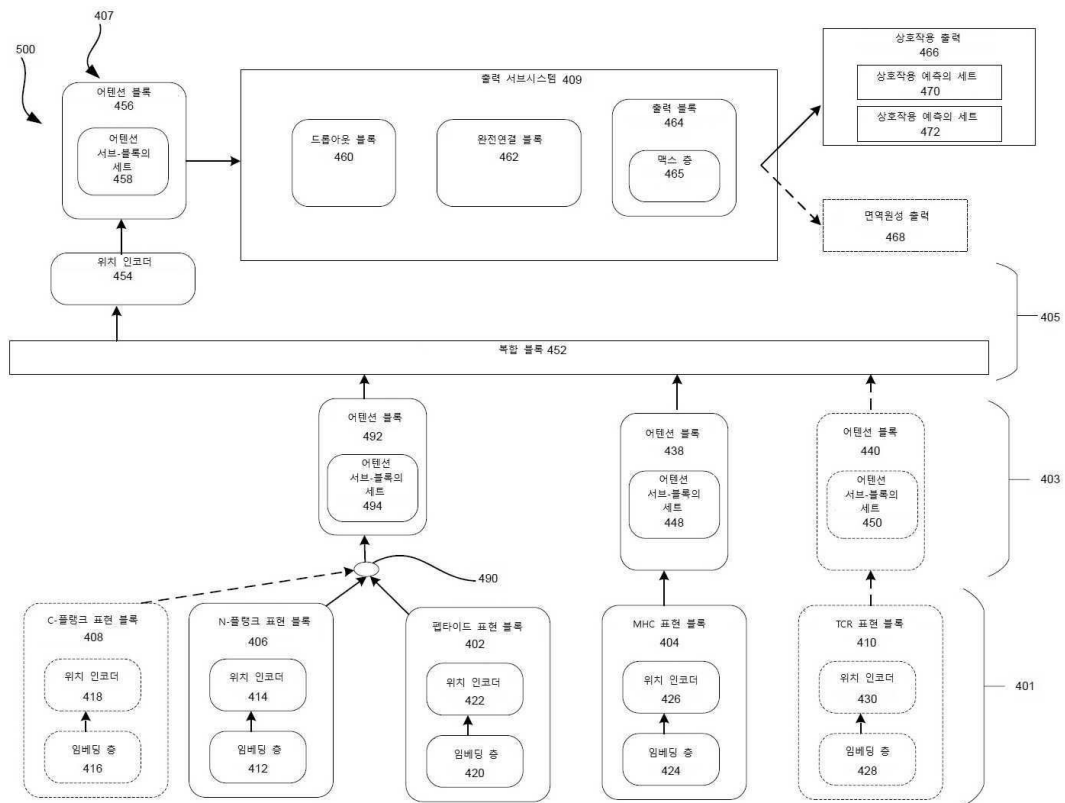
도면4a



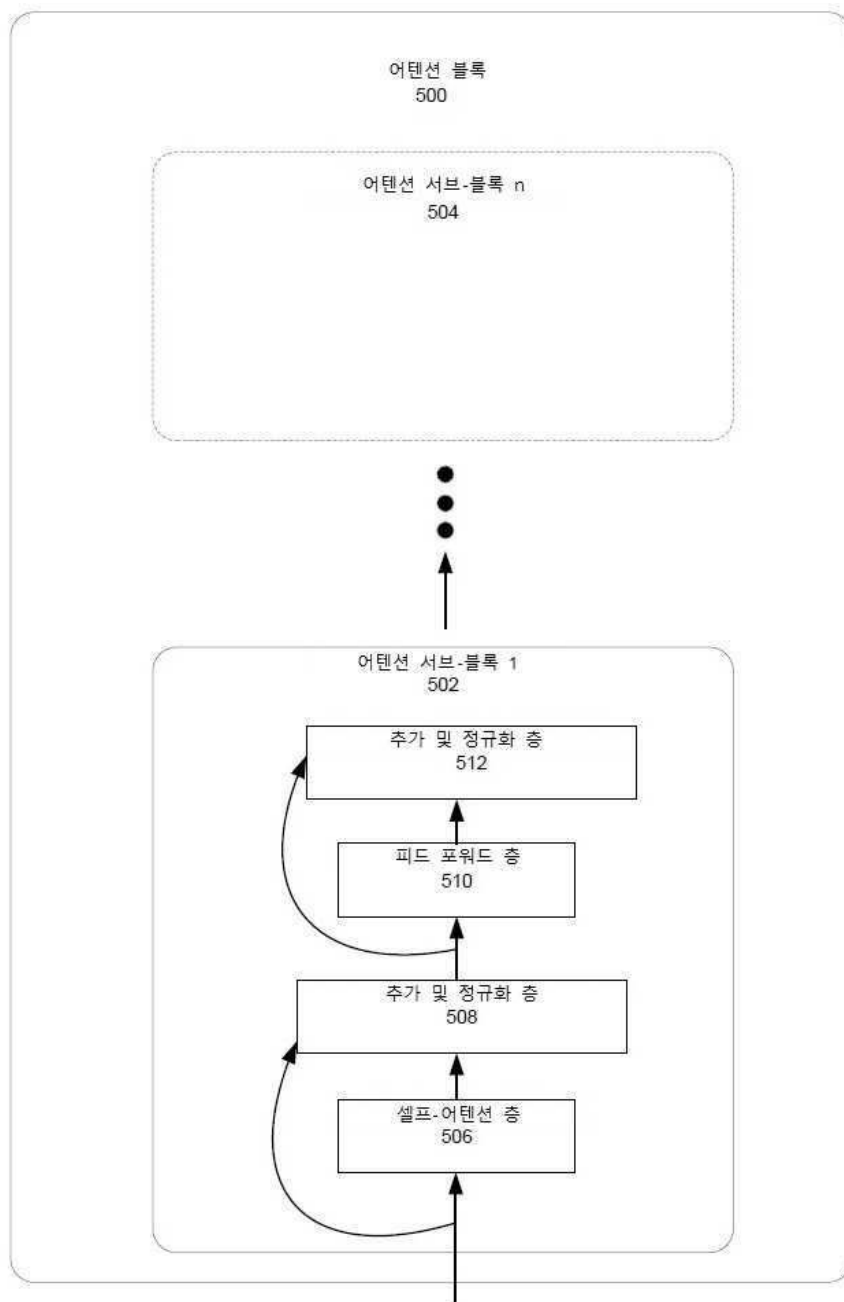
도면4b



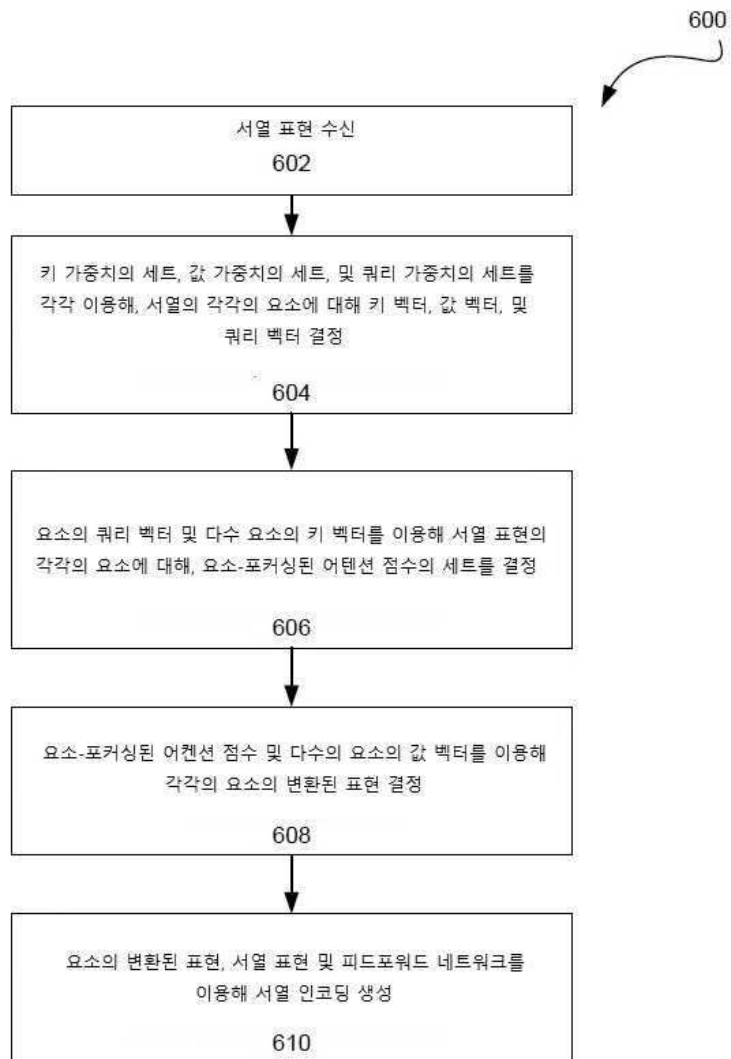
도면4c



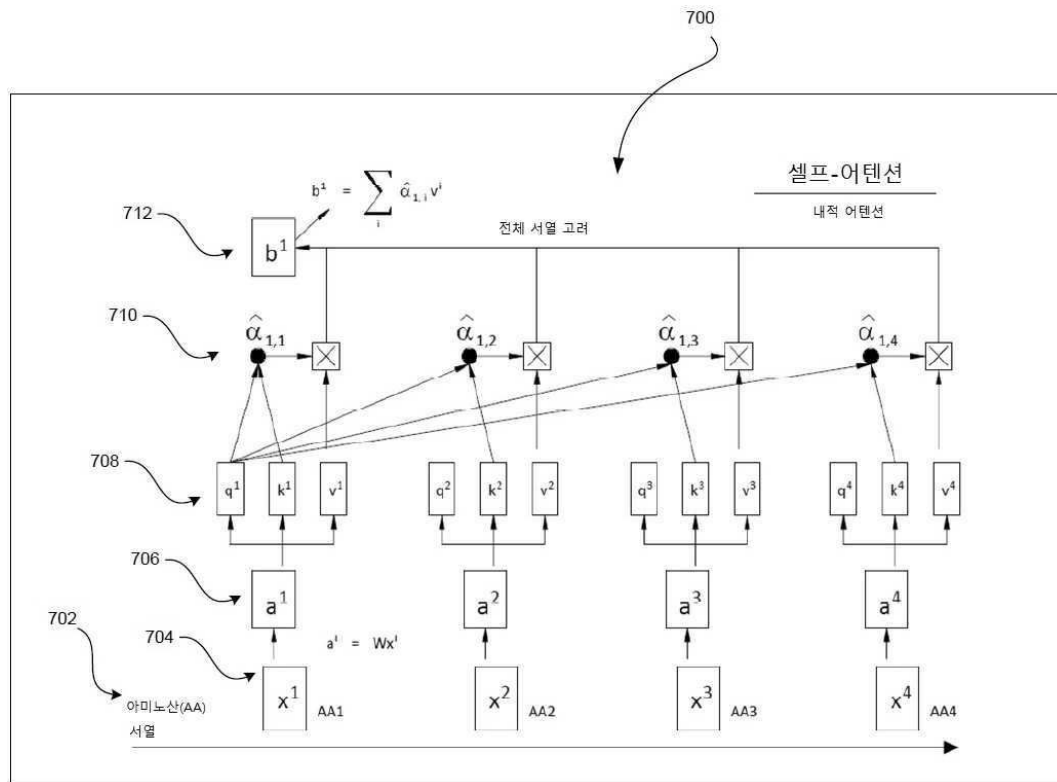
도면5



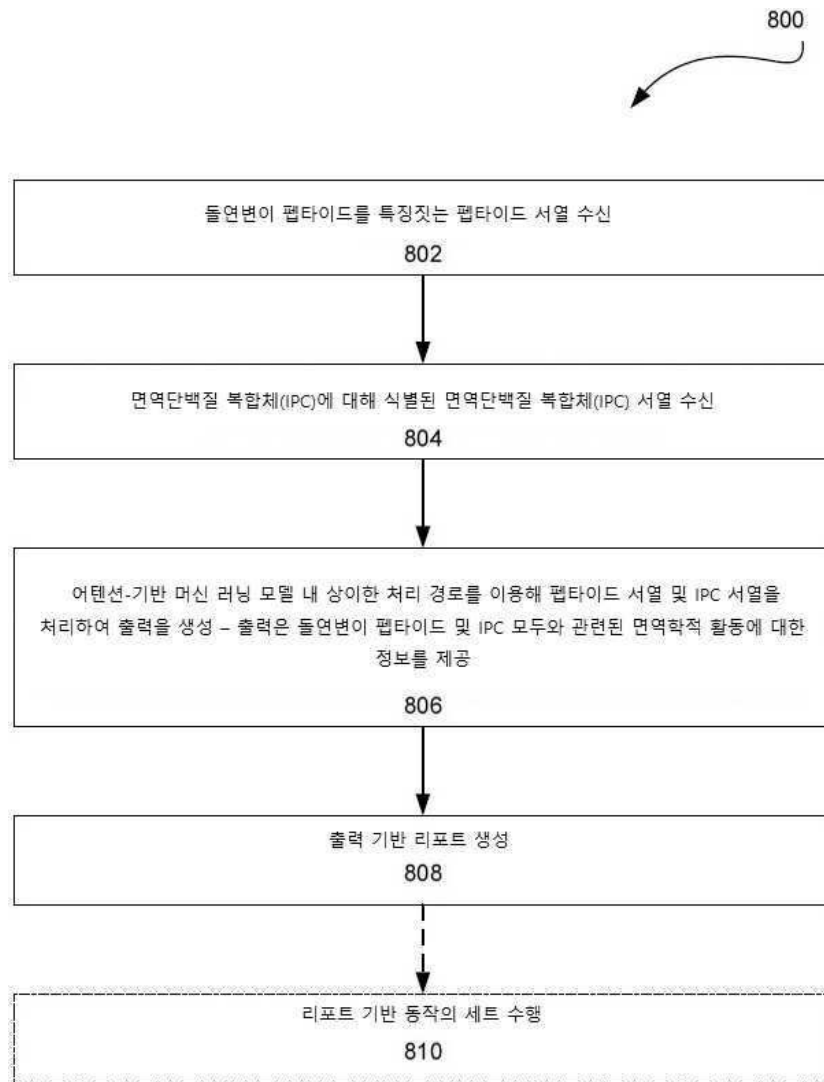
도면6



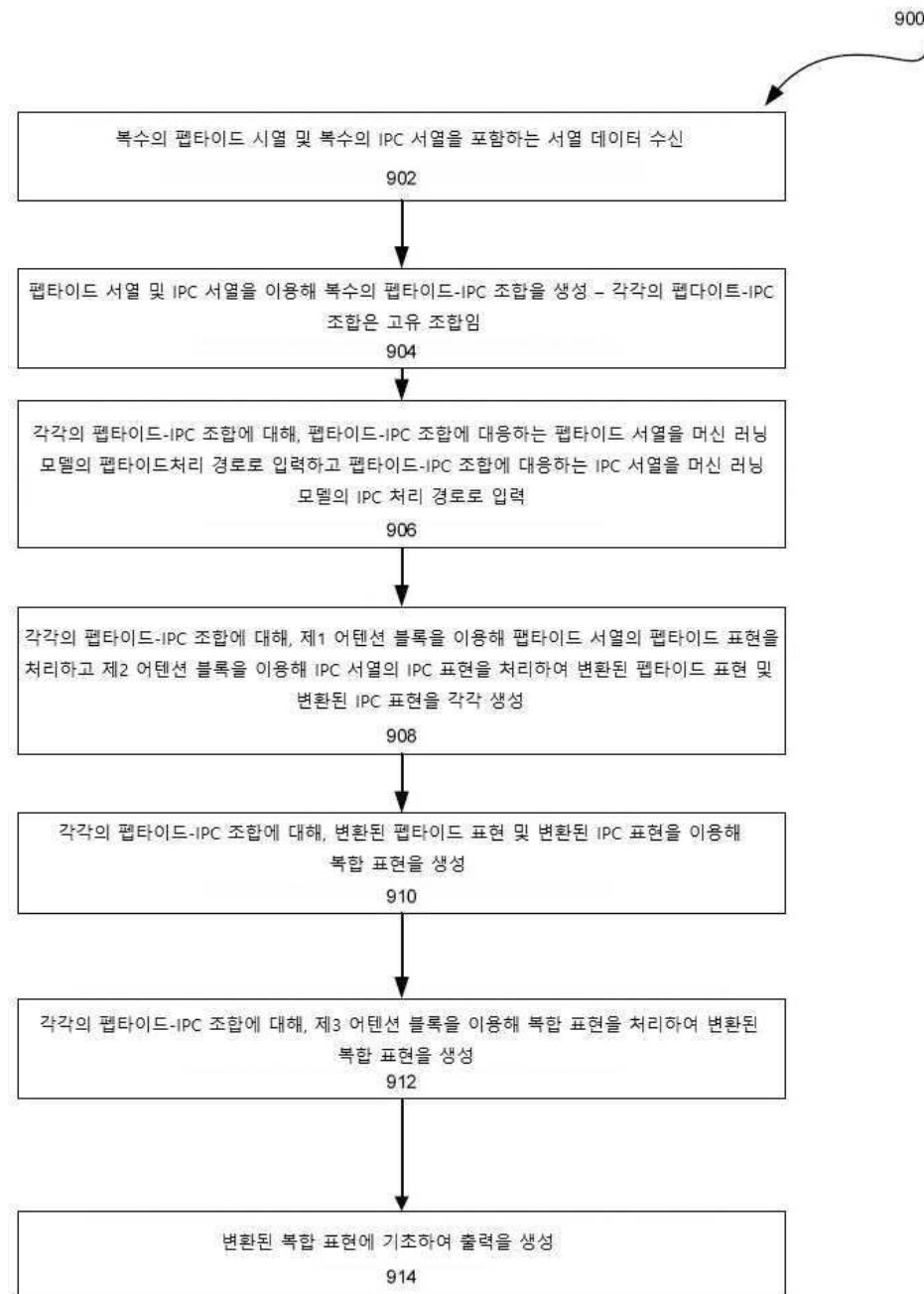
도면7



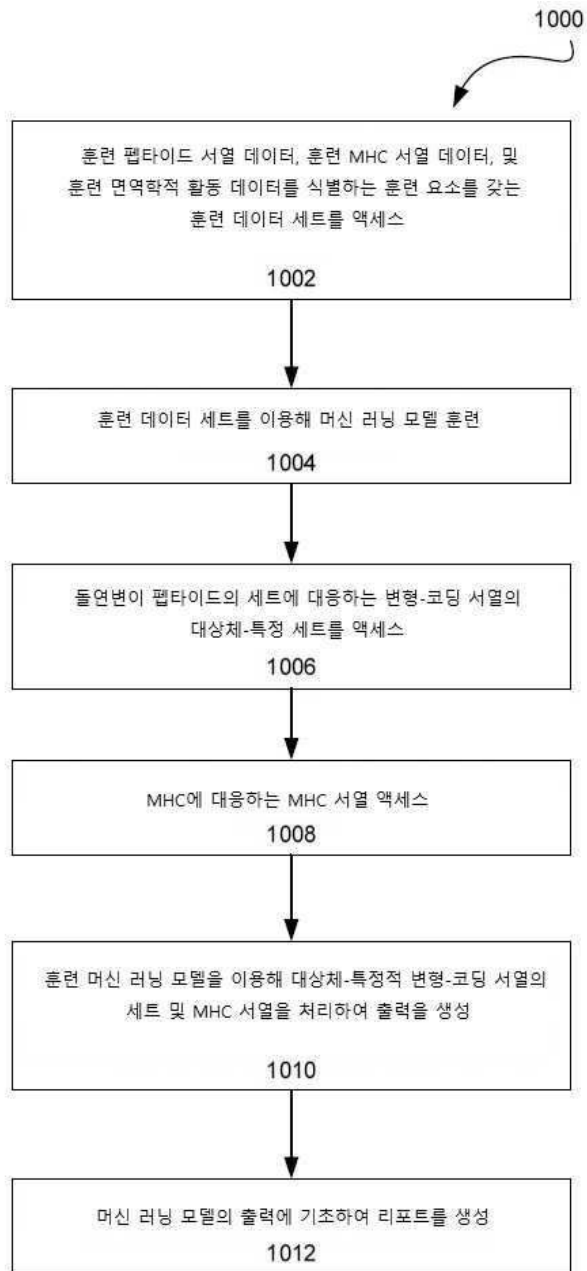
도면8



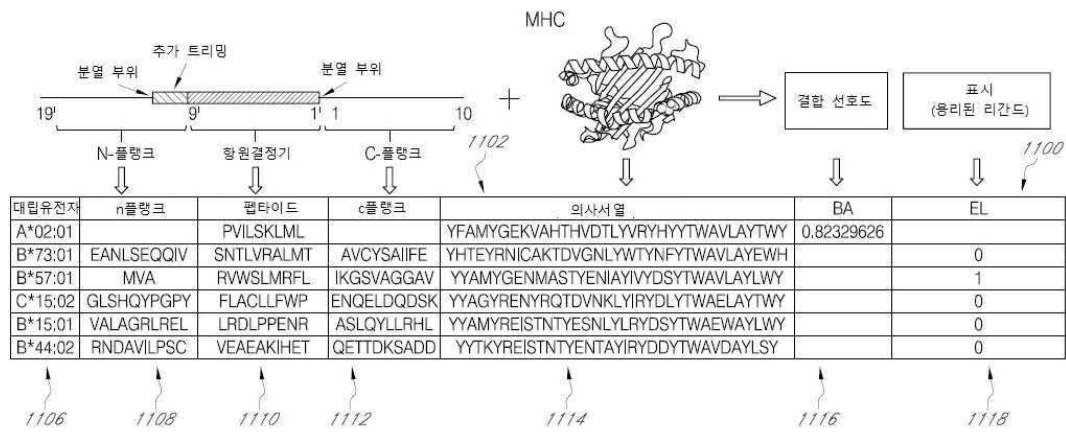
도면9



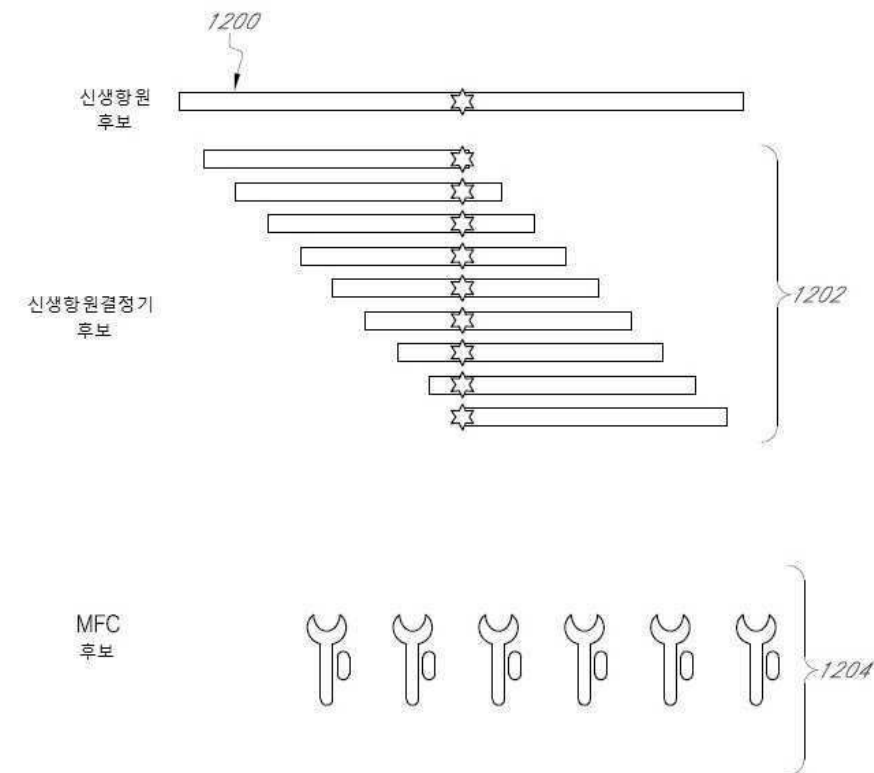
도면10



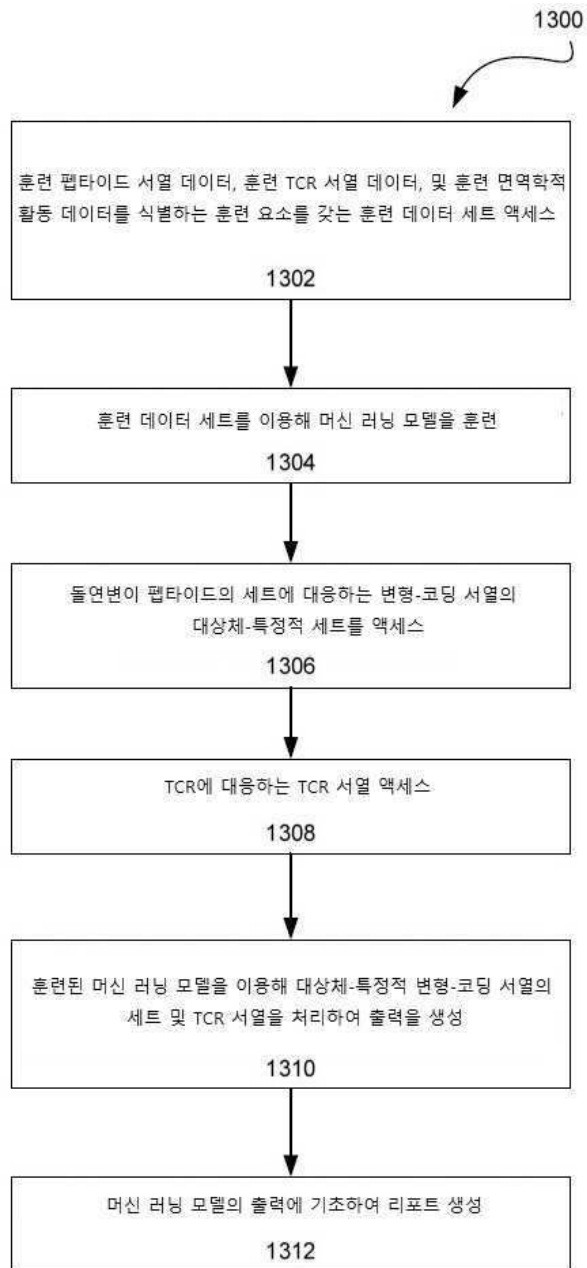
도면11



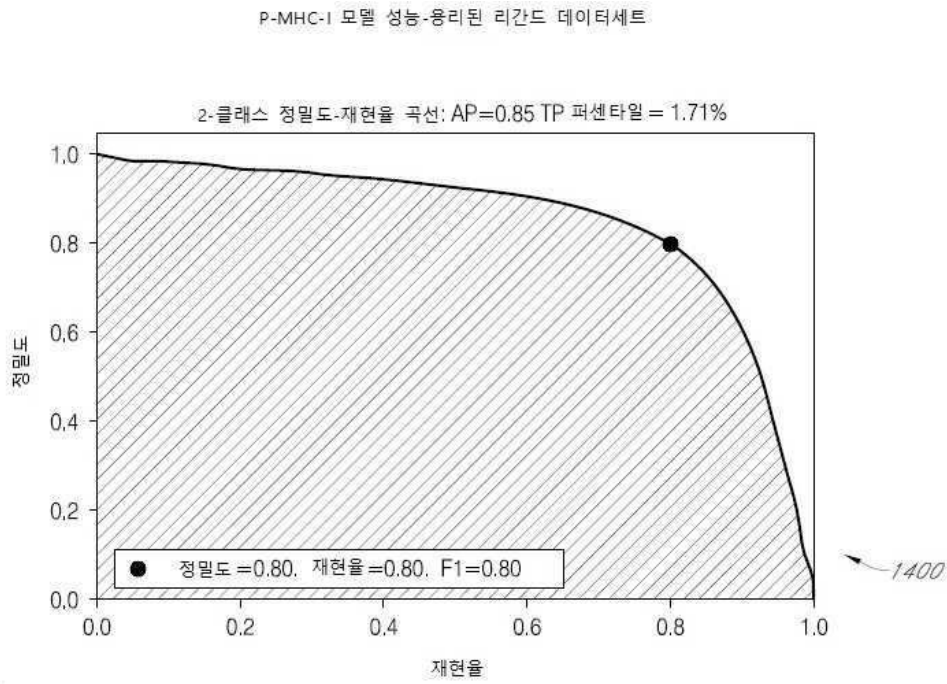
도면12



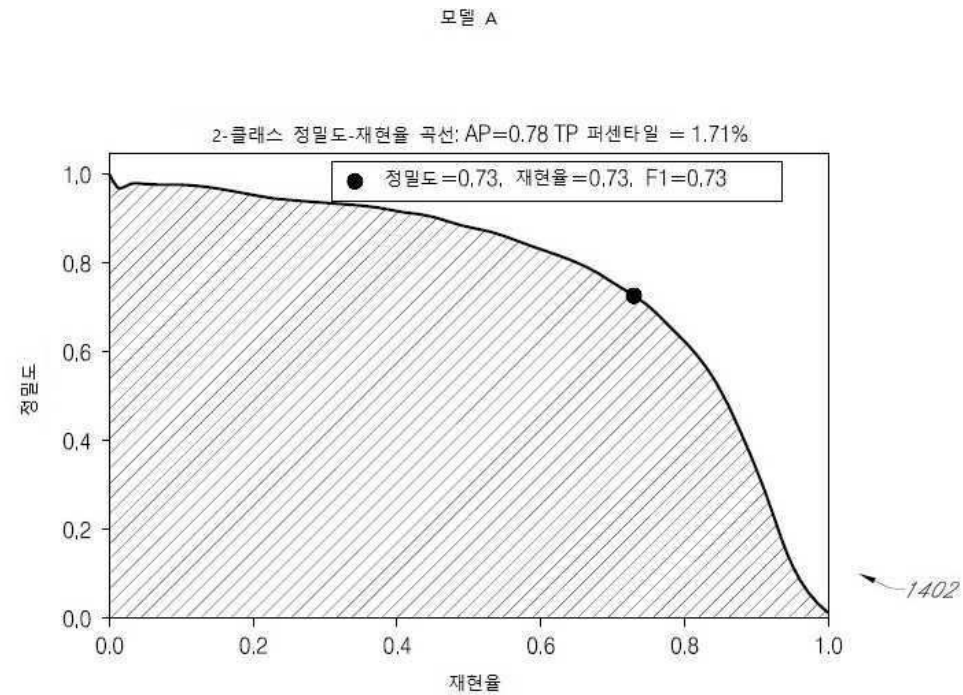
도면13



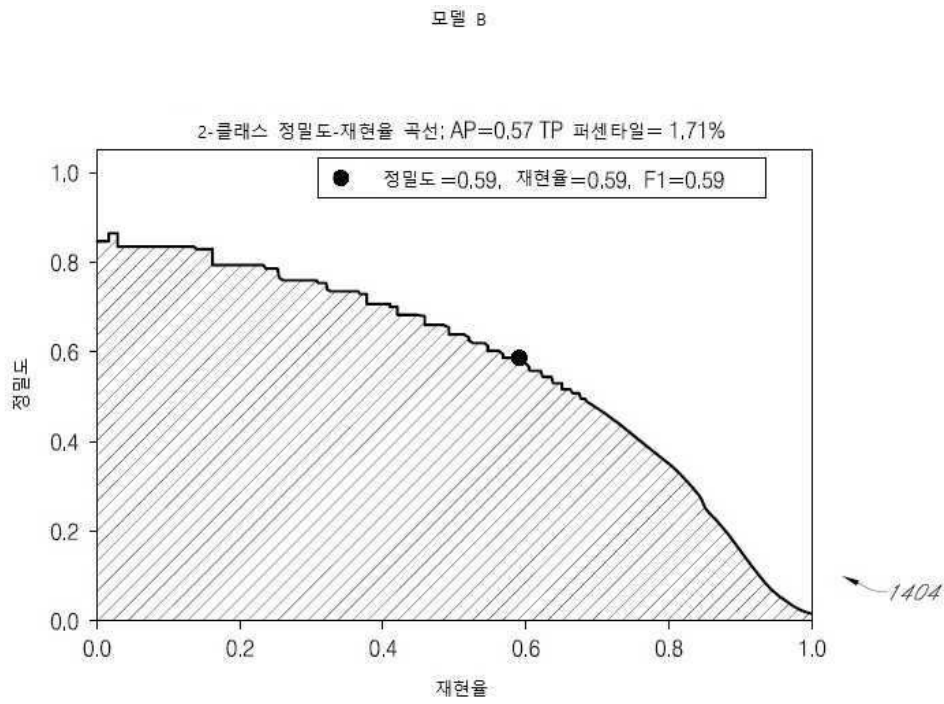
도면14a



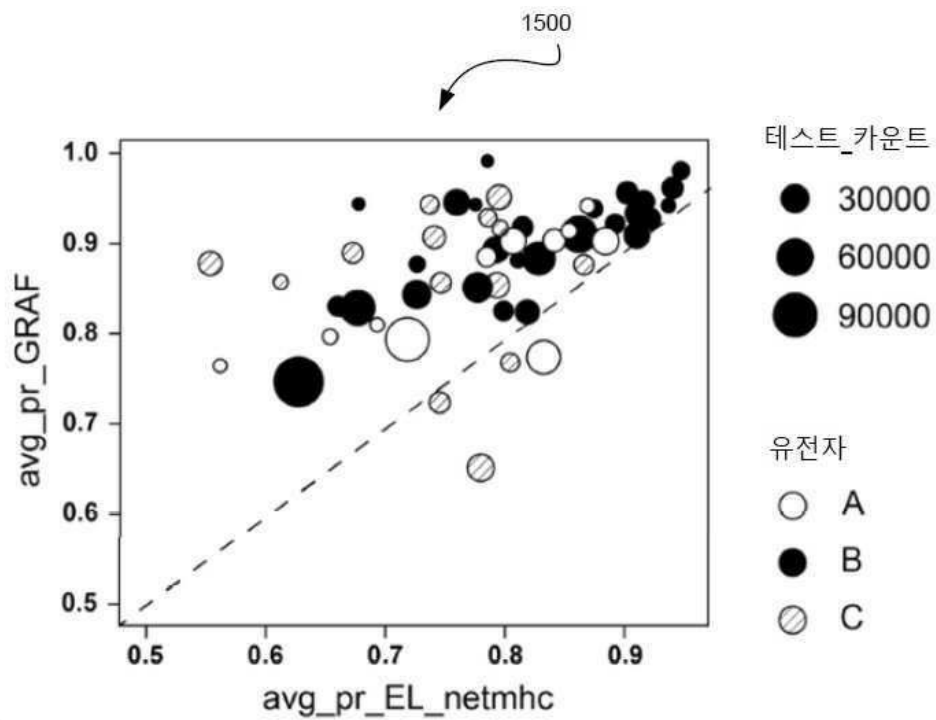
도면14b



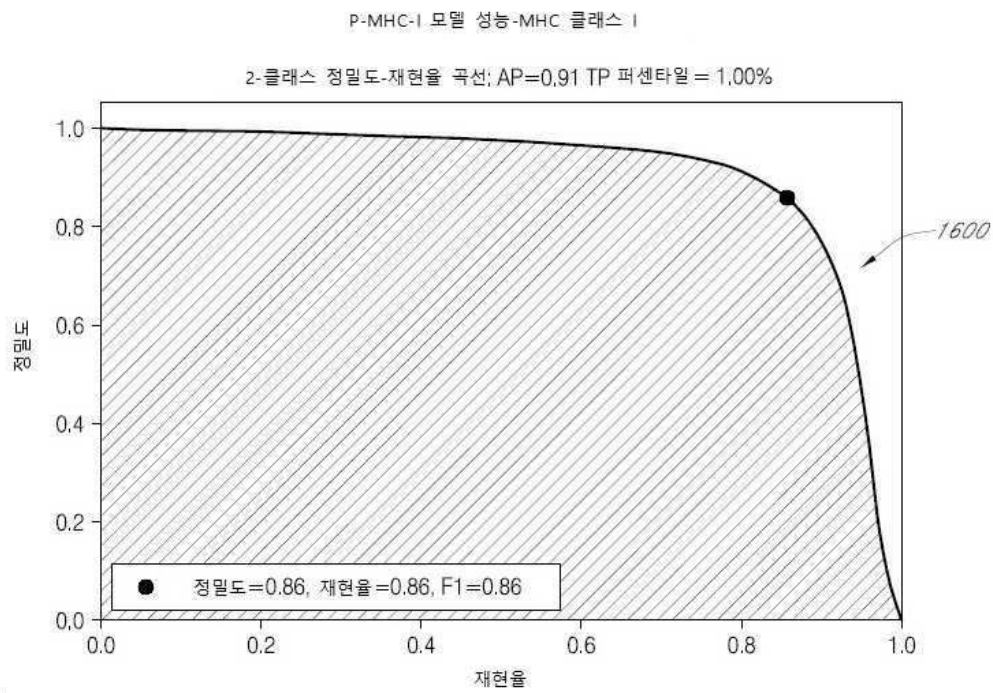
도면14c



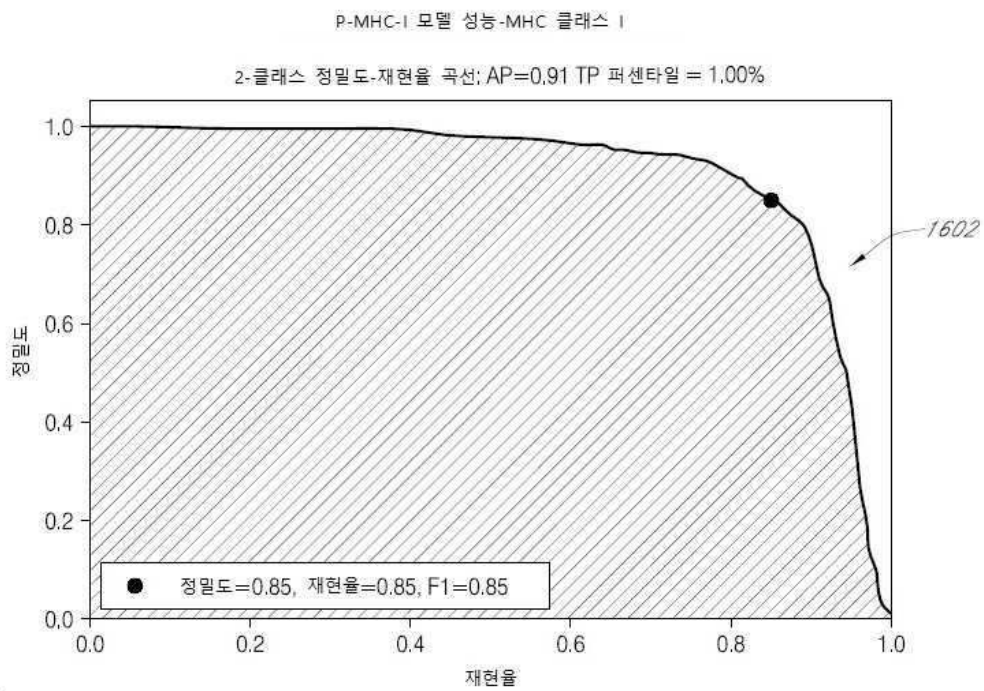
도면15



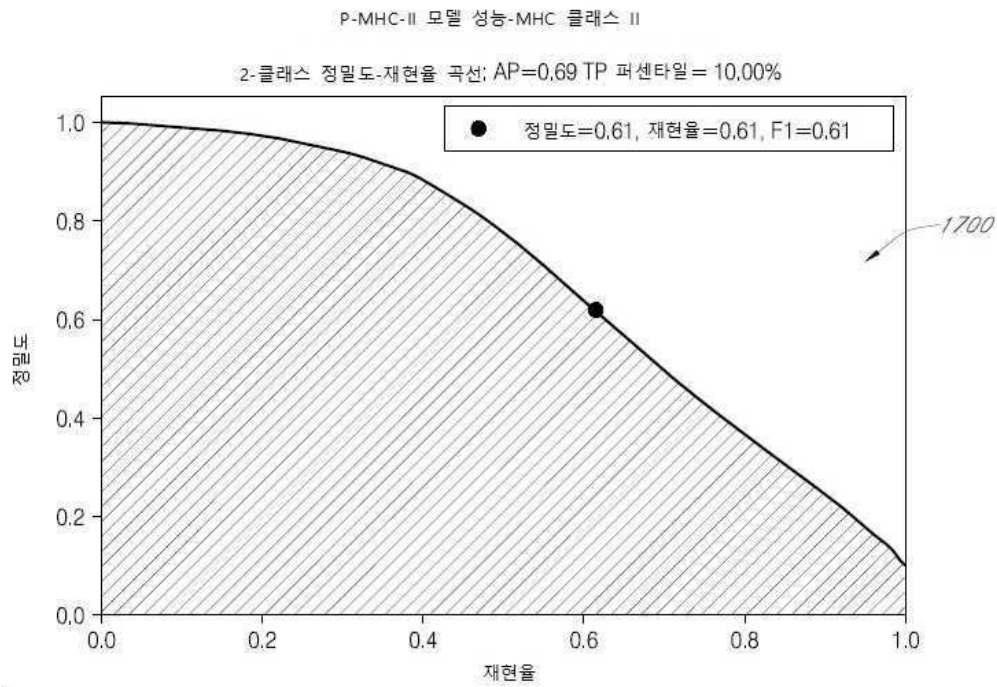
도면16a



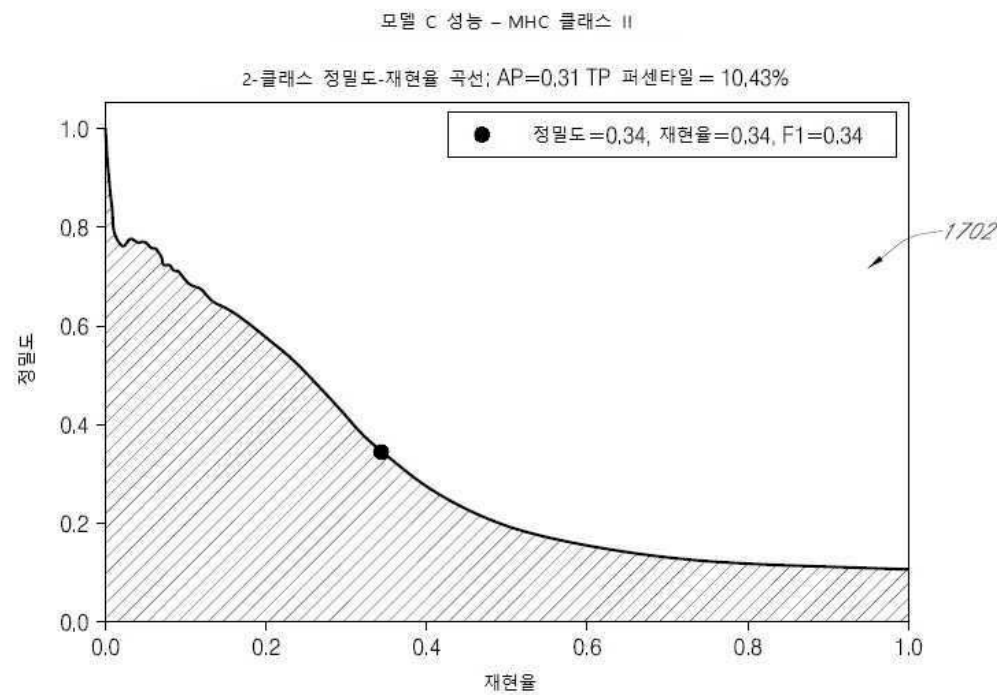
도면16b



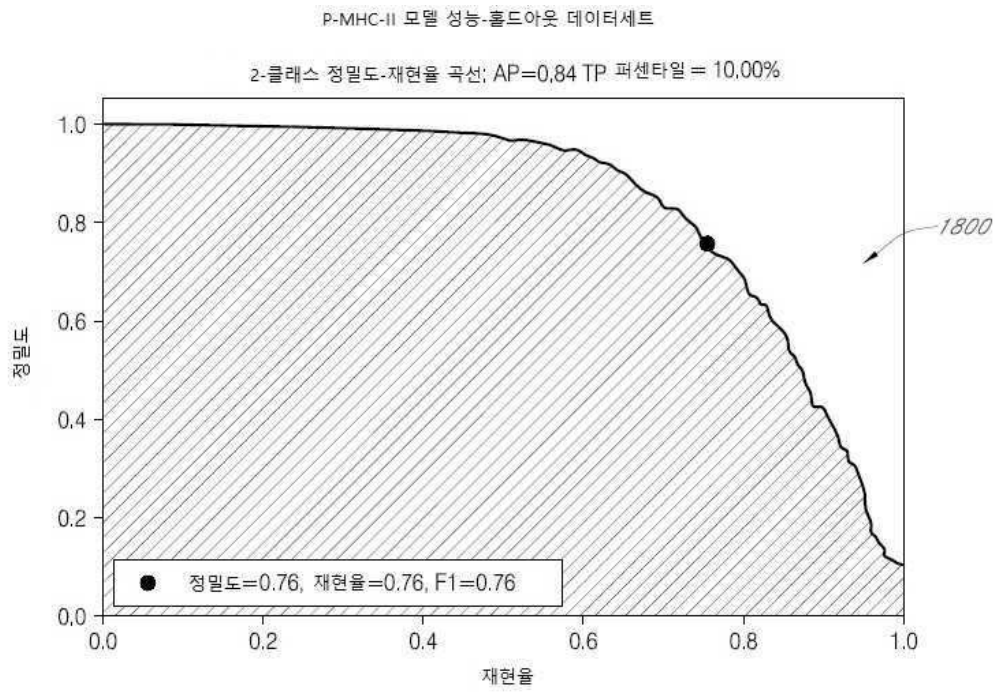
도면17a



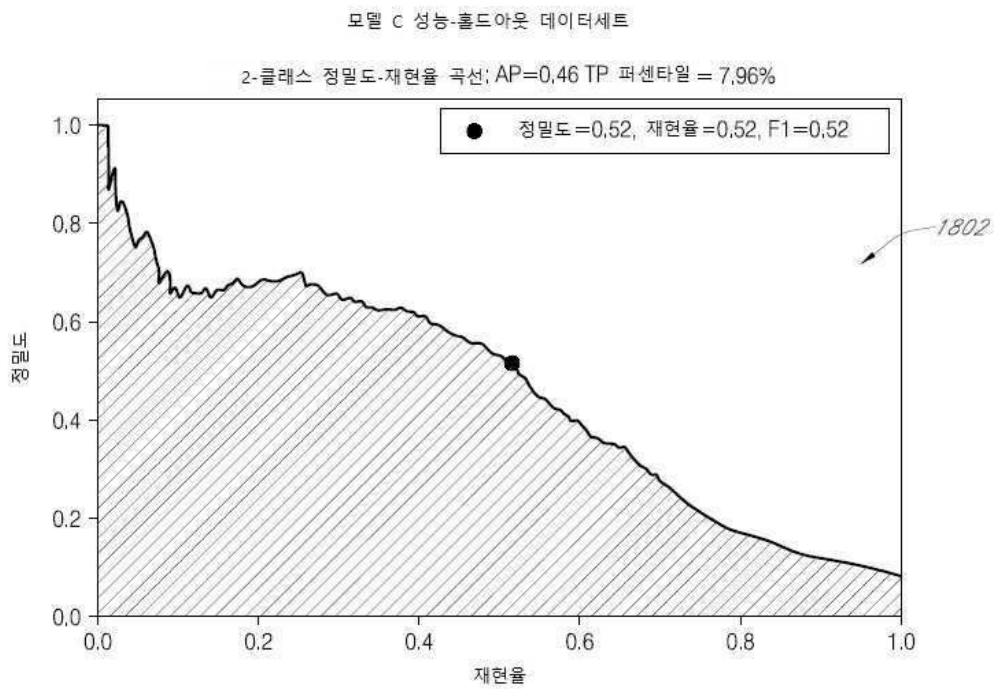
도면17b



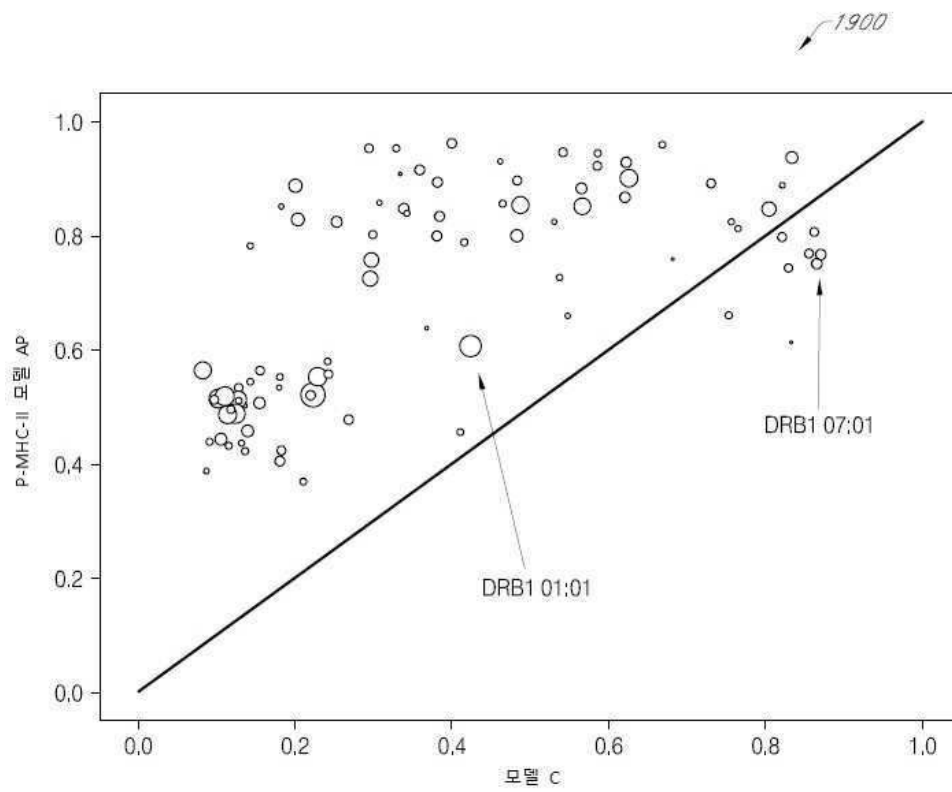
도면18a



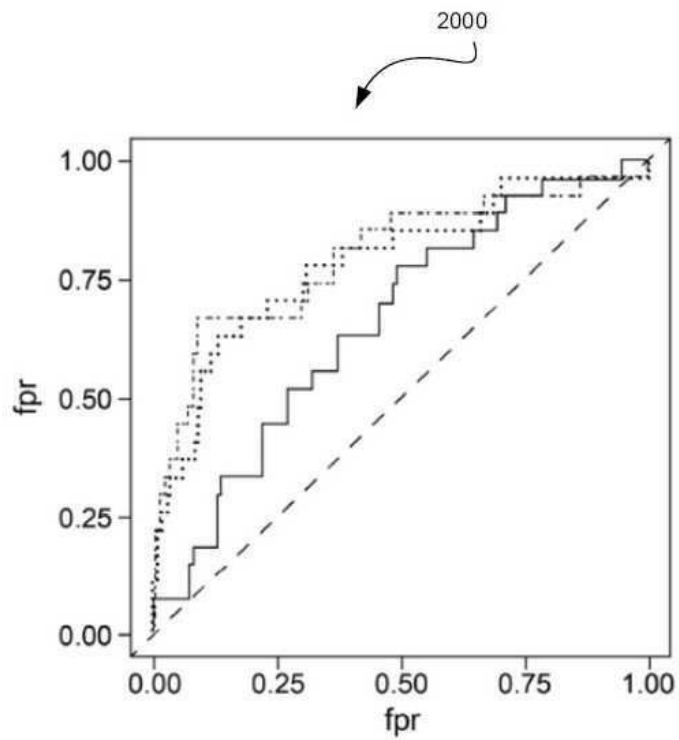
도면18b



도면19



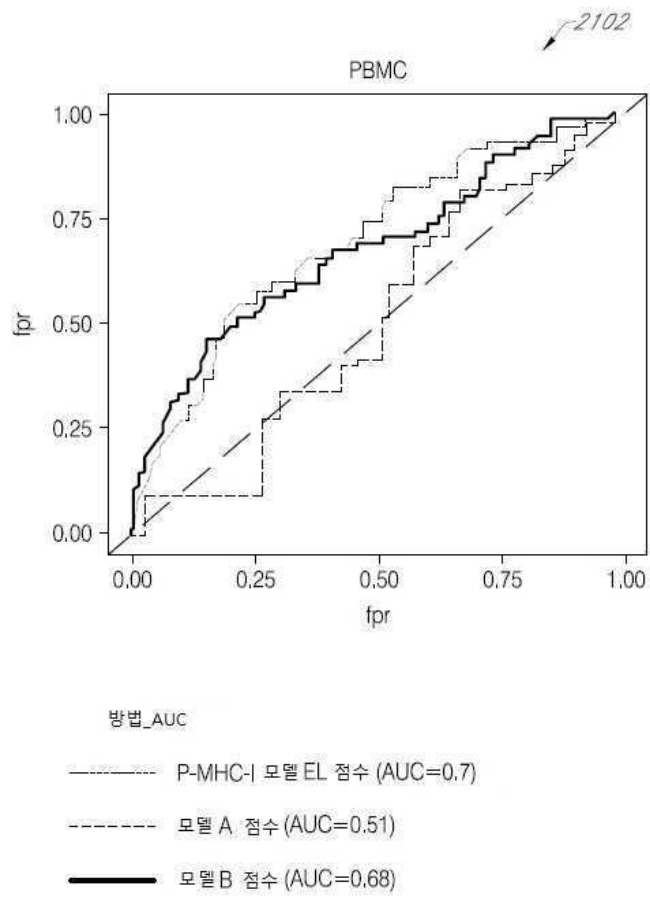
도면20



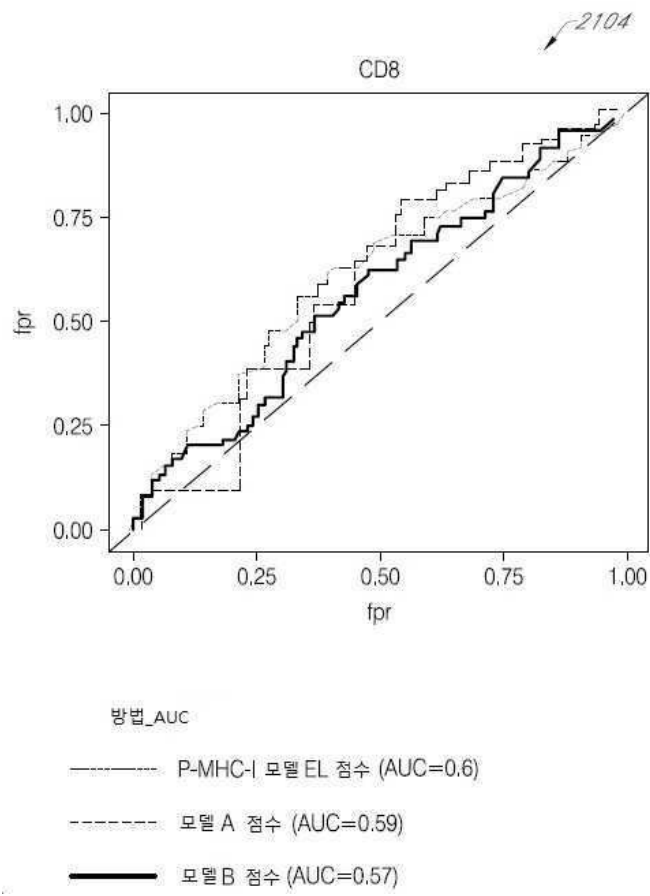
방법_AUC

- GRAF_EL_점수 (AUC=0.8)
- IEDB_BA_점수 (AUC=0.66)
- NetMHCpan_EL_점수 (AUC = 0.79)

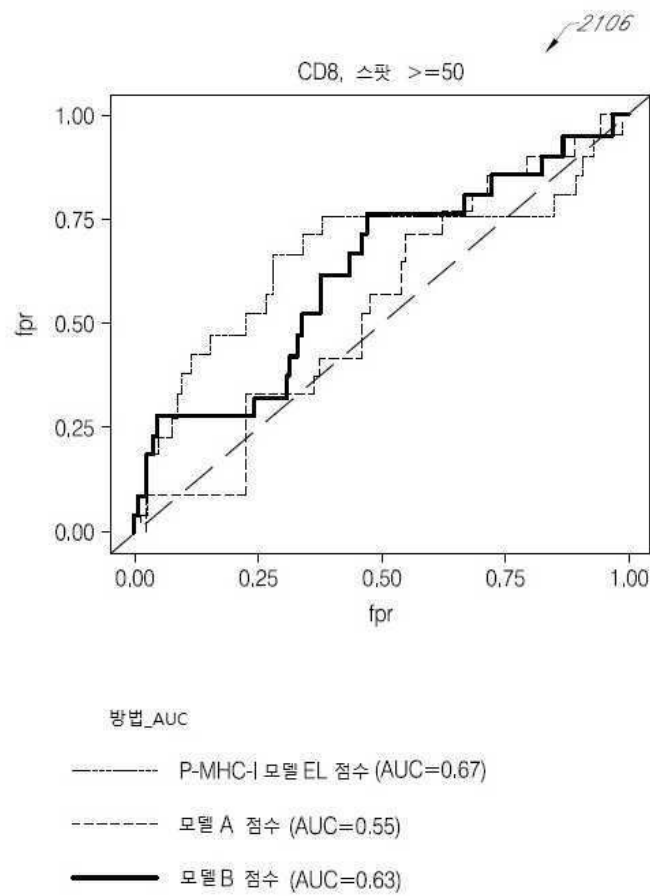
도면21a



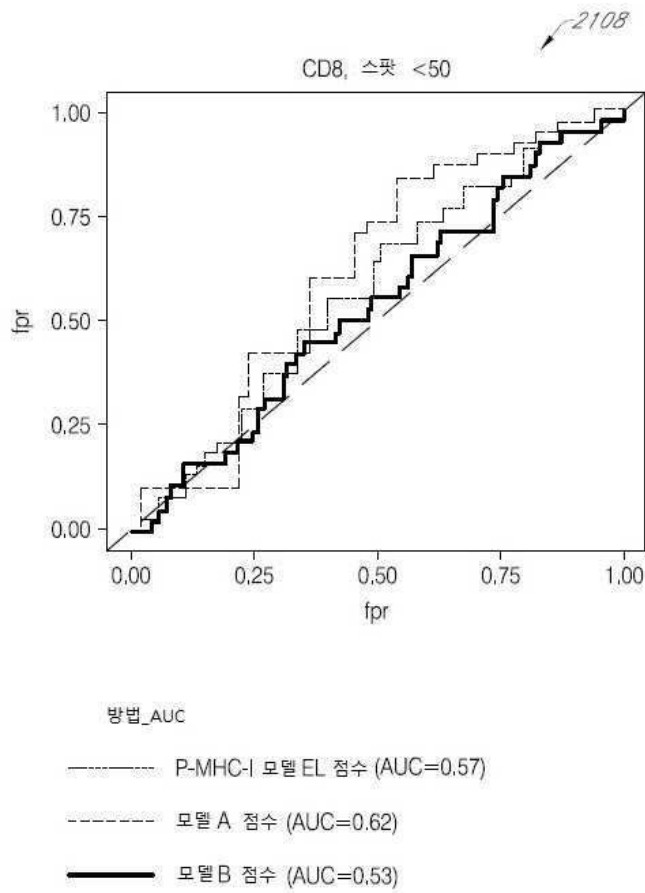
도면21b



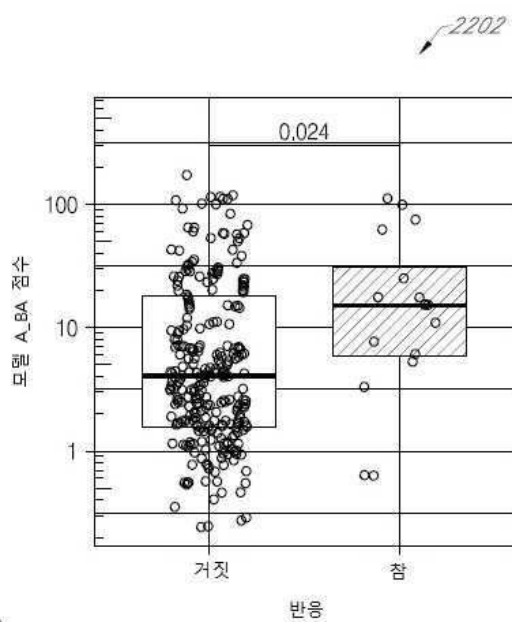
도면21c



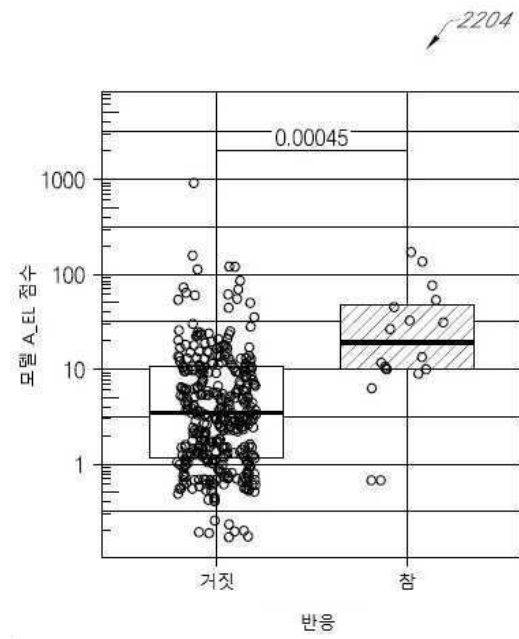
도면21d



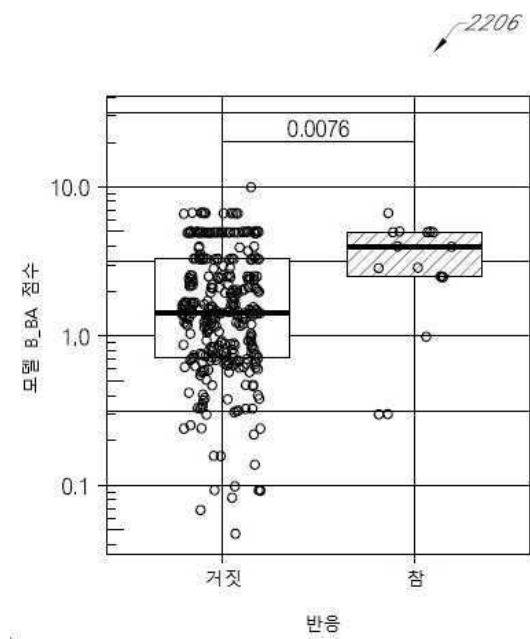
도면22a



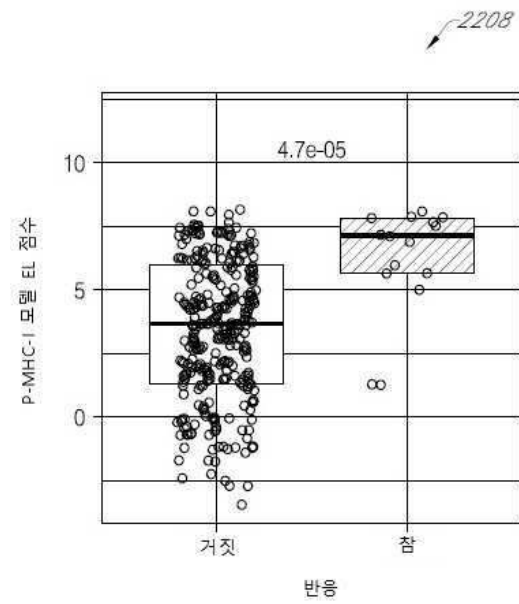
도면22b



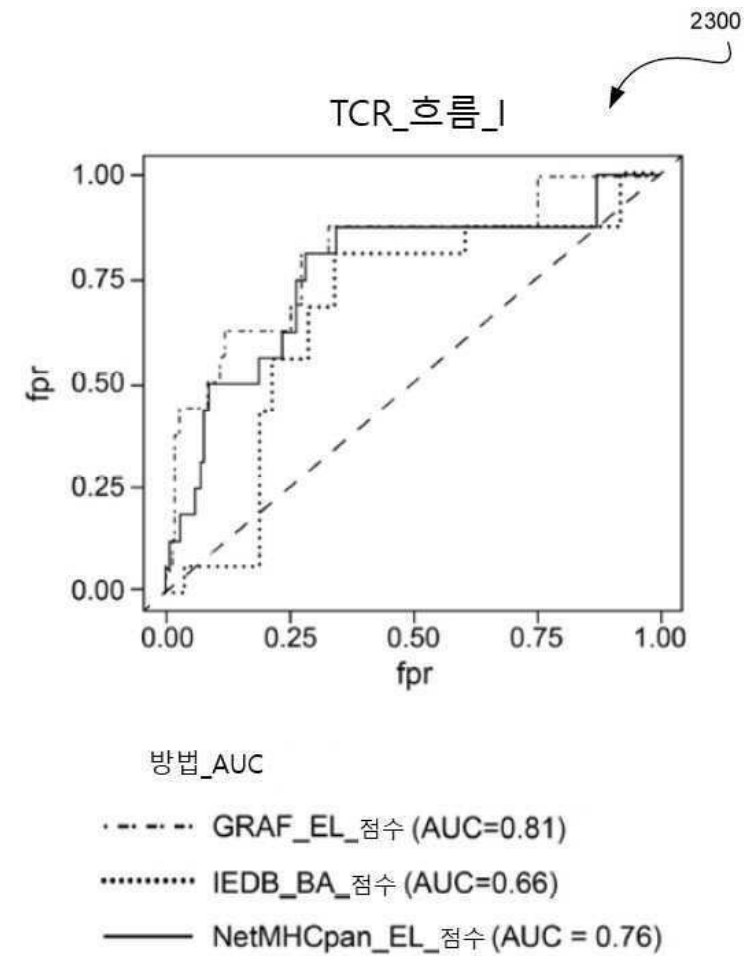
도면22c



도면22d



도면23



도면24

