



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0120945
(43) 공개일자 2024년08월08일

(51) 국제특허분류(Int. Cl.)
G16B 30/00 (2019.01) G16B 20/30 (2019.01)
G16B 40/20 (2019.01)
(52) CPC특허분류
G16B 30/00 (2019.02)
G16B 20/30 (2019.02)
(21) 출원번호 10-2023-0013563
(22) 출원일자 2023년02월01일
심사청구일자 2023년02월01일

(71) 출원인
충남대학교산학협력단
대전광역시 유성구 대학로 99 (궁동, 충남대학교)
(72) 발명자
구성우
대전광역시 유성구 온천북로33번길 22-4, 806호(봉명동, 신영 미소랑 도시형생활주택)
김진아
대전광역시 유성구 전민로 71, 112동 407호(전민동, 삼성푸른아파트)
(74) 대리인
특허법인다움

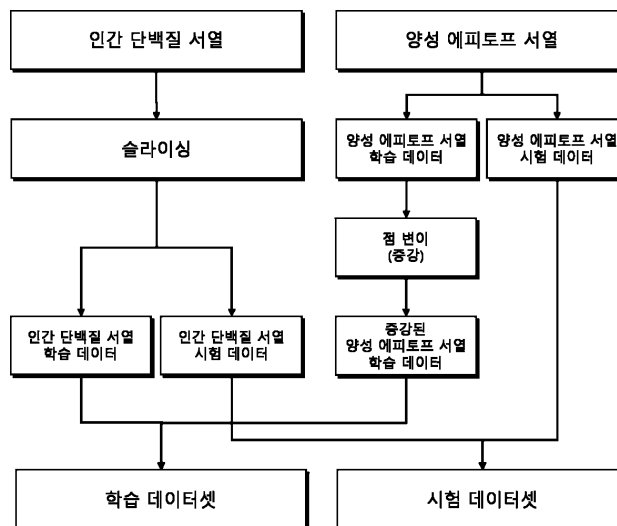
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 데이터 증강 방법 및 이를 이용한 양성 에피토프 예측 모델 생성 방법

(57) 요약

본 발명은 신약 개발을 위한 데이터 증강 방법 및 에피토프 활성 예측 방법에 관한 것이다. 본 발명은 음성 에피토프 서열 대신에 인간의 몸에 일반적으로 존재하는 인간 단백질 서열을 슬라이싱하고, 점 변이 기전으로 양성 에피토프 서열 데이터를 전처리하여, 생물화학적 근거를 기반으로 양성 에피토프 예측을 위한 데이터를 효과적으로 증강시킬 수 있다. 또한, 본 발명은 방대한 데이터로 사전 학습된 자연어 처리 모델을 이용하여 활성 에피토프에 대한 복합적인 패턴을 학습하고, 증강된 데이터로 에피토프 예측 모델을 학습시켜 임의의 아미노산 서열에 대한 에피토프 양성 여부를 예측할 수 있다. 이에 각종 질병별로 최적화된 에피토프를 효과적으로 발굴할 수 있다.

대표도 - 도2



(52) CPC특허분류

G16B 40/20 (2019.02)

(72) 발명자

윤휘열

대전광역시 유성구 죽동로 321, 104동 2103호(죽동, 금성백조예미지)

정상근

대전광역시 서구

채정우

대전광역시 유성구 봉명로 94, 701동 1302호(봉명동, 도안신도시7단지예미지백조의호수)

김동일

서울특별시 마포구 광성로 17, 103동 505호(신수동, 신촌숲아이파크)

강현석

대전광역시 유성구 와룡로136번길 97, 303동 1806호(봉산동)

이현정

대전광역시 중구 대종로 713, 1동 904호(충촌동, 금호아파트)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711152822
과제번호	2020-0-01441-003
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	인공지능융합선도프로젝트
연구과제명	인공지능융합연구센터지원(충남대학교)
기 여 율	1/4
과제수행기관명	충남대학교 산학협력단
연구기간	2020.04.01 ~ 2022.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711174038
과제번호	00155857
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	인공지능융합혁신인재양성
연구과제명	인공지능융합혁신인재양성(충남대학교)
기 여 율	1/4
과제수행기관명	충남대학교 산학협력단
연구기간	2022.07.01 ~ 2025.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711154253
과제번호	2022R1A2C1010929
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	중견연구(개인연구)
연구과제명	전주기 신약개발 통합 인공지능 시스템 개발
기 여 율	1/4
과제수행기관명	충남대학교 산학협력단
연구기간	2022.03.01 ~ 2026.02.28

이 발명을 지원한 국가연구개발사업

과제고유번호	1711174955
과제번호	2022R1A5A7085156
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	선도연구센터(CRC)지원사업
연구과제명	라이프사이클 기반 시니어헬스 융합연구센터
기 여 율	1/4
과제수행기관명	충남대학교 산학협력단
연구기간	2022.09.01 ~ 2029.02.28

명세서

청구범위

청구항 1

- (1) 인간 단백질 서열을 슬라이싱하는 단계; 및
- (2) 양성 에피토프 서열에서 어느 하나를 변환하는 단계를 포함하는, 에피토프 예측 데이터 증강 방법.

청구항 2

청구항 1에 있어서,

상기 (1) 단계는,

상기 인간 단백질 서열을 에피토프와 유사한 길이로 슬라이싱하는 것을 특징으로 하는, 에피토프 예측 데이터 증강 방법.

청구항 3

청구항 2에 있어서,

상기 (1) 단계는,

상기 인간 단백질 서열을 아미노산 12개 내지 16개를 포함하는 길이로 슬라이싱하는 것을 특징으로 하는, 에피토프 예측 데이터 증강 방법.

청구항 4

청구항 3에 있어서,

상기 (1) 단계는,

상기 인간 단백질 서열을 아미노산 12개 내지 16개 중에서 선택되는 어느 하나의 길이로 동일하게 각각의 자리별로 순차적으로 슬라이싱하는 것을 특징으로 하는, 에피토프 예측 데이터 증강 방법.

청구항 5

청구항 1에 있어서,

상기 (2) 단계는,

상기 양성 에피토프 서열에서 각각의 자리별로 하나의 아미노산을 상기 아미노산 이외의 다른 아미노산 중 어느 하나의 아미노산으로 변환하는 것을 특징으로 하는, 에피토프 예측 데이터 증강 방법.

청구항 6

청구항 5에 있어서,

상기 (2) 단계는,

FASTA 포맷에서 지정되어 있는 아미노산 목록 중에서 선택되는 다른 아미노산으로 변환하는 것을 특징으로 하는, 에피토프 예측 데이터 증강 방법.

청구항 7

(a) 인간 단백질 서열 및 양성 에피토프 서열을 이용하여 데이터를 증강하는 단계; 및

(b) 증강된 데이터를 이용하여, 아미노산 서열을 입력 받고 양성 에피토프 확률을 출력하는 양성 에피토프 예측 모델을 학습하는 단계를 포함하는 것을 특징으로 하는, 양성 에피토프 예측 모델 생성 방법.

청구항 8

청구항 7에 있어서,

상기 (b) 단계는,

(b-1) 상기 아미노산 서열을 사전 학습된 자연어 처리 모델에 입력하고, 하나 이상의 토큰을 포함하는 출력 임베딩을 출력하는 단계; 및

(b-2) 상기 출력 임베딩 중 첫 번째 토큰을 인공 신경망에 입력하고, 양성 에피토프 확률을 출력하는 단계를 포함하는 것을 특징으로 하는, 양성 에피토프 예측 모델 생성 방법.

청구항 9

청구항 8에 있어서,

상기 인공 신경망은,

완전 연결 레이어를 포함하고, 마지막 레이어의 활성화 함수는 시그모이드 함수인 것을 특징으로 하는, 양성 에피토프 예측 모델 생성 방법.

청구항 10

명령어를 로드하는 메모리; 및

상기 명령어를 실행하는 프로세서를 포함하고,

상기 명령어는,

상기 프로세서가 청구항 7 내지 청구항 9 중 어느 한 청구항의 단계를 수행하도록 하는 것을 특징으로 하는, 양성 에피토프 예측 모델 생성 장치.

발명의 설명

기술 분야

[0001] 본 발명은 신약 개발을 위한 데이터 증강 방법 및 에피토프 활성 예측 모델 생성 방법에 관한 것이다.

배경 기술

[0002] 여러 질병 중에서 암은 주요 사망 원인으로 지목되고 있다. WHO(World Health Organization)가 2020년 12월 9일에 발표한 10대 사망 원인(The top 10 causes of death)에서도 호흡기암 및 폐암(Trachea, bronchus, lung cancers)이 6위로 집계되었다. 이처럼 현대 사회에서는 암을 치료하기 위해 천문학적인 금액이 소요되고 있는 실정이다.

[0004] 기존의 항암 치료법은 메스꺼움, 구토, 피로, 및 탈모 등 수많은 부작용을 유발하는 문제점이 있다. 더욱이, 항암 치료 이후에도 암의 재발 및 전이 등이 발생할 수 있다. 이에 항원 제시 세포(Antigen Presenting Cells: APC)를 사용하여 면역 시스템에 진행되는 특이 항원에 대한 반응을 유도하는 암 치료용 백신이 주목받고 있다.

[0006] 항원-항체 반응을 이용한 신약은 암 치료용 백신 이외에도 각종 바이러스 백신이나 백혈병, 림프종, 및 골수종 등 다양한 질병 치료에도 활용될 수 있다. 이로 인해 제약 산업 분야에서는 각종 질병별로 최적화된 안티젠(항원)과 에피토프(항원결정기)를 발굴하기 위해 수많은 연구가 진행되고 있다. 그러나 활성 에피토프에 대한 데이터가 절대적으로 부족한 상황이다. 또한, 부족한 데이터로 인해 높은 예측력을 가지는 에피토프 활성 예측 모델을 생성할 수 없는 문제점이 있었다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 대한민국 공개특허공보 제10-2011-0128008호 "RNA 전이 항원 제시 세포를 이용한 신규 종양항원의 규명 방법"

발명의 내용

해결하려는 과제

[0008] 양성 에피토프 예측을 위한 데이터를 증강시키고, 증강된 데이터로 높은 예측력을 가지는 에피토프 예측 모델을 생성하는 방법을 제안한다.

과제의 해결 수단

[0009] 본 발명에 따른 에피토프 예측 데이터 증강 방법은 (1) 인간 단백질 서열을 슬라이싱하는 단계; 및 (2) 양성 에피토프 서열에서 어느 하나를 변환하는 단계를 포함한다.

[0011] 또한, 본 발명에 따른 양성 에피토프 예측 방법은 (a) 인간 단백질 서열 및 양성 에피토프 서열을 이용하여 데이터를 증강하는 단계; (b) 증강된 데이터를 이용하여 아미노산 서열을 입력 받고, 양성 에피토프 확률을 출력하는 양성 에피토프 예측 모델을 학습하는 단계를 포함한다.

발명의 효과

[0012] 본 발명은 음성 에피토프 서열 대신에 인간의 몸에 일반적으로 존재하는 인간 단백질 서열을 슬라이싱하고, 점 변이 기전으로 양성 에피토프 서열 데이터를 전처리하여, 생물화학적 근거를 기반으로 양성 에피토프 예측을 위한 데이터를 효과적으로 증강시킬 수 있다.

[0014] 또한, 본 발명은 방대한 데이터로 사전 학습된 자연어 처리 모델을 이용하여 활성 에피토프에 대한 복합적인 패턴을 학습하고, 증강된 데이터로 에피토프 예측 모델을 학습시켜 임의의 아미노산 서열에 대한 에피토프 양성 여부를 예측할 수 있다. 이에 각종 질병별로 최적화된 에피토프를 효과적으로 발굴할 수 있다.

[0016] 여기에 직접적으로 기재되지 않은 효과라도, 발명의 설명에 의해 예상되거나 기대되는 효과는 발명의 효과에 기재된 것으로 이해되어야 한다.

도면의 간단한 설명

[0017] 도 1 및 도 2는 본 발명의 실시예에 따른 데이터 증강 방법의 구성을 나타내는 도면이다.

도 3은 본 발명의 실시예에 따른 데이터 증강 방법에서 인간 단백질 서열을 슬라이싱하는 모습을 나타내는 도면이다.

도 4는 본 발명의 실시예에 따른 데이터 증강 방법에서 점 변이로 데이터를 증강하는 모습을 나타내는 도면이다.

도 5 및 도 6은 본 발명의 실시예에 따른 예측 모델 생성 방법의 구성을 나타내는 도면이다.

도 6은 본 발명의 실시예에 따른 약물-표적 상호작용 예측 방법의 학습 데이터를 나타내는 도면이다.

발명을 실시하기 위한 구체적인 내용

[0018] 본 발명은 발명의 설명에서 기재하는 실시예로 한정되지 않고 다양하게 구현될 수 있다. 발명의 설명에서 사용하는 용어는 실시예를 설명하기 위한 것으로, 본 발명을 한정하려는 의도로 사용하는 것이 아니다. 발명의 설명에서 "포함하다"의 용어는 기재된 특징이 조합된 구성이 존재함을 지정하려는 것이다. 그러므로 하나 또는 그 이상의 다른 특징이나 이들을 조합한 구성의 존재 또는 부가 가능성을 배제하지 않는 것으로 이해되어야 한다. 이하에서는 본 발명이 속한 분야의 기술자에게 명백한 내용으로서 공개된 구성이나 기능에 대한 상세한 설

명은 생략한다.

- [0020] 본 발명의 실시예에 따른 에피토프 예측 데이터 증강 방법(이하에서는 '데이터 증강 방법'이라 함)은 생물화학적 인 근거에 기반한 방법으로 데이터를 증강시켜, 양질의 학습 데이터를 확보할 수 있는 방법에 관한 것이다. 또한, 본 발명의 실시예에 따른 양성 에피토프 예측 모델 생성 방법(이하에서는 '예측 모델 생성 방법'이라 함)은 사전 학습된 자연어 처리 모델을 이용하여 아미노산의 서열 정보만으로 에피토프 활성 여부를 예측하는 모델을 생성하는 방법에 관한 것이다.
- [0022] 도 1 및 도 2는 본 발명의 실시예에 따른 데이터 증강 방법의 구성을 나타내는 도면이다. 도 1을 참조하면, 본 발명의 실시예에 따른 데이터 증강 방법은 (1) 인간 단백질 서열을 슬라이싱하는 단계(S10) 및 (2) 양성 에피토프 서열에서 어느 하나를 변환하는 단계(S20)를 포함한다. 각각의 단계는 컴퓨팅 디바이스(Computing device)에 의해 수행될 수 있으며, 각각의 단계에 대한 설명에서 주어가 생략될 수 있다. 여기서, 컴퓨팅 디바이스는 입력, 연산, 및 출력을 수행하여 데이터의 처리가 가능한 디바이스이다.
- [0024] 도 1 및 도 2를 참조하면, (1) 단계(S10)에서는 인간 단백질 서열(Human protein sequence)을 슬라이싱한다. 본 발명에서는 기존의 방법과 달리, 음성 에피토프 서열(Negative epitope sequence)을 사용하지 않고 인간 단백질 서열을 사용한다. 인간의 면역 시스템은 화이트 리스트(White list) 방식으로 작동하며, 인간의 몸에 없는 단백질을 양성 에피토프로 인식하여 공격한다. 즉, 음성 에피토프는 인간의 몸에서 정상(Normal)으로 인식되고, 양성 에피토프는 이상(Anormal)로 인식된다. 본 발명에서는 이러한 생물화학적 근거를 기반으로, 음성 에피토프 데이터 대신, 인간 몸에서 정상으로 존재하는 인간 단백질 데이터를 사용한다. 인간 단백질 데이터는 음성 에피토프 데이터보다 쉽게 획득이 가능하면서도 음성 에피토프와 유사한 성질을 가지기 때문에, 훨씬 더 많은 데이터 확보가 가능해진다.
- [0026] (1) 단계(S10)에서는 인간 단백질 서열을 에피토프와 유사한 길이로 슬라이싱한다. 인간 단백질은 에피토프에 비해 아미노산 서열이 길기 때문에, 에피토프 서열의 길이와 유사하게 슬라이싱하는 전처리 과정을 수행한다. 이러한 슬라이싱은 데이터 증강에도 기여한다.
- [0028] 도 3은 본 발명의 실시예에 따른 데이터 증강 방법에서 인간 단백질 서열을 슬라이싱하는 모습을 나타내는 도면이다. (1) 단계(S10)에서는 인간 단백질 서열을 아미노산 12개 내지 16개를 포함하는 길이로 슬라이싱한다. 도 3을 참조하면, (1) 단계(S10)에서는 인간 단백질 서열을 에피토프 길이의 5 퍼센타일 내지 95 퍼센타일인 길이로 슬라이싱한다. 이때, 에피토프 길이의 5 퍼센타일은 아미노산 12개이며, 95 퍼센타일은 아미노산 16개에 해당한다.
- [0030] (1) 단계(S10)에서는 인간 단백질 서열을 아미노산 12개 내지 16개 중에서 선택되는 어느 하나의 길이로 동일하게 각각의 자리별로 순차적으로 슬라이싱한다. 도 3의 예시에서, CAMPATH-1 안티젠을 에피토프 길이의 5 퍼센타일의 길이인 아미노산 12개로 슬라이싱하는 경우, 전체 서열에서 첫 번째 자리부터 아미노산 12개를 슬라이싱한다. 첫 번째 슬라이싱의 결과는 'MKRFLFLLLTIS' 서열이다. 그리고 두 번째 자리로 이동하여 아미노산 12개를 슬라이싱한다. 두 번째 슬라이싱의 결과는 'KRFLFLLLTISL' 서열이다. 동일한 방식으로 전체 서열에서 마지막 자리의 아미노산이 12개에 해당하는 자리까지 슬라이싱한다. 마지막 슬라이싱의 결과는 'FVANAIHLFC' 서열이다. CAMPATH-1 안티젠을 에피토프 길이의 95 퍼센타일의 길이인 아미노산 16개로 슬라이싱하는 경우도 방식은 동일하다. 또한, CAMPATH-1 안티젠을 에피토프 길이의 9 퍼센타일과 95 퍼센타일 사이의 정수 길이들로 슬라이싱할 수 있다. 슬라이싱된 인간 단백질 서열은 하나의 인간 단백질 서열로부터 다수가 생성될 수 있으며, 각각은 에피토프와 유사한 길이를 가진다. 즉, (1) 단계(S10)에서는 하나의 인간 단백질 서열로부터 음성 에피토프와 유사한 데이터를 다수개로 증강할 수 있다.

- [0032] 도 4는 본 발명의 실시예에 따른 데이터 증강 방법에서 점 변이로 데이터를 증강하는 모습을 나타내는 도면이다. 다시 도 1 및 도 2를 참조하면, (2) 단계(S20)에서는 양성 에피토프 서열에서 각각의 자리별로 하나의 아미노산을 그 아미노산 이외의 다른 아미노산 중 어느 하나의 아미노산으로 변환한다. 본 발명에서 양성 에피토프 서열의 데이터 증강 방식은 점 변이(Point mutation) 기전을 기반으로 한다. 점 변이는 인간 몸에서 매일 수조 건이 발생할 정도로 자주 발생한다. 그렇지만 대부분의 점 변이는 양성이며, 이러한 변이는 게놈(Genome)의 중요하지 않은 부분에 있다. 또한, 세포는 점 변이를 스스로 수정하며 처리한다. 따라서 점 변이로 변이된 단백질 또한 변이 전의 단백질과 유사한 성질을 가진다고 할 수 있다. 본 발명에서는 에피토프 서열 중에서 하나의 아미노산을 그 아미노산 이외의 다른 아미노산을 하나씩 넣어주어 점 변이를 인위적으로 부여한다.
- [0034] 도 4를 참조하면, (2) 단계(S20)에서는 FASTA 포맷에서 지정되어 있는 아미노산 목록 중에서 선택되는 다른 아미노산으로 변환한다. FASTA 포맷에서는 아미노산별로 대응하는 코드가 부여되어 있다. FASTA 포맷의 양성 에피토프 서열은 각각의 자리마다 어느 한 아미노산과 대응하는 코드를 가지므로, (2) 단계(S20)에서는 FASTA 포맷에서 지정되어 있는 아미노산 목록 중 그 코드 이외의 다른 코드로 변환한다. 따라서 (2) 단계(S20)를 통해 하나의 에피토프 서열의 각각의 자리별로 $n(= \text{자연계에 존재하는 아미노산 종류의 수} - 1)$ 배 이하만큼의 데이터를 얻을 수 있게 된다. (2) 단계(S20)에서는 점 변이를 인위적으로 부여하여, 양성 에피토프 서열에서 각각의 자리별로 기존의 아미노산 외의 다른 아미노산으로 변환한다. 도 4의 예시에서 양성 에피토프 서열 'EGSFDEDGFYAKVGLDAFSADELK'는 첫 번째 자리의 'E' 아미노산(Glutamic acid; 글루탐산)을 'E' 아미노산 이외의 다른 아미노산으로 변환한다. 예를 들어, 'R' 아미노산(Arginine; 아르기닌) 그리고 'H' 아미노산(Histidine; 히스티딘) 등으로 변환한다. 이러한 방식으로 (2) 단계(S20)에서는 각각의 자리별로 변환된 양성 에피토프 서열을 생성하여 데이터를 증강할 수 있다.
- [0036] 다시 도 2를 참조하면, 본 발명에서는 (1) 단계(S10)에서 슬라이싱된 인간 단백질 서열을 학습 데이터와 시험 데이터로 분류할 수 있다. 그리고 양성 에피토프 서열을 학습 데이터와 시험 데이터로 분류하고, (2) 단계(S20)에서 학습 데이터를 증강한 후, 증강된 양성 에피토프 서열 학습 데이터를 생성할 수 있다. 그리고 인간 단백질 서열 학습 데이터와 증강된 양성 에피토프 서열 학습 데이터를 이용하여 학습 데이터셋으로 구성할 수 있다. 이 학습 데이터셋은 슬라이싱과 점 변이를 통해 생물화학적인 근거에 기반하여 증강된 데이터를 포함한다.
- [0038] 도 5 및 도 6은 본 발명의 실시예에 따른 예측 모델 생성 방법의 구성을 나타내는 도면이다. 도 5를 참조하면, 본 발명의 실시예에 따른 예측 모델 생성 방법은 (a) 인간 단백질 서열 및 양성 에피토프 서열을 이용하여 데이터를 증강하는 단계(S100), (b) 증강된 데이터를 이용하여 아미노산 서열을 입력 받고, 양성 에피토프 확률을 출력하는 양성 에피토프 예측 모델을 생성하는 단계(S200)를 포함한다. 각각의 단계는 컴퓨팅 디바이스에 의해 수행될 수 있으며, 각각의 단계에 대한 설명에서 주어가 생략될 수 있다.
- [0040] (a) 단계(S100)에서는 인간 단백질 서열 및 양성 에피토프 서열을 이용하여 데이터를 증강한다. (a) 단계(S100)는 전술한 데이터 증강 방법과 구성이 동일하므로, 자세한 설명은 생략한다.
- [0042] (b) 단계(S200)에서는 증강된 데이터를 이용하여, 양성 에피토프 예측 모델을 학습한다. 양성 에피토프 예측 모델은 아미노산 서열을 입력 받고, 양성 에피토프 확률을 출력하는 분류 모델이다. 양성 에피토프 예측 모델은 사전 학습된 자연어 처리 모델과 인공 신경망 모델을 포함한다. 이때, 증강된 데이터는 인공 신경망의 파인 튜닝(Fine tuning)에 사용된다.
- [0044] 구체적으로, (b) 단계(S200)는 (b-1) 아미노산 서열을 사전 학습된 자연어 처리 모델에 입력하고, 하나 이상의 토큰을 포함하는 출력 임베딩을 출력하는 단계(S210), 및 (b-2) 출력 임베딩 중 첫 번째 토큰을 인공 신경망에

입력하고, 양성 에피토프 확률을 출력하는 단계(S220)를 포함한다.

- [0046] 도 6을 참조하면, (b-1) 단계(S210)에서는 아미노산 서열을 사전 학습된 자연어 처리 모델에 입력하고, 하나 이상의 토큰을 포함하는 출력 임베딩을 출력한다. 본 발명에서는 아미노산 서열을 임베딩하는 과정에서 사전 학습된 자연어 처리 모델을 이용한다. 여기서, 임베딩(Embedding)은 텍스트를 벡터화하는 것으로서, 컴퓨팅 디바이스가 텍스트를 효율적으로 처리할 수 있도록 하는 전처리 과정이다. 이러한 임베딩 과정을 통해 텍스트 데이터를 활용하는 인공지능 모델의 성능을 더욱 향상시킬 수 있다. 특히, 데이터가 충분히 많은 경우, 임베딩이 인공지능 모델의 성능에 큰 영향을 미치므로, 데이터의 특성에 따른 임베딩 과정이 매우 중요하다.
- [0048] (b-1) 단계(S210)에서 입력 받는 아미노산 서열은 FASTA 포맷의 서열이다. FASTA는 단백질의 구조를 표현하는 텍스트 포맷이다. 따라서 단백질을 FASTA 포맷으로 인코딩하면, 그 구조를 텍스트 타입인 문자열로 표현할 수 있다.
- [0050] 이하에서는 사전 학습된 자연어 처리 모델로서 트랜스포머(Transformer) 계열 모델을 ELECTRA 방식으로 학습시킨 모델을 사용하는 실시예를 기준으로 설명한다. 실시예에 따라서, 사전 학습된 자연어 처리 모델은 Prot-Electra 모델일 수 있다. 그러나 본 발명에서 자연어 처리 모델이 Prot-Electra 모델로 한정되는 것은 아니며, 다양한 자연어 처리 모델을 이용할 수 있다.
- [0052] (b-1) 단계(S210)에서는 아미노산 서열을 토큰화(Tokenization)한다. 임베딩을 위해 먼저 텍스트 데이터를 전 처리해야 하는데, 데이터를 토큰(Token) 단위로 나누는 작업을 토큰화라고 한다. 본 발명에서는 FASTA 포맷의 문자열 서열을 토큰화하므로, 단어 토큰화(Word Tokenization)를 수행한다. 이때, 토큰화를 수행하는 토큰나이저(Tokenizer)는 하나의 단어를 더 작은 단위의 여러 서브 워드들로 분리하는 서브 워드 토큰나이저일 수 있다.
- [0054] (b-1) 단계(S210)에서는 각각의 토큰을 자연어 처리 모델에 입력하고, 하나 이상의 토큰을 포함하는 출력 임베딩을 출력한다. (b-1) 단계(S210)는 문자열을 토큰화한 후, 사전 학습된 자연어 처리 모델을 이용하여 임베딩하는 단계이다. 사전 학습된 자연어 처리 모델은 전체 문장의 문맥을 반영하여 단어를 임베딩할 수 있다. 사전 학습된 자연어 처리 모델에 입력하는 입력 임베딩은 임베딩 레이어(Embedding layer)를 지난 임베딩 벡터들이고, 모델의 연산을 거친 후의 출력 임베딩은 전체 문장의 문맥을 반영한 임베딩 벡터들이 된다. 이때, 출력 임베딩이 포함하는 각각의 벡터를 토큰이라 할 수 있다.
- [0056] 사전 학습된 자연어 처리 모델은 사전 학습을 위해서 인공 신경망의 입력으로 들어가는 입력 텍스트의 일부 단어를 랜덤으로 마스킹(Masking)한다. 그리고 인공 신경망이 가려진 단어들(Masked words)을 예측하도록 한다. 자연어 처리 모델은 FASTA 포맷의 아미노산 서열을 입력 받고, 일부분이 삭제된 FASTA 포맷의 아미노산 서열에서 삭제된 부분에 가장 확률이 높은 아미노산을 FASTA 포맷으로 출력한다. 이를 정확하게 복구할 때까지 자연어 처리 모델의 가중치를 변경하며 학습한다.
- [0058] (b-2) 단계(S220)에서는 출력 임베딩 중 첫 번째 토큰을 인공 신경망에 입력한다. CLS 토큰은 사전 학습된 자연어 처리 모델의 출력 임베딩 중 첫 번째 토큰 임베딩에 해당한다. 본 발명에서는 출력 임베딩 중 특정 단어와 대응하는 토큰이 아닌, 문장의 첫 번째 토큰으로 삽입/출력되는 CLS 토큰을 인공 신경망의 입력으로 사용한다. 각각의 CLS 토큰은 토큰화된 아미노산의 서열 전체를 참고하여 생성되는 벡터이므로, 서열에 대한 요약 정보를 가진다. 따라서 이러한 CLS 토큰을 이용하여 양성 에피토프 확률을 예측하는데 활용할 수 있다. 즉, 전체 벡터를 사용하지 않고, 일부 토큰만을 사용하여 연산에 소요되는 리소스를 최소화하면서도 분류 모델의 성능을 대폭 향상시킬 수 있다.

[0060] 인공 신경망은 CLS 토큰을 입력 받고, 양성 에피토프 확률을 출력하는 모델이다. 인공 신경망은 심층 학습(Deep Learning) 모델일 수 있고, 완전 연결 신경망(Fully Connected Network; FCN)이거나 적어도 완전 연결 레이어(Fully Connected Layer; FCL)를 포함할 수 있다. 인공 신경망은 히든 레이어의 활성화 함수로 ReLU(Rectified Linear Unit) 함수를 사용할 수 있다. 그리고 마지막 레이어의 활성화 함수는 시그모이드(Sigmoid) 함수를 사용할 수 있다. 인공 신경망은 양성 에피토프 확률을 출력하나, 양성 에피토프 확률을 이용하여 레이블을 예측할 수 있다. 에피토프는 양성 또는 음성으로 분류되므로, 확률값이 높은 레이블로 에피토프를 분류할 수 있기 때문이다.

[0062] 실시예에 따라서, (b-2) 단계(S310)는 인공 신경망의 임의의 노드를 미리 정해진 확률로 드롭할 수 있다. 이는 드롭아웃(Dropout)을 구현한 것으로, 학습 데이터에 대한 과적합을 완화시킬 수 있다. 이러한 드롭아웃은 히든 레이어에서 이루어질 수 있다.

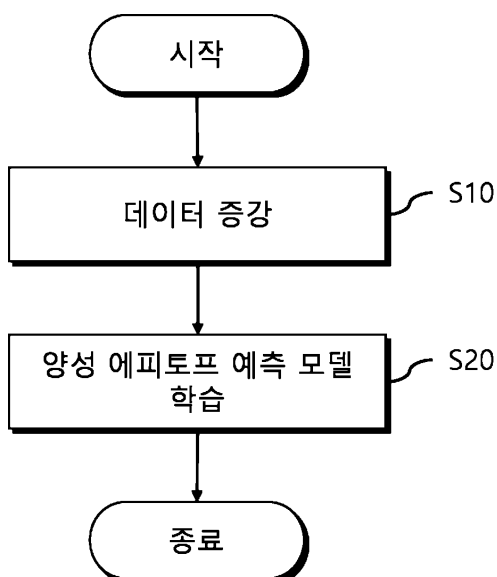
[0064] 이와 같이, 학습 데이터를 증강한 후, 사전 학습된 자연어 처리 모델을 통해 아미노산 서열을 임베딩하고, 인공 신경망을 파인 튜닝(Fine tuning)하여 양성 에피토프 확률을 출력하는 인공지능 모델의 구성은 발명자가 반복적인 시뮬레이션을 통해 도출한 구성으로서, 방대한 자연어 데이터를 기반으로 양성 에피토프 확률을 예측할 수 있도록 하는 구성에 해당한다.

[0066] 실시예에 따라서, 양성 에피토프 예측 모델 생성 장치(이하에서는 '예측 모델 생성 장치'라 함)는 명령어를 저장하는 메모리 및 명령어를 실행하는 프로세서를 포함한다. 여기서, 명령어는 본 발명의 실시예에 따른 예측 모델 생성 방법의 각각의 단계를 포함한다. 또는, 실시예에 따라서, 각각의 단계에서 수행하는 기능은 각각의 구성으로 변환되고, 이러한 구성을 포함하는 장치로 구현될 수 있다. 각각의 구성은 전술한 진단 방법의 단계에 대한 설명과 중복되므로, 그 자세한 설명은 생략한다.

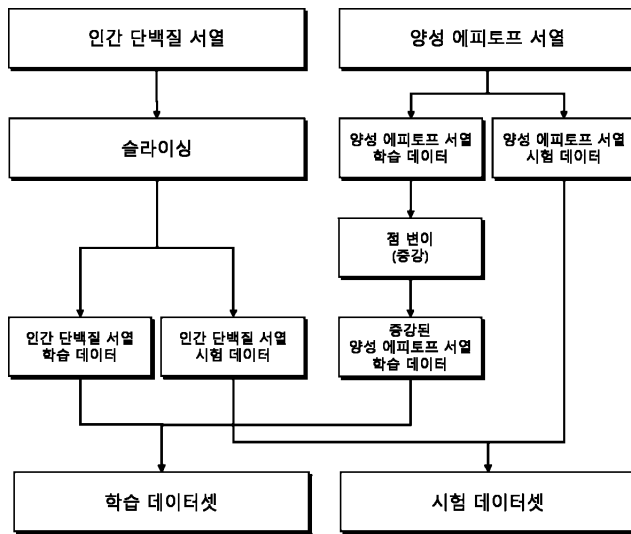
[0068] 발명의 설명에 기재된 내용은 예시에 불과하며, 본 발명은 이 분야의 기술자에 의하여 다양하게 변형되어 실시될 수 있다. 따라서 본 발명의 보호범위는 설명된 실시예의 기재와 표현으로 제한되지 않는다.

도면

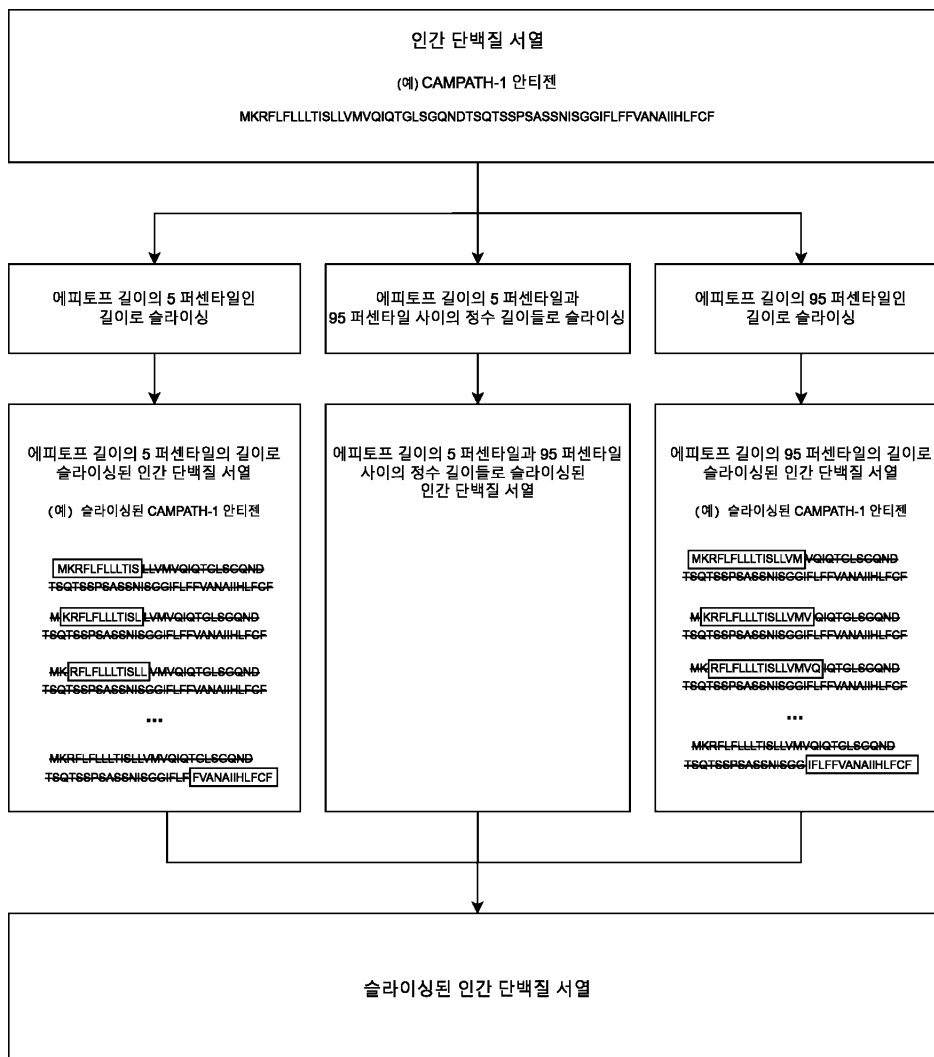
도면1



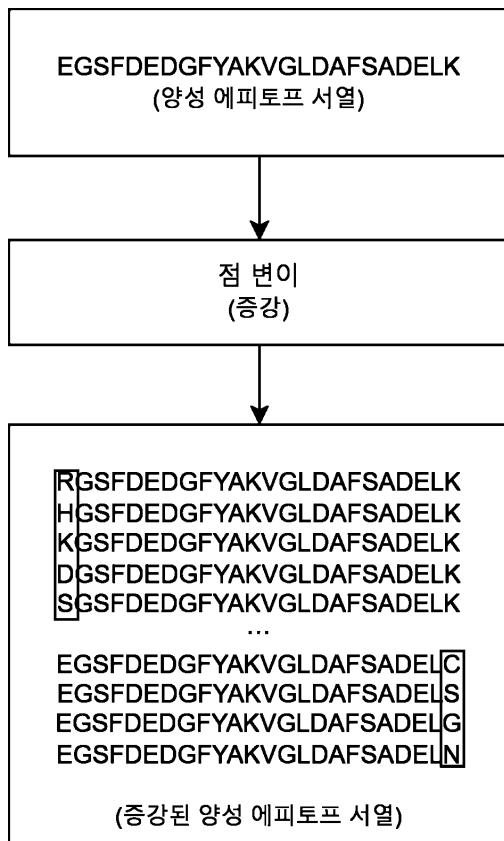
도면2



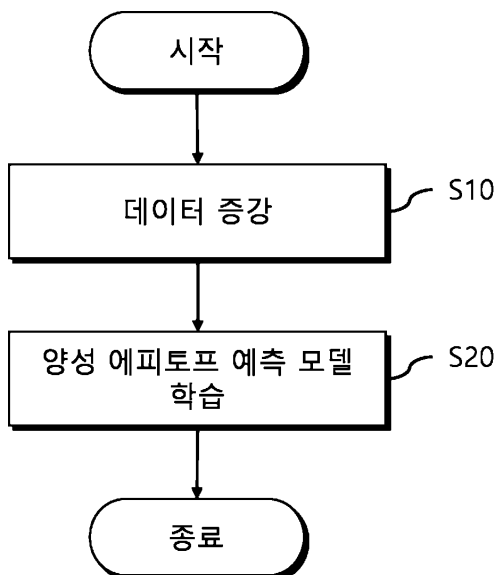
도면3



도면4



도면5



도면6

