



(19) 대한민국특허청(KR)  
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0104566  
(43) 공개일자 2024년07월05일

(51) 국제특허분류(Int. Cl.)  
G16B 40/20 (2019.01) G16B 30/00 (2019.01)  
(52) CPC특허분류  
G16B 40/20 (2019.02)  
G16B 30/00 (2019.02)  
(21) 출원번호 10-2022-0186934  
(22) 출원일자 2022년12월28일  
심사청구일자 2022년12월28일

(71) 출원인  
조선대학교산학협력단  
광주광역시 동구 조선대길 126(서석동, 산학협력단)  
(72) 발명자  
이상웅  
서울특별시 강남구 삼성로 212, 7동 908호 (대치동, 은마아파트)  
방일수  
광주광역시 동구 용산1길 2, 104동 204호  
안성윤  
경기도 성남시 수정구 복정로16번길 22  
(74) 대리인  
김전수

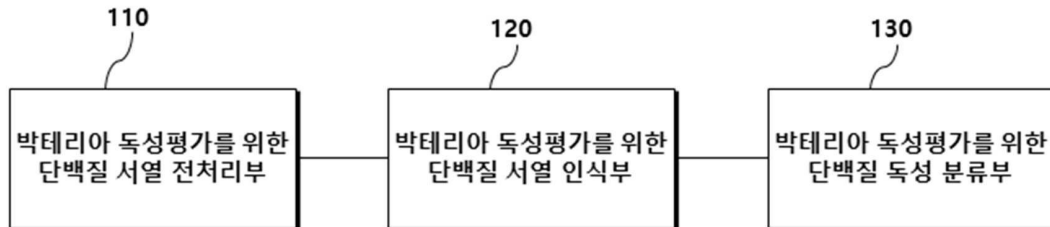
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 박테리아 독성 평가를 위한 딥러닝 장치 및 그 방법

### (57) 요약

본 발명은 박테리아 독성 평가를 위한 딥러닝 장치 및 그 방법에 관한 것으로, 더욱 상세하게는 박테리아 독성 평가를 위해 수집한 상기 박테리아가 생성한 단백질 서열을 일정한 크기로 토큰화하는 전처리부와 상기 전처리부에서 처리한 토큰의 독성 여부를 판단하기 위한 단백질 서열 인식부 및 상기 단백질 서열 인식부에서 인식한 결과를 바탕으로 상기 박테리아의 독성 여부를 판단하는 박테리아 독성 분류부를 포함하는 것을 특징으로 하는 박테리아 독성 평가를 위한 딥러닝 장치 및 그 방법에 관한 것이다.

대표도 - 도1



박테리아 독성 평가를 위한 딥러닝 장치 (100)

이 발명을 지원한 국가연구개발사업

|             |                                               |
|-------------|-----------------------------------------------|
| 과제고유번호      | 1485018633                                    |
| 과제번호        | ARQ202101648002                               |
| 부처명         | 환경부                                           |
| 과제관리(전문)기관명 | 한국환경산업기술원                                     |
| 연구사업명       | 실내공기생물학적위해인자관리기술개발사업                          |
| 연구과제명       | 실내공기 생물 유해인자에 의한 세포 및 생물체 반응 평가를 위한 다각적 모델링 기 |
| 법 개발        |                                               |
| 기 여 율       | 1/1                                           |
| 과제수행기관명     | 조선대학교 산학협력단                                   |
| 연구기간        | 2022.01.01 ~ 2022.12.31                       |

---

## 명세서

### 청구범위

#### 청구항 1

박테리아 독성 평가를 위해 수집한 상기 박테리아가 생성한 단백질 서열을 일정한 크기의 토큰으로 토큰화하는 전처리부;

상기 전처리부에서 처리한 토큰에 대해서 단백질 서열을 인식하기 위한 단백질 서열 인식부; 및

상기 단백질 서열 인식부에서 인식한 결과를 바탕으로 상기 박테리아의 독성 여부를 판단하는 박테리아 독성 분류부;를 포함하는 것을 특징으로 하는 박테리아 독성 평가를 위한 딥러닝 장치.

#### 청구항 2

청구항 1에 있어서,

상기 딥러닝 장치는,

서로 다른 유형의 단백질 서열을 포함하는 개별 데이터 세트를 두 개 이상 결합한 데이터 세트로 학습된 것이며,

상기 서로 다른 유형의 단백질 서열을 포함하는 개별 데이터 세트는, 각각 독성 인자로 분류된 박테리아 단백질로 구성된 양성 데이터 세트와 비독성 인자로 분류된 박테리아 단백질로 구성된 음성 데이터 세트를 포함하는 것을 특징으로 하는 박테리아 독성 평가를 위한 딥러닝 장치.

#### 청구항 3

청구항 1에 있어서,

상기 딥러닝 장치는,

적어도 20회 이상 사전 트레이닝을 거친 데이터 세트를 사용하며, 상기 사전 트레이닝된 결과에서 두번째로 가장 낮은 손실을 선택하여, 서로 다른 학습 속도로 미세하게 튜닝되는 것을 특징으로 하는 박테리아 독성 평가를 위한 딥러닝 장치.

#### 청구항 4

청구항 1에 있어서,

상기 박테리아 독성 분류부는,

상기 인식한 단백질 인식 결과에 대해 과적합을 줄이기 위해 드롭아웃 계층(Dropout Layer)을 추가한 것을 특징으로 하는 박테리아 독성 평가를 위한 딥러닝 장치.

#### 청구항 5

청구항 1에 있어서,

상기 박테리아 독성 분류부는,

차원을 줄여서 단백질 서열의 특성이 더욱 뚜렷하게 나타나도록 하기 위한 완전 연결 계층(Fully Connected Layer)을 포함하는 것을 특징으로 하는 박테리아 독성 평가를 위한 딥러닝 장치.

#### 청구항 6

박테리아 독성 평가를 위한 딥러닝 장치에서, 박테리아 독성 평가를 위해 상기 박테리아가 생성한 단백질 서열을 일정한 크기의 토큰으로 토큰화하는 전처리 단계;

상기 딥러닝 장치에서, 상기 토큰에 대해서 단백질 서열을 인식하기 위한 단백질 서열 인식 단계; 및

상기 딥러닝 장치에서, 상기 인식한 결과를 바탕으로 상기 박테리아의 독성 여부를 판단하는 박테리아 독성 분류 단계;를 포함하는 것을 특징으로 하는 박테리아 독성 평가를 위한 딥러닝 방법.

#### 청구항 7

청구항 6에 있어서,

상기 딥러닝 방법은,

상기 딥러닝 장치를 서로 다른 유형의 단백질 서열을 포함하는 개별 데이터 세트를 두 개 이상 결합한 데이터 세트로 학습하며,

상기 서로 다른 유형의 단백질 서열을 포함하는 개별 데이터 세트는, 각각 독성 인자로 분류된 박테리아 단백질로 구성된 양성 데이터 세트와 비독성 인자로 분류된 박테리아 단백질로 구성된 음성 데이터 세트를 포함하는 것을 특징으로 하는 박테리아 독성 평가를 위한 딥러닝 방법.

#### 청구항 8

청구항 6에 있어서,

상기 딥러닝 방법은,

적어도 20회 이상 사전 트레이닝을 거친 데이터 세트를 사용하며, 상기 사전 트레이닝된 결과에서 두번째로 가장 낮은 손실을 선택하여, 서로 다른 학습 속도로 미세하게 상기 딥러닝 장치를 튜닝하는 것을 특징으로 하는 박테리아 독성 평가를 위한 딥러닝 방법.

#### 청구항 9

청구항 6에 있어서,

상기 박테리아 독성 분류 단계는,

드롭아웃 계층(Dropout layer)을 이용하여 상기 인식한 단백질 인식 결과에 대해 과적합을 줄이는 드롭아웃 단계;를 포함하는 것을 특징으로 하는 박테리아 독성 평가를 위한 딥러닝 방법.

#### 청구항 10

청구항 6에 있어서,

상기 박테리아 독성 분류 단계는,

완전 연결 계층(Fully Connected Layer)을 이용하여 차원을 줄여서 단백질 서열의 특성이 더욱 뚜렷하게 나타나도록 하기 위한 완전 연결 단계;를 포함하는 것을 특징으로 하는 박테리아 독성 평가를 위한 딥러닝 방법.

### 발명의 설명

#### 기술 분야

[0001]

본 발명은 박테리아 독성 평가를 위한 딥러닝 장치 및 그 방법에 관한 것으로, 더욱 상세하게는 박테리아 독성 평가를 위해 수집한 상기 박테리아가 생성한 단백질 서열을 일정한 크기로 토큰화하는 전처리부와 상기 전처리부에서 처리한 토큰의 독성 여부를 판단하기 위한 단백질 서열 인식부 및 상기 단백질 서열 인식부에서 인식한 결과를 바탕으로 상기 박테리아의 독성 여부를 판단하는 박테리아 독성 분류부를 포함하는 것을 특징으로 하는 박테리아 독성 평가를 위한 딥러닝 장치 및 그 방법에 관한 것이다.

#### 배경 기술

[0002]

사람은 집에서는 직장에서는 90% 이상의 시간을 실내에서 보낸다. 실내 공간에 들어가면 열린 문을 통해 또는 질병을 옮길 가능성이 있는 숙주로부터 내부로 이동했을 수 있는 병원균으로 공기가 즉시 오염될 수 있다. 오염된 공기는 공기 전파를 통한 감염 가능성을 크게 높이는 경향이 있다. 결핵, 천식 및 COVID-19과 같은 널리 퍼진 질병은 때때로 숙주의 죽음으로 이어지기 때문에 심각한 건강의 위험성을 초래할 수 있다. 세계보건기구(WHO)는 오염된 실내 공기로 인해 매년 평균 380만명이 목숨을 잃는다고 발표했다.

- [0003] 박테리아와 곰팡이는 가장 흔한 공기 중 병원균이다. 이 병원균은 다양한 단백질을 생산한다. 단백질은 많은 유용한 기능을 수행하지만 때때로 인간에게 해를 끼치기도 한다. 이러한 단백질을 독성단백질 또는 독소로 알려져 있으며, 인체에 독성 인자로 작용하여 질병을 유발하게 된다.
- [0004] 독성 단백질을 분석하는 전통적인 방법은 시험관내에서 직접 실험하고 배양한 후 분석하는 방법이 활용되고 있다. 이는 정확한 분석 결과를 도출할 수 있지만, 시간이 오래 걸린다는 단점이 있다. 근래에는 게놈 및 아미노산 서열 데이터를 딥러닝 기반으로 한 시뮬레이션 분석법이 많이 활용되고 있다. 시뮬레이션 분석법은 상대적으로 정확도가 떨어지나 빠른 시간 내에 결과를 도출할 수 있는 특징이 있다.
- [0005] 본 발명에서는 독성 요인으로 작용할 수 있는 박테리아 단백질을 예측하기 위해 딥러닝 기반의 독성평가 장치 및 방법을 제안하고자 한다.
- [0006] 다음으로 본 발명의 기술분야에 존재하는 선행기술에 대하여 간단하게 설명하고, 이어서 본 발명이 상기 선행기술에 비해서 차별적으로 이루고자 하는 기술적 사항에 대해서 기술하고자 한다.
- [0007] 먼저 한국공개특허 제1020200126715호(2020.11.09.)는 단백질의 독성 여부를 판단하는 점에서 본 발명의 목적과 일부 유사하지만, 단일 단백질의 독성 여부를 판단하는 것을 그 특징으로 한다. 반면에 본 발명은 박테리아의 복수의 단백질 서열에서 잠재적인 독성을 가지는 단백질 서열을 분류하는 것으로, 한국공개특허 제1020200126715호는 다수의 단백질 서열에 대한 잠재적인 독성을 분석하는 본 발명의 기술적 특징을 전혀 기재하거나 시사 혹은 그 어떠한 암시도 없다.
- [0008] 한국등록특허 제2030055호(2019.10.01.)는 바이러스의 단백질 서열에서 특정 단백질 서열을 추출하는 점에서 본 발명의 목적이 일부 유사하다. 다만, 한국등록특허 제2030055호는 단백질 서열을 비교를 통해 공통적으로 보유하고 있는 단백질 서열을 추출하는 것으로 본 발명과 같이 박테리아의 각 단백질 서열에 대한 독성 여부를 분류하는 기술적 특징에 대해서는 전혀 기재하거나 시사 혹은 그 어떠한 암시도 없다.
- [0009] 한국등록특허 제2299220호(2022.09.01.)는 학습을 통해 단백질의 특정 기능을 예측하는 점에서 본 발명의 목적과 일부 유사하다. 다만, 한국등록특허 제2299220호는 단백질과 화합물 간의 상호 작용을 예측하는 것으로, 본 발명과 같이 박테리아의 단백질 서열에서 독성을 가지는 단백질 서열을 분류함으로써 잠재적인 독성을 가지는 박테리아를 선별할 수 있도록 하는 기술적 특징에 대해서는 전혀 기재하거나 시사 혹은 그 어떠한 암시도 없다.
- [0010] 학술논문 Deep learning using primary structure and domain embeddings for assessing protein toxicity.(Bioinformatics 2021, 36, 5159-5168., Pan, X.; Zuallaert, J.; Wang, X.; Shen, H.B.; Campos, E.P.; Marushchak, D.O.; De Neve, W. ToxDL:)은 특정 학습네트워크를 이용하여 학습을 수행함으로써 단백질의 독성을 예측하는 점에서 본 발명의 목적과 일부가 유사하나, 상기 논문은 일반적인 CNN을 이용하는 반면에, 본 발명은 과적합을 방지하기 위한 드롭아웃 계층, 완전 연결계층 및 독성 여부를 분류하기 위한 하이퍼볼릭 탄젠트 함수를 적용한 학습 네트워크를 이용하여 독성 여부가 레이블링된 박테리아의 각 단백질 서열을 학습하는 기술적 구성에 대해서는 전혀 기재하고 있지 않다. 따라서 양 발명은 그 기술적 구성, 목적 및 효과에 있어서 현저한 차이점이 있다.

## 발명의 내용

### 해결하려는 과제

- [0011] 본 발명은 상기와 같은 문제점을 해결하기 위해 창작된 것으로서, 박테리아 독성 평가를 위한 고성능의 딥러닝 장치 및 그 방법을 제공하는 것을 그 목적으로 한다.
- [0012] 또한 본 발명은 서로 다른 유형의 단백질 서열을 포함하는 개별 데이터 세트를 두 개 이상 결합한 데이터 세트로 학습하여 실제 박테리아에서 발생하는 단백질의 독성여부를 고성능으로 판단하는 것을 그 목적으로 한다.
- [0013] 또한 본 발명은 적어도 20회 이상 사전 트레이닝을 거친 데이터 세트를 사용하며, 상기 사전 트레이닝된 결과에서 두번째로 가장 낮은 손실(second lowest loss)을 선택하여, 서로 다른 학습 속도로 미세하게 튜닝함으로써(fine-grained tuning), 성능을 향상시키는 것을 그 목적으로 한다.
- [0014] 또한 본 발명은 단백질 인식 결과에 대해 과적합을 줄이기 위해 드롭아웃 계층을 추가하고, 차원을 줄여서 단백질 서열의 특성이 더욱 뚜렷하게 나타나도록 하기 위한 완전 연결 계층을 포함함으로써, 성능을 높이는 것을 목적으로 한다.

## 과제의 해결 수단

- [0015] 본 발명의 일 실시예에 따른 박테리아 독성 평가를 위한 딥러닝 장치는 박테리아 독성 평가를 위해 수집한 상기 박테리아가 생성한 단백질 서열을 일정한 크기의 토큰으로 토큰화하는 전처리부, 상기 전처리부에서 처리한 토큰에 대해서 단백질 서열을 인식하기 위한 단백질 서열 인식부, 및 상기 단백질 서열 인식부에서 인식한 결과를 바탕으로 상기 박테리아의 독성 여부를 판단하는 박테리아 독성 분류부를 포함하는 것을 특징으로 한다.
- [0016] 또한 상기 딥러닝 장치는, 서로 다른 유형의 단백질 서열을 포함하는 개별 데이터 세트를 두 개 이상 결합한 데이터 세트로 학습된 것이며, 상기 서로 다른 유형의 단백질 서열을 포함하는 개별 데이터 세트는, 각각 독성 인자로 분류된 박테리아 단백질로 구성된 양성 데이터 세트와 비독성 인자로 분류된 박테리아 단백질로 구성된 음성 데이터 세트를 포함하는 것을 특징으로 한다.
- [0017] 또한 상기 딥러닝 장치는, 적어도 20회 이상 사전 트레이닝을 거친 데이터 세트를 사용하며, 상기 사전 트레이닝된 결과에서 두번째로 가장 낮은 손실(second lowest loss)을 선택하여, 서로 다른 학습 속도로 미세하게 튜닝되는 것을 특징으로 한다.
- [0018] 여기서 상기 박테리아 독성 분류부는, 상기 인식한 단백질 인식 결과에 대해 과적합을 줄이기 위해 드롭아웃 계층(Dropout Layer)을 추가한 것을 특징으로 한다.
- [0019] 또한 상기 박테리아 독성 분류부는, 차원을 줄여서 단백질 서열의 특성이 더욱 뚜렷하게 나타나도록 하기 위한 완전 연결 계층(Fully Connected Layer)을 포함하는 것을 특징으로 한다.
- [0020] 한편, 본 발명의 또 다른 일 실시예에 따른 박테리아 독성 평가를 위한 딥러닝 방법은, 박테리아 독성 평가를 위한 딥러닝 장치에서, 박테리아 독성 평가를 위해 상기 박테리아가 생성한 단백질 서열을 일정한 크기의 토큰으로 토큰화하는 전처리 단계; 상기 딥러닝 장치에서, 상기 토큰에 대해서 단백질 서열을 인식하기 위한 단백질 서열 인식 단계; 및 상기 딥러닝 장치에서, 상기 인식한 결과를 바탕으로 상기 박테리아의 독성 여부를 판단하는 박테리아 독성 분류 단계를 포함하는 것을 특징으로 한다.
- [0021] 또한 상기 딥러닝 방법은, 상기 딥러닝 장치를 서로 다른 유형의 단백질 서열을 포함하는 개별 데이터 세트를 두 개 이상 결합한 데이터 세트로 학습하며, 상기 서로 다른 유형의 단백질 서열을 포함하는 개별 데이터 세트는, 각각 독성 인자로 분류된 박테리아 단백질로 구성된 양성 데이터 세트와 비독성 인자로 분류된 박테리아 단백질로 구성된 음성 데이터 세트를 포함하는 것을 특징으로 한다.
- [0022] 또한 상기 딥러닝 방법은, 적어도 20회 이상 사전 트레이닝을 거친 데이터 세트를 사용하며, 상기 사전 트레이닝된 결과에서 두번째로 가장 낮은 손실(second lowest loss)을 선택하여, 서로 다른 학습 속도로 미세하게 상기 딥러닝 장치를 튜닝하는 것을 특징으로 한다.
- [0023] 상기 박테리아 독성 분류 단계는, 드롭아웃 계층(Dropout layer)을 이용하여 상기 인식한 단백질 인식 결과에 대해 과적합을 줄이는 드롭아웃 단계를 포함하는 것을 특징으로 한다.
- [0024] 상기 박테리아 독성 분류 단계는, 완전 연결 계층(Fully Connected Layer)을 이용하여 차원을 줄여서 단백질 서열의 특성이 더욱 뚜렷하게 나타나도록 하기 위한 완전 연결 단계를 포함하는 것을 특징으로 한다.

## 발명의 효과

- [0025] 이상에서와 같이 본 발명의 박테리아 독성평가를 위한 딥러닝 장치 및 그 방법은 서로 다른 유형의 단백질 서열을 포함하는 개별 데이터 세트를 두 개 이상 결합한 데이터 세트로 학습하여 실제 박테리아에서 발생하는 단백질의 독성여부를 고성능으로 판단할 수 있도록 하는 효과가 있다.
- [0026] 또한 본 발명은 적어도 20회 이상 사전 트레이닝을 거친 데이터 세트를 사용하며, 상기 사전 트레이닝된 결과에서 두번째로 가장 낮은 손실(second lowest loss)을 선택하여, 서로 다른 학습 속도로 미세하게 튜닝함으로써(fine-grained tuning), 성능을 향상시키는 효과가 있다.
- [0027] 또한 본 발명은 단백질 인식 결과에 대해 과적합을 줄이기 위해 드롭아웃 계층을 추가하고, 차원을 줄여서 단백질 서열의 특성이 더욱 뚜렷하게 나타나도록 하기 위한 완전 연결 계층을 포함함으로써, 성능을 높이는 효과가 있다.

## 도면의 간단한 설명

- [0028] 도 1은 본 발명의 일 실시예에 따른 박테리아 독성 평가를 위한 딥러닝 장치 및 그 방법을 설명하기 위해 나타낸 도면이다.
- 도 2는 본 발명의 일 실시예에 따른 학습 데이터 세트를 생성하는 과정을 설명하기 위해 나타낸 도면이다.
- 도 3은 본 발명의 일 실시예에 따른 알려지지 않은 단백질 서열에 대한 검증 데이터 세트 구성을 설명하기 위해 나타낸 도면이다.
- 도 4는 본 발명의 일 실시예에 따른 알려지지 않은 단백질 서열에 검증 데이터 세트를 구성 방법을 설명하기 위해 나타낸 도면이다.
- 도 5은 본 발명의 일 실시예에 따른 독성 동물성 단백질 평가 성능을 설명하기 위한 도면이다.
- 도 6은 본 발명의 일 실시예에 따른 독성 박테리아 단백질 평가 성능을 설명하기 위한 도면이다.
- 도 7은 본 발명의 일 실시예에 따른 서로 다른 유형의 단백질 서열을 포함하는 개별 데이터 세트를 두 개 이상 결합한 데이터 세트로 학습한 독성 박테리아 단백질 평가 성능을 설명하기 위해 나타낸 도면이다.
- 도 8은 본 발명의 일 실시예에 따른 서로 다른 유형의 단백질 서열을 포함하는 개별 데이터 세트를 두 개 이상 결합한 데이터 세트로 학습한 모델과 독성 박테리아 단백질 평가를 위한 학습 데이터 세트로 학습한 독성 박테리아 단백질 평가 성능을 비교 설명하기 위해 나타낸 도면이다.
- 도 9는 본 발명의 일 실시예에 따른 실제 독성 박테리아 단백질 평가 성능을 상세하게 설명하기 위해 모델의 혼동행렬을 나타낸 도면이다.
- 도 10은 본 발명의 일 실시예에 따른 실제 독성 박테리아 단백질 평가에서 위험성이 있는 단백질로 평가한 단백질 서열 수를 나타낸 도면이다.
- 도 11은 본 발명의 일 실시예에 따른 박테리아 독성 평가를 위한 딥러닝 과정을 나타낸 흐름도이다.
- 도 12는 본 발명의 일 실시예에 따른 박테리아 독성 평가를 위한 딥러닝 장치의 블록도를 나타낸 흐름도이다.

### 발명을 실시하기 위한 구체적인 내용

- [0029] 이하, 첨부한 도면을 참조하여 본 발명의 박테리아 독성 평가를 위한 딥러닝 장치 및 그 방법에 대한 바람직한 실시예를 상세히 설명한다. 각 도면에 제시된 동일한 참조부호는 동일한 부재를 나타낸다. 또한 본 발명의 실시예들에 대해서 특정한 구조적 내지 기능적 설명들은 단지 본 발명에 따른 실시예를 설명하기 위한 목적으로 예시된 것으로, 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는 것이 바람직하다.
- [0030] 도 1은 본 발명의 일 실시예에 따른 박테리아 독성 평가를 위한 딥러닝 장치 및 그 방법을 설명하기 위해 나타낸 도면이다.
- [0031] 도 1에 도시한 바와 같이, 본 발명은 박테리아 독성평가를 위한 단백질 서열 전처리부(이하, 전처리부)(110), 박테리아 독성평가를 위한 단백질 서열 인식부(120) 및 박테리아 독성평가를 위한 박테리아 독성 분류부(130)를 포함하여 구성된다.
- [0032] 상기 단백질 서열 인식부(120)는 Hugging Face Mani의 ProtBERT 모델을 기반으로 구성한다. ProtBERT 모델은 입력 데이터가 토큰화 되어야 하므로, 상기 전처리부(110)는 단백질 입력 데이터를 토큰화 한다. 토큰화된 입력 데이터는 단백질 서열 인식부를 통해 단백질 서열이 인식되며, 인식된 단백질 서열은 독성 결정을 위해 박테리아 독성 분류부(130)로 입력된다. 상기 박테리아 독성 분류부(130)는 데이터의 과적합을 줄이기 위해 드롭아웃 계층(Dropout Layer)을 포함하며, 차원을 줄이기 위해 완전 연결 계층(Fully Connected Layer)을 포함한다. 차원을 줄이면 단백질 서열의 특성이 더욱 뚜렷하게 나타나게 하는 효과가 있다.
- [0033] 상기 박테리아 독성 분류부(130)는 독성 분류를 위해 하이퍼볼릭탄젠트(tanh) 함수를 사용하며 tanh함수는 [수학식 1]과 같다

[0034] [수학식 1]

$$\tanh(x) = \frac{2}{1 + e^{-2x}} - 1$$

[0035]

[0036] [수학식 1]에서 x는 이전 레이어에서 생성된 특징 벡터이고 tanh의 예측 결과 범위는 -1에서 1사이로 표시된다. 양수는 독성 단백질을 나타내고, 음수는 독성이 없음을 나타낸다.

[0037] 본 발명에서는 적어도 20회 이상 사전 트레이닝을 거친 데이터 세트를 사용하며, 상기 사전 트레이닝된 결과에서 두번째로 가장 낮은 손실(second lowest loss)을 선택하여, 서로 다른 학습 속도로 미세하게 상기 딥러닝 장치를 튜닝(fine-grained tuning)한다.

[0038] 도 2는 본 발명의 일 실시예에 따른 학습 데이터 세트를 생성하는 과정을 설명하기 위해 나타낸 도면이다.

[0039] 본 발명의 성능을 검증하기 위해 기존의 다른 독성 단백질 예측모델에서 사용한 데이터 세트 2개에 대한 설명과 본 발명에서 제안하는 데이터 세트를 구성하는 방법을 설명한다.

[0040] 첫 번째 데이터 세트는 알려진 동물 독성 단백질로 구성된 데이터 세트로 UniPort의 Animal Toxion Annotation Project에서 4,472개의 양성 샘플과 6,341개의 음성 샘플로 구성되어 있다. 첫 번째 데이터 세트는 ToxDL로 표기한다. ToxDL 데이터 세트에서 1002개 이상의 아미노산 잔기로 구성된 단백질을 포함하는데, 본 발명의 단백질 서열 인식부(110)에서는 1002개 이상의 서열 길이를 처리할 수 없다. 이를 해결하기 위해서는 1002개 이상의 서열을 자르는 방법과 1002개 이하의 데이터만 사용하는 방법이 있는데, 데이터 서열을 임의로 1002개에서 자를 경우 잘못된 학습이 될 가능성이 있어, 1002개 이하의 서열만을 포함하는 데이터만 추출하여 데이터 세트를 구성한다.

[0041] 두 번째 데이터 세트는 알려진 박테리아 종의 독성 단백질을 예측하기 위해 수집된 공개 데이터 세트이다. 이 데이터 세트는 183개의 양성 샘플과 382개의 음성 샘플로 구성되어 있으며, BTXpred 데이터로 표기한다.

[0042] 세 번째 데이터 세트는 서로 다른 유형의 단백질 서열을 포함하는 두 개의 개별 데이터 세트를 결합한다. VFDB 데이터 세트는 독성 인자로 표시된 25,288개의 박테리아 단백질로 구성되며, 세 번째 데이터 세트의 양성 샘플로 구성한다. toxinpred2는 17,821개의 비독소 박테리아 단백질로 구성되며, 세 번째 데이터 세트의 음성 샘플로 구성한다. 세 번째 데이터 세트는 Combined 데이터로 표기한다.

[0043] UniportKB로 표기된 데이터 세트는 검증용 데이터 세트로 SwissProt의 18,194개의 데이터 중 373개의 단백질이 독성 단백질로 레이블링 되어 있다. 이중 두 번째 데이터 세트인 BTXpred와 중복되는 127개를 제외해서 216개를 양성 샘플로 사용한다. 독성 단백질로 표기되지 않은 나머지 17,821개 중 무작위 1,000개를 선택하여 음성 샘플로 구성한다.

[0044] 도 3은 본 발명의 일 실시예에 따른 알려지지 않은 단백질 서열에 대한 검증 데이터 세트 구성을 설명하기 위해 나타낸 도면이다.

[0045] 독성이 표지되지 않은 단백질 서열에서 잠재적인 독성 관련 서열을 찾기 위해 실내 공기에서 흔히 발견되는 4개의 잘 알려진 박테리아 종을 선택한다. 선정된 종은 황색포도구균(*Staphylococcus aureus*), 황색소구균(*Micrococcus luteus*), 표피포도구균(*Staphylococcus epidermidis*) 및 고초균(*Bacillus subtilis*)이다.

[0046] 도 4는 본 발명의 일 실시예에 따른 알려지지 않은 단백질 서열에 검증 데이터 세트를 구성 방법을 설명하기 위해 나타낸 도면이다.

[0047] 도 3에서 선택한 박테리아 4종의 염기서열을 수집하기 위해 GenBank를 검색하여 단백질 염기서열을 다운로드 한다. 다운로드한 데이터에서 독성 단백질에 대한 종별 특성을 찾기 위해 다중으로 표시된 단백질을 제외하고, 이미 표지된 기능이 있는 단백질도 제외하는 데이터 프로세싱 과정을 진행한다.

[0048] 이렇게 진행되어 수집된 데이터는 도 3에 표기된 단백질 수를 나타낸다.

[0049] 도 5은 본 발명의 일 실시예에 따른 독성 동물성 단백질 평가 성능을 설명하기 위한 도면이다.

[0050] 도 5에서부터 도 8까지 본 발명의 성능을 검증하기 위해 4가지 지표를 사용한다. 4가지 지표는 F1 Score, MCC(Matthews correlation coefficient), AUPRC(area under the precision-recall curve) 및 AUROC(area under the receiver operating characteristic curve)이다. 평가에 사용된 메트릭은 TP(True Positive)는 독성 단백질을 독성 단백질로 분류한 진양성, TN(True Negative)은 무독성 단백질을 무독성 단백질로 분류한 진음

성이다. FP(False Positive)의 경우에는 독성 단백질을 무독성 단백질로 평가한 가양성을 FN(False Negative)는 무독성 단백질을 독성 단백질로 평가한 가음성을 나타낸다.

[0051] F1 Score를 계산하기 위한 계산식은 [수학식 2]와 같다

[0052] [수학식 2]

$$F1\ score = 1 + \frac{2TP}{FP+FN}$$

[0053]

[0054] MCC를 계산하기 위한 계산식은 [수학식 3]과 같다

[0055] [수학식 3]

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}}$$

[0056]

[0057] 도 5는 독성 동물 단백질 데이터 세트에 대해 본 발명과 이전 방법의 결과를 나타낸다. 이전 방법의 수치는 각 방법별 논문에 발표된 수치를 인용하였으며, 본발명의 결과는 가장 하단에 표시되어 있다. 도 5에서는 독성 동물 단백질 데이터에 대해 본 발명과 이전 방법과 성능 차이가 없음을 나타낸다.

[0058] 도 6은 본 발명의 일 실시예에 따른 독성 박테리아 단백질 평가 성능을 설명하기 위한 도면이다.

[0059] 독성 박테리아 단백질 평가 성능은 학습 데이터 세트를 활용한 BTXpred 방법과 비교한다. BTXpred 방법과 동일한 지표를 사용하여 평가하며, 본 발명이 정확도에서 +1.91%, MCC에서 +0.0141만큼 좋은 성능을 보이는 것을 확인할 수 있다. BTXpred는 독성 박테리아 단백질 평가라는 측면에서 본 발명과 목적이 동일하며, 향후 알려지지 않은 독성 단백질 평가에 비교 방법으로 계속 활용한다.

[0060] 도 7은 본 발명의 일 실시예에 따른 서로 다른 유형의 단백질 서열을 포함하는 개별 데이터 세트를 두 개 이상 결합한 데이터 세트로 학습한 독성 박테리아 단백질 평가 성능을 설명하기 위해 나타낸 도면이다.

[0061] 본 발명은 4개의 평가지표(F1, MCC, auROC, auPRC)에서 모두 우수한 결과를 보여준다. 도 5에서 확인한 동물 독성 단백질 검증보다 결과가 좋은 이유는 박테리아 단백질은 동물성 단백질 보다 30~40% 짧아 모델이 단백질 서열의 특징을 식별하기 쉽기 때문이다.

[0062] 도 8은 본 발명의 일 실시예에 따른 서로 다른 유형의 단백질 서열을 포함하는 개별 데이터 세트를 두 개 이상 결합한 데이터 세트로 학습한 모델과 독성 박테리아 단백질 평가를 위한 학습 데이터 세트로 학습한 독성 박테리아 단백질 평가 성능을 비교 설명하기 위해 나타낸 도면이다.

[0063] 검증 데이터 세트에 대해 본 발명과 BTXpred 방법을 비교하였으며, F1 +0.075, MCC +0.1205, auPRC에서 +0.1013 만큼 좋은 검증 결과를 확인할 수 있다.

[0064] 도 9는 본 발명의 일 실시예에 따른 실제 독성 박테리아 단백질 평가 성능을 상세하게 설명하기 위해 모델의 혼동행렬을 나타낸 도면이다.

[0065] 두 모델 모두 TP(진양성) 비율이 높았으며 대부분 독성 단백질을 독성 단백질로 올바르게 분류한다. 본 발명의 경우 더 높은 TP 비율을 보여주며, FP는 32개로 BTXpred보다 좋은 성능을 보여준다.

[0066] 도 10은 본 발명의 일 실시예에 따른 실제 독성 박테리아 단백질 평가에서 위험성이 있는 단백질로 평가한 단백질 서열 수를 나타낸 도면이다.

[0067] 4가지 박테리아가 생산한 단백질에 대해 BTXpred와 본 발명의 테스트 결과이다. 두 모델 역시 비슷한 성능으로 독성 단백질을 검출했으며, 특히 황색포도구균(*Staphylococcus aureus*)의 경우 218개의 독성 단백질을 더 검출하는 것을 확인할 수 있다.

[0068] 도 11은 본 발명의 일 실시예에 따른 박테리아 독성 평가를 위한 딥러닝 방법 과정을 나타낸 흐름도이다.

[0069] S110에서 박테리아가 생성한 단백질 서열이 입력되면, S120 단계에서 모델이 처리할 수 있도록 토큰화를 진행한다. 토큰화를 진행하면 모든 입력 데이터는 동일한 길이로 정리된다.

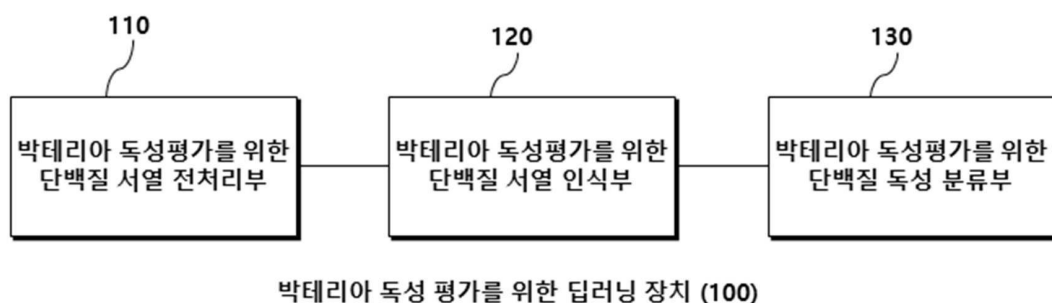
[0070] S130에서 토큰화된 데이터를 protBERT에 입력하면 단백질 서열을 인식한다. 인식한 단백질 서열의 독성 평가를 위해 독성 분류 단계로 입력한다. 이때 데이터의 결과를 뚜렷하게 하기 위해 S140의 드롭아웃 처리 단계와 S150의 완전 연결 처리 단계를 거쳐 S160의 독성 여부 최종판단 단계를 진행한다. 독성 여부 최종 판단은 tanh 함수

를 이용하여 자세한 방법은 상기 도 1에서 설명하였으므로 생략한다.

- [0071] 도 12는 본 발명의 일 실시예에 따른 박테리아 독성 평가를 위한 딥러닝 장치의 블록도이다.
- [0072] 도 12에서 도시한 바와 같이, 본 발명의 일 실시예에 따른 박테리아 독성 평가를 위한 딥러닝 장치는 입력 데이터 토큰 생성 계층(210), ProtBERT 기반 단백질 서열 인식 계층(220), 과적합 방지를 위한 드롭아웃 계층(230), 차원 축소를 위한 완전 연결 계층(240) 및 최종 결과 계층(250)으로 구성된다. 차원을 줄이면 단백질 서열의 특성이 더욱 뚜렷하게 나타난다.
- [0073] 입력 데이터 토큰 생성 계층(210)은 ProtBERT 모델을 사용하기 위한 전처리 계층으로 입력 단백질 서열을 동일한 길이의 토큰으로 변환한다. 변환된 토큰은 ProtBERT 모델의 입력값으로 사용된다.
- [0074] 독성 분류 모델(270)은 ProtBERT 기반 단백질 서열 인식계층(220) 및 독성 분류 계층(260)으로 구성된다. 독성 분류 모델(270)은 입력된 단백질 서열을 인식하고 독성 여부를 판단하는 모델이다. ProtBERT 기반 단백질 서열 인식 계층(220)은 단백질 서열을 인식하는 계층이므로 단백질 서열 인식 후 독성 여부를 판단하는 독성 분류 계층(230)이 필수요소이다.
- [0075] 독성 분류 계층(270)은 과적합 방지를 위한 드롭아웃 계층(230) 및 차원축소를 위한 완전 연결 계층(240)으로 구성된다.
- [0076] 독성 분류 계층(270)은 PortBERT 기반 단백질 서열 인식 계층(220)에서 인식한 단백질 서열의 독성 여부를 판단한다. 독성 여부를 보다 명확하게 판별하기 위해 과적합 방지를 위한 드롭아웃 계층(230)을 추가하여 모델의 미세조정을 실시한다. 본 발명의 최종 결과물은 독성 또는 무독성으로 분류되어야 하므로, 최종 판단 단계 전에 차원을 축소하여 독성 여부를 판단할 수 있도록 차원축소를 위한 완전 연결 계층(240)을 배치한다. 완전 연결 계층에서 사용하는 하이퍼볼릭탄젠트 함수(tanh)는 도 1에서 설명하였으므로 생략한다.
- [0077] 독성 분류 계층(260)에서는 단백질 서열에 대한 독성 여부를 최종 판단하여 최종 판단 계층(250)에 독성 또는 무독성으로 분류한다.
- [0078] 이상에서와 같이 본 발명의 박테리아 독성평가를 위한 딥러닝 장치 및 그 방법은 서로 다른 유형의 단백질 서열을 포함하는 개별 데이터 세트를 두 개 이상 결합한 데이터 세트로 학습하여 실제 박테리아에서 발생하는 단백질의 독성여부를 판단할 수 있도록 하는 효과가 있다.
- [0079] 또한 상기에서는 본 발명에 따른 바람직한 실시예를 위주로 상술하였으나 본 발명의 기술적 사상은 이에 한정되는 것은 아니며 본 발명의 각 구성요소는 동일한 목적 및 효과의 달성을 위하여 본 발명의 범위 내에서 변경 또는 수정될 수 있을 것이다.
- [0080] 아울러 이상에서는 본 발명의 바람직한 실시예에 대하여 도시하고 설명하였지만, 본 발명은 상술한 특정의 실시예에 한정되지 아니하며, 청구범위에서 청구하는 본 발명의 요지를 벗어남이 없이 당해 발명이 속하는 기술분야에서 통상의 지식을 가진 자에 의해 다양한 변형 실시가 가능한 것은 물론이고, 이러한 변형 실시들은 본 발명의 기술적 사상이나 전망으로부터 개별적으로 이해되어서는 안 될 것이다.

## 도면

### 도면1



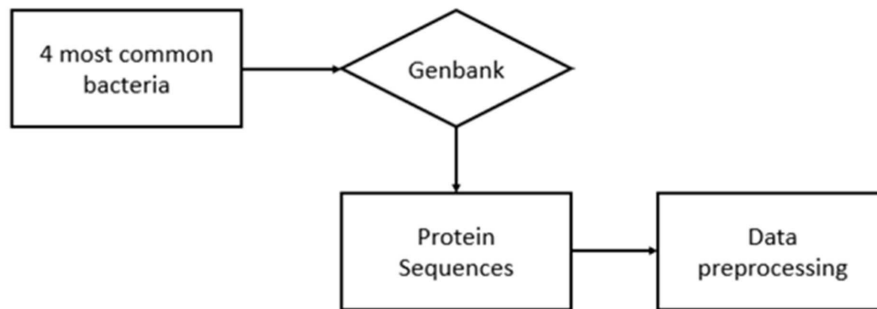
도면2

| Dataset   | Purpose        | Positive | Negative |
|-----------|----------------|----------|----------|
| ToxDL     | Training set   | 4413     | 5671     |
|           | Validation set | 59       | 670      |
| BTXpred   | Training set   | 140      | 402      |
|           | Validation set | 43       | 92       |
| Combined  | Training set   | 20,229   | 14,258   |
|           | Validation set | 5059     | 3563     |
| UniprotKB | Test set       | 216      | 1000     |

도면3

| Bacteria Species           | Number of Proteins |
|----------------------------|--------------------|
| Staphylococcus aureus      | 1497               |
| Miccroccus luteus          | 236                |
| Staphylococcus epidermidis | 27                 |
| Bacillus subtilis          | 344                |

도면4



도면5

| Method        | F1-Score | MCC   | auROC | auPRC |
|---------------|----------|-------|-------|-------|
| BLAST         | 0.800    | 0.801 | -     | -     |
| BLAST-score   | 0.789    | 0.775 | 0.868 | 0.818 |
| InterProScan  | 0.347    | 0.402 | -     | -     |
| Hmmsearch     | 0.185    | 0.307 | -     | -     |
| ClanTox       | 0.620    | 0.604 | 0.903 | 0.612 |
| ToxinPred-RF  | 0.667    | 0.638 | 0.948 | 0.716 |
| ToxinPred-SVM | 0.677    | 0.648 | 0.939 | 0.712 |
| ToxDL         | 0.809    | 0.793 | 0.989 | 0.913 |
| ToxIBTL       | 0.830    | 0.816 | 0.953 | 0.847 |
| 본 발명          | 0.833    | 0.818 | 0.915 | 0.814 |

도면6

| Method  | Accuracy | F1     | MCC    | auROC  | auPRC  |
|---------|----------|--------|--------|--------|--------|
| BTXpred | 96.07%   | -      | 0.9293 | -      | -      |
| TOXIFY  | 83.35%   | -      | -      | -      | -      |
| 본 발명    | 97.98%   | 0.9579 | 0.9434 | 0.9671 | 0.9647 |

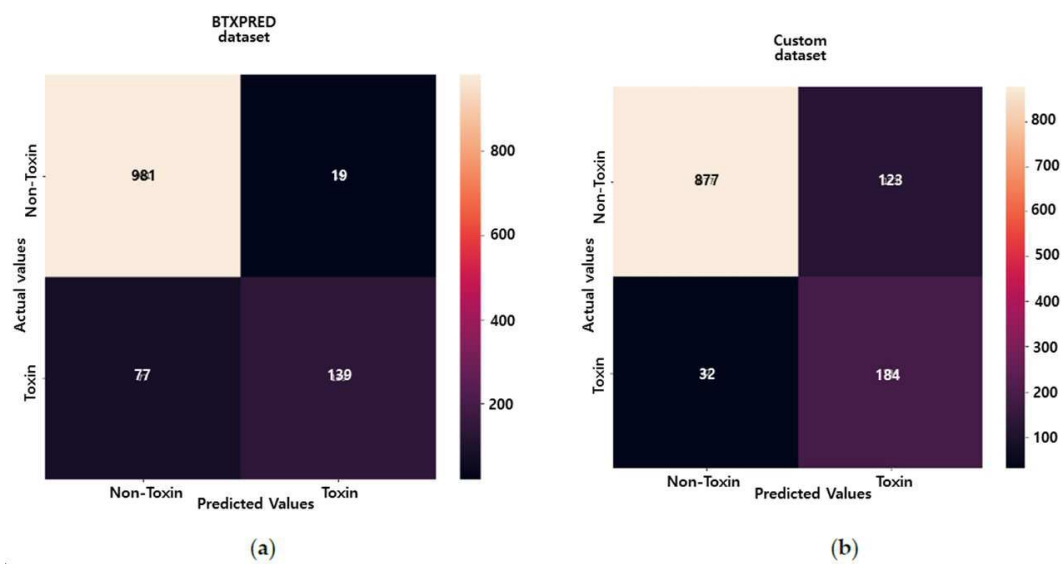
도면7

| Method | F1     | MCC    | auROC  | auPRC  |
|--------|--------|--------|--------|--------|
| 본 발명   | 0.9527 | 0.8845 | 0.9645 | 0.9408 |

도면8

| Dataset Used for Training | F1     | MCC    | auPRC  | auROC  |
|---------------------------|--------|--------|--------|--------|
| BTXpred                   | 0.7790 | 0.7617 | 0.8401 | 0.8239 |
| Combined                  | 0.7036 | 0.6412 | 0.7388 | 0.8644 |

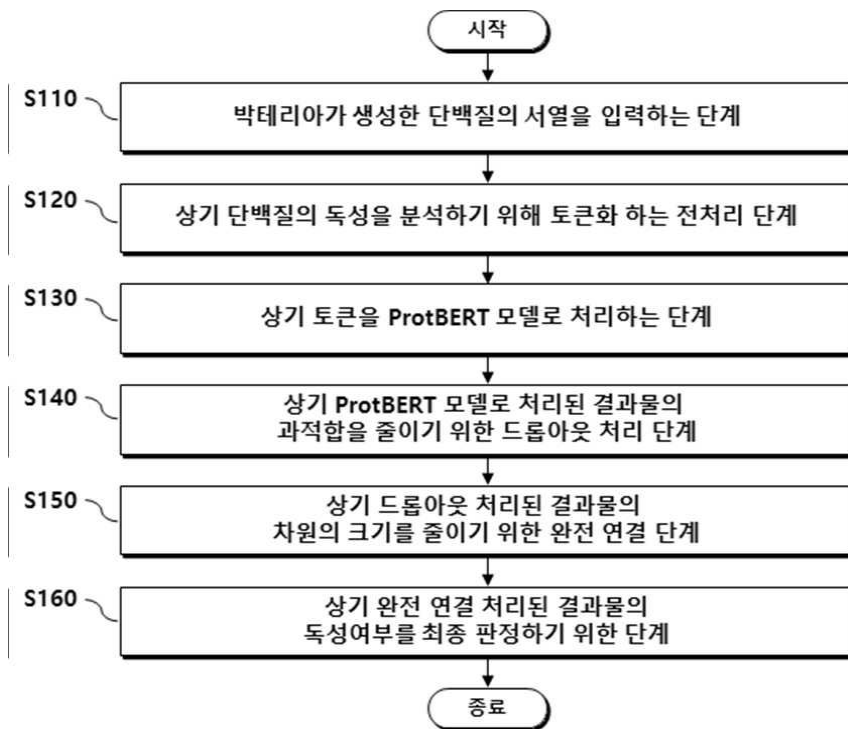
도면9



도면10

| Dataset Used for Training | Staphylococcus Aureus | Micrococcus Luteus | Staphylococcus Epidermidis | Bacillus Subtilis |
|---------------------------|-----------------------|--------------------|----------------------------|-------------------|
| BTXpred                   | 45                    | 43                 | 5                          | 21                |
| Combined                  | 263                   | 49                 | 5                          | 27                |

도면11



도면12

