



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0145210
(43) 공개일자 2021년12월01일

(51) 국제특허분류(Int. Cl.)
G16B 40/00 (2019.01) G06N 3/08 (2006.01)
G16B 5/00 (2019.01) G16H 10/40 (2018.01)
G16H 50/20 (2018.01) G16H 50/70 (2018.01)
(52) CPC특허분류
G16B 40/00 (2019.02)
G06N 3/08 (2013.01)
(21) 출원번호 10-2021-7034694
(22) 출원일자(국제) 2020년03월27일
심사청구일자 없음
(85) 번역문제출일자 2021년10월26일
(86) 국제출원번호 PCT/US2020/025502
(87) 국제공개번호 WO 2020/205649
국제공개일자 2020년10월08일
(30) 우선권주장
62/826,228 2019년03월29일 미국(US)

(71) 출원인
벤 바이오사이언시스 코포레이션
미국, 캘리포니아 94080, 사우스샌프란시스코, 5
플로어, 투 타워 플레이스
(72) 발명자
세리에 다니엘
미국, 캘리포니아 94080, 사우스샌프란시스코, 5
플로어, 투 타워 플레이스 벤 바이오사이언시스
코포레이션 내
우 첸친
미국, 캘리포니아 94080, 사우스샌프란시스코, 5
플로어, 투 타워 플레이스 벤 바이오사이언시스
코포레이션 내
(74) 대리인
김태홍, 김진희

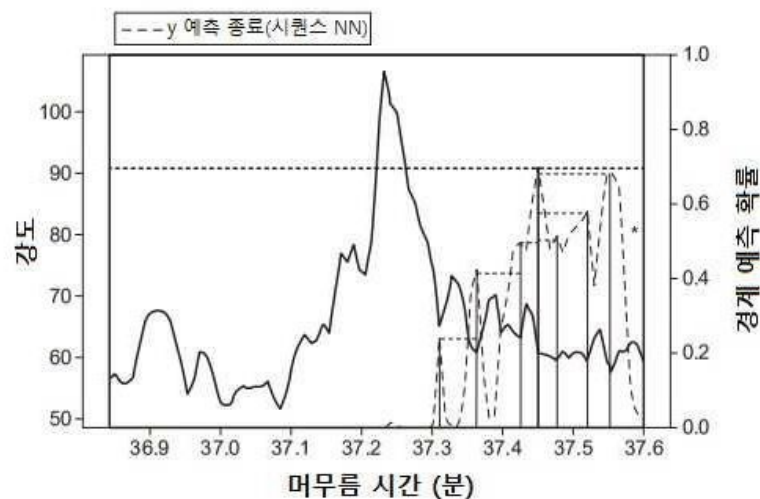
전체 청구항 수 : 총 26 항

(54) 발명의 명칭 질량 분석 데이터의 경계 자동 검출

(57) 요약

빈도 또는 시간 종속 데이터로 표현되거나 파생된 강도를 기반으로 양의 유무를 자동으로 감지하는 시스템 및 방법. 일 예시에 따르면, 질량 분석법으로부터의 강도는 존재비가 결정될 수 있는 강도의 시작 및 중지 피크를 찾도록 훈련된 인공 신경망과 같은 비선형 수학적 모델을 사용하여 식별된다.

대표도



(52) CPC특허분류

G16B 5/00 (2019.02)

G16H 10/40 (2021.08)

G16H 50/20 (2018.01)

G16H 50/70 (2018.01)

명세서

청구범위

청구항 1

질량 분석 피크(mass spectrometry peak)들을 식별하기 위해 모델을 훈련시키는 방법에 있어서,

복수의 라벨링된(labeled) 시퀀스 데이터(sequential data)를 사용하는 단계-상기 복수의 라벨링된 시퀀스 데이터의 각각은 글리코펩티드 또는 펩티드, 또는 이의 단편들의 특징적인 머무름 시간(retention time) 값들을 나타내는 피크 시작 및 피크 중지 라벨을 포함 함-, 및

상기 라벨링된 시퀀스 데이터를 사용하여 적어도 하나의 기계 학습 모델(machine learning model)을 훈련시키는 단계를 포함하고, 여기서 상기 훈련된 모델은 생물학적 샘플에 존재하는 하나 이상의 글리코펩티드 또는 펩티드, 또는 이의 단편의 존재를 나타내는 질량 분석 시작 및 중지 피크들을 식별하도록 적용되는, 방법.

청구항 2

제1항에 있어서, 상기 생물학적 샘플로부터의 상기 질량 분석 시작 및 중지 피크들 사이의 시간 경과에 따른 통합(integration)에 의해 상기 하나 이상의 글리코펩티드 또는 펩티드의 존재를 정량화하는 단계를 더 포함하는, 방법.

청구항 3

제1항에 있어서, 상기 기계 학습 모델은 순환 신경망(recurrent neural network) 모델인, 방법.

청구항 4

제3항에 있어서, 상기 순환 신경망은 관심 계층(attention layer)을 갖는 양방향 LSTM 순환 신경망 모델인, 방법.

청구항 5

제1항의 방법에 의해 훈련된 기계 학습 모델을 사용하여 글리코펩티드 또는 펩티드에 대한 존재비(abundance)를 계산하는 단계를 포함하는, 방법.

청구항 6

시스템에 있어서,

라벨링된 시퀀스 데이터의 데이터 저장소;

상기 저장소에 결합된 하나 이상의 프로세서들;

상기 하나 이상의 프로세서들에 의해 실행될 때 제1항 내지 제5항 중 어느 한 항의 방법을 수행하는, 상기 하나 이상의 프로세서들에 액세스 가능한 비일시적 메모리에 상주하는, 기계 실행 가능 코드(machine executable code)를 포함하는, 시스템.

청구항 7

질량 분석 피크들을 식별하는 방법에 있어서,

입력 파라미터들의 함수로서 출력을 생성하는 적어도 하나의 훈련된 신경망 모델을 제공하는 단계-여기서 상기 입력 파라미터들에 대한 값들은 m/z (질량 대 전하) 비율을 갖는 글리코펩티드에 대한 머무름 시간 값들의 시퀀스를 포함 함-; 및

생물학적 샘플의 질량 분석법으로부터의 상기 머무름 시간 값들의 시퀀스에 대한 질량 대 전하(m/z) 비율들을 포함하는 질량 분석(MS) 데이터를 상기 입력 파라미터들에 대한 값들로 제공하는 단계를 포함하고;

상기 출력은 상기 생물학적 샘플에서 하나 이상의 분석물 또는 이의 단편의 존재를 나타내는 하나 이상의 피크

시작 및 피크 중지 머무름 시간 값들에 대한 위치를 포함하는, 방법.

청구항 8

제7항에 있어서, 상기 생물학적 샘플로부터의 상기 질량 분석 시작 및 중지 피크들 사이의 머무름 시간 값들을 합산하거나 통합함으로써 상기 하나 이상의 글리코펩티드 또는 펩티드의 존재를 정량화하는 단계를 더 포함하는, 방법.

청구항 9

제8항에 있어서, 상기 정량화 단계는 상기 예측된 시작 및 중지 피크들에 대해 가중 평활 함수(weighted smoothing function)를 적용하는 단계를 포함하는, 방법.

청구항 10

시스템에 있어서,

질량 분석(MS) 데이터의 데이터 저장소;

상기 저장소에 결합된 하나 이상의 프로세서들;

상기 하나 이상의 프로세서에 의해 실행될 때 제7항 내지 제9항 중 어느 한 항의 방법을 수행하는, 상기 하나 이상의 프로세서들에 액세스 가능한 비일시적 메모리에 상주하는, 기계 실행 가능 코드를 포함하는, 시스템.

청구항 11

방법에 있어서,

원시 MS 데이터로부터의 데이터를 특징화하는 단계(featurizing)를 포함하고, 상기 원시 MS 데이터는, 각각이 질량/전하 강도를 나타내는, 복수의 아날로그 샘플들을 포함하고, 상기 특징화 단계는, 상기 아날로그 샘플들의 각각에 대해,

머무름 시간 윈도우 내에서 상기 샘플을 중앙에 배치하는 단계,

상기 샘플을 강도 값들을 나타내는 포인트들의 시퀀스로 이산화하고, 상기 강도 값들을 표준화하는 단계, 및

라벨링된 포인트들의 시퀀스를 생성하기 위해 상기 포인트들의 시퀀스 중 피크 시작 및 피크 중지 시간들에 대응하는 포인트들에 라벨들을 할당하는 단계를 포함하고;

상기 라벨링된 포인트들의 시퀀스는 보이지 않는(unseen) MS 데이터의 존재비를 예측하기 위해 기계 학습 모델을 훈련하도록 구성되는, 방법.

청구항 12

제11항에 있어서, 상기 중앙에 배치하는 단계는 상기 머무름 윈도우의 중심에 곡선의 정점을 위치시키는 단계를 포함하는, 방법.

청구항 13

제11항에 있어서, 상기 강도 값들을 표준화하는 단계는 임계값보다 큰 강도 값을 제외하는, 방법.

청구항 14

제13항에 있어서, 상기 임계 값이 약 500인, 방법.

청구항 15

시스템에 있어서,

복수의 아날로그 샘플들을 포함하는 원시 질량 분석(MS) 데이터의 데이터 저장소;

상기 저장소에 결합된 하나 이상의 프로세서들;

상기 하나 이상의 프로세서에 의해 실행될 때 제11항 내지 제14항 중 어느 한 항의 방법을 수행하는, 상기 하나

이상의 프로세서들에 액세스 가능한 비일시적 메모리에 상주하는, 기계 실행 가능 코드를 포함하는, 시스템.

청구항 16

방법에 있어서,

주어진 m/z 비율 및 머무름 기간에 대한 질량 분석(MS) 데이터를 나타내는 포인트들의 시퀀스에 대한 확률들의 분포를 계산하는 단계-상기 확률들은 피크 시작 또는 피크 중지 머무름 시간을 나타냄-, 및

상기 확률들의 분포를 사용하여 존재비에 대한 통계를 계산하는 단계를 포함하고, 상기 통계는 상기 MS 데이터에서 생물학적 샘플의 존재를 나타내는, 방법.

청구항 17

제16항에 있어서, 상기 통계는 적분을 평가하기 위한 시작 및 끝 포인트들의 가능한 모든 쌍들을 사용하여 적분들의 가중 합으로 계산된 평균 존재비인, 방법.

청구항 18

제1항, 제7항 또는 제11항 중 어느 한 항에 있어서, 상기 기계 학습 모델은 시퀀스 모델(모델 A) 또는 참조 기반 모델(모델 B)인, 방법.

청구항 19

방법에 있어서,

복수의 라벨링된 시퀀스 데이터를 사용하는 단계, 및

상기 라벨링된 시퀀스 데이터를 사용하여 적어도 하나의 기계 학습 모델을 훈련시키는 단계를 포함하고,

상기 훈련된 모델은 보이지 않는 샘플에서 강도의 시작 및 중지 피크가 발생하는 시간 주기 또는 주파수 범위를 식별하도록 적응되는, 방법.

청구항 20

제19항에 있어서, 상기 시작 및 중지 피크는 잡음에서 신호로의 전환 기간에 포함되는, 방법.

청구항 21

제19항에 있어서, 상기 기계 학습 모델은 주의 계층을 갖는 순환 신경망인, 방법.

청구항 22

제19항에 있어서, 상기 샘플은 상기 강도의 시작 피크에서 중지 피크까지의 시간 또는 주파수에 대한 합산 또는 적분이 물리적인 것(physical thing)을 정량화하도록 상기 물리적인 것을 나타내는 방법.

청구항 23

제19항에 있어서, 상기 샘플은 지정된 시간 기간 동안 질량 대 전하 강도들의 시퀀스 시리즈들이고 상기 합산 또는 적분은 상기 물리적인 것의 존재비를 나타내는, 방법.

청구항 24

방법에 있어서,

입력 파라미터들의 함수로서 출력을 생성하는 적어도 하나의 훈련된 신경망 모델을 제공하는 단계-상기 입력 파라미터들에 대한 값들은 양을 나타내는 강도 값들의 시간적 또는 스펙트럼 시퀀스임-; 및

상기 입력 파라미터들의 각각에 대한 값들을 제공하는 단계를 포함하고, 상기 값들은 양에 대해 보이지 않는 강도들의 시퀀스를 나타내고;

상기 출력은 상기 양을 정량화하는 데 사용되는 하나 이상의 피크 시작 및 피크 중지 시간들 또는 주파수들을 포함하는, 방법.

청구항 25

제24항에 있어서, 상기 양은 상기 피크 시작 및 상기 피크 중지 시간들 또는 주파수들 사이의 시간 또는 주파수에 대한 합산 또는 적분에 의해 각각 정량화되는, 방법.

청구항 26

제24항 또는 제22항에 있어서, 여기서 상기 강도들은 질량 분석법, MRM, 액체 또는 기체 크로마토그래피, X선 회절, 라만 분광법, UV-VIS 분광법, 형광분석 분광법 또는 인광분석 분광법, 핵 자기 공명(NMR) 분광법, 또는 자기 공명 영상(MRI)의 세트로부터 선택된 프로세스로부터 얻어지는, 방법.

발명의 설명

기술 분야

[0001] 반복된 출원에 대한 상호 참조

[0002] 본 출원은 2019년 3월 29일에 출원된 미국 가특허출원 제62/826,228호의 우선권 및 이익을 주장하며, 그 전체 내용은 모든 목적을 위해 그 전체가 참고로 본원에 포함된다.

[0003] 분야

[0004] 본 개시는 일반적으로 질량 분석법(mass spectrometry)을 사용한 물리적 사물의 검출, 특히 다중 반응 모니터링 질량 분석법(MRM-MS)을 사용한 생물학적 양(biological quantity)의 자동화된 검출에 관한 것이다.

배경 기술

[0005] 다중 반응 모니터링(MRM)은 탠덤(tandem) 질량 분석(MS)에 사용되는 기술이며, 이는 다른 분석물 중에서 단백질, 지질 및 번역 후 변형(PTM)의 매우 민감하고 특이적인 검출을 가능하게 한다. 이 기술은 새로운 임상 분석 및 진단에 사용할 가능성이 높지만, 정확한 분석 소프트웨어의 가용성은 여전히 병목 상태이다.

[0006] MRM 실험에 대한 기존의 분석에는 이전에 데이터 의존적 수집 또는 기타 발견 기술로 특징지어졌던 전환에서 작업하여 손으로 크로마토그래피 피크(chromatographic peak)의 시작과 정지를 선택하는 작업이 포함되었다. 일반적인 구현은 애질런트(Agilent)의 메스헌터(MassHunter) 소프트웨어에서 볼 수 있으며, 여기서 강도는 주어진 전구체(precursor) 및 제품 질량 대 전하 비율에 대해 미리 지정된 머무름 시간(RT) 범위에 대해 표시된다. 피크의 시작과 끝을 선택하면 관찰된 RT, 피크 폭 및 통합된 존재비 값(abundance value)이 산출된다. 평가의 개인 간 변동성, 인적 오류 및 많은 시간 투자로 인해 이 방법은 연구 이외의 용도로는 부적절하다. 최근 몇 년 동안 많은 소프트웨어 패키지가 전환들의 리스트에서 피크들을 자동으로 통합하여 이러한 갭을 메우려고 시도하였다.

[0007] 기존 소프트웨어 패키지는, 질량 분석 데이터에서 상대적 존재비(abundance present)가 높을 때, 예를 들어 펩타이드(peptide)를 정량화하는 데 매우 유용한 것으로 입증되었다. 그러나 이러한 소프트웨어 패키지는 글리코실화된 펩티드와 같이 상대적으로 풍부하지 않고 매우 이질적인 종(heterogeneous species)을 통합하는 데 덜 정확하다.

발명의 내용

해결하려는 과제

과제의 해결 수단

[0008] 본 개시는 시계열 데이터로 표현되거나 시계열 데이터로부터 유도된 양(quantity)의 존재를 자동으로 검출하기 위한 방법 및 시스템을 설명한다. 본 개시는 또한 주파수 의존적 데이터로 표현되거나 그로부터 유도된 양의 존재를 자동으로 검출하기 위한 방법 및 시스템을 제시한다.

[0009] 일 구현에서, 자동화된 검출은 시간에 대한 강도의 시계열에 적용된다. 강도는 양의 존재 또는 부재를 감지하는 검출기로부터 나온다. 예를 들어, 자동 검출은 강도가 시간의 함수, 예를 들어 머무름 시간으로서 전환 값을 검

출하는 MS 검출기로부터 나오는 질량 분석(MS)에 적용될 수 있다. 이와 같이, 본원에 기술된 자동화된 검출은 시간의 함수 또는 주파수의 함수로서 관찰된 강도 값에 기초하여 광범위한 데이터 입력에 적용될 수 있다.

- [0010] 일 구현에서 자동화된 검출은 시간에 대한 질량 대 전하 비율 강도의 시계열에 적용되며, 이는 질량 분석(MS)을 포함하는 공정에서 생산된 것이다. 다른 구현에서 자동화된 검출은 시간에 대한 강도의 시계열에 적용되며, 이는 액체 또는 기체 크로마토그래피와 관련된 공정에서 생산된 것이다.
- [0011] 다른 예에 따르면, 자동화된 검출은 시간 또는 주파수 데이터로 표현되거나 유도된 다음 강도 중 임의의 것에 적용될 수 있다:
- [0012] · 강도가 x-선 검출기, 예를 들어, 스캔 각도가 시간의 함수로 관찰되는 x-선 스캔 각도, 예를 들어, $2 - \theta$ 의 함수로서 x-선의 반사를 검출하는 전하 결합 디바이스(CCD)에서 파생되는 x-선 회절.
- [0013] · 광 파장, 예를 들어 cm^{-1} , 흡수 또는 방출의 함수로 진동 모드를 검출하는 적외선 검출기에서 강도가 파생되는 라만 분광법.
- [0014] · 자외선 및 가시광선 파장 또는 주파수의 함수로 전자 전환의 강도를 검출하는 광 검출기에서 강도가 파생되는 UV-VIS 분광법, 방출된 형광 및 인광 광 강도가 입사, 즉 흡수 또는 여기 파장의 함수로 검출되는 형광 분광법 또는 인광 분광법.
- [0015] · 강도가 강한 자기장 내에서 복사 주파수 흡수 또는 방출의 함수로 핵 전환을 감지하는 전파 검출기에서 파생되는 NMR(Nuclear Magnetic Resonance) 분광법. NMR에서, 강도 대 주파수 검출은 예를 들어 푸리에 변환을 사용하여 동등한 강도 대 시간으로 변환될 수 있다. NMR에 적용할 수 있는 동일한 원리는 자기 공명 영상(MRI)에도 적용할 수 있다.
- [0016] 하나의 특정 구현에서 자동 검출은 질량 대 전하 비율 강도 대 시간의 시계열에 적용되며, 이는 질량 분석을 포함하는 공정에서 생성된다. 이 구현에 따르면 자동화된 검출은 MS 데이터에서 시작 피크와 종료 피크가 발생하는 위치를 감지하기 위해 하나 이상의 수학적 모델을 사용하는 것을 포함한다. 이러한 경계 포인트의 정확한 선택을 제공함으로써 존재비(abundance)에 대한 보다 정확한 평가가 이루어진다.
- [0017] 일 양태에 따르면, 질량 분석 피크를 식별하기 위해 기계 학습 모델을 훈련시키는 방법은 피크 시작 및 피크 중지 라벨링되는 질량/전하 비율을 나타내는 포인트의 훈련 세트 시퀀스를 사용한다. 훈련된 모델은 생물학적 샘플에 존재하는 하나 이상의 글리코펩티드 또는 펩티드, 또는 이의 단편의 존재를 나타내는 질량 분석 시작 및 중지 피크를 식별하도록 조정된다.
- [0018] 다른 양태에 따르면, 질량 분석 피크를 식별하는 방법은 입력 파라미터의 함수로서 출력을 생성하는 적어도 하나의 훈련된 신경망 모델을 사용하고, 여기서 입력 파라미터에 대한 값은 m/z (질량 대 전하) 비율을 갖는 글리코펩티드에 대한 체류 시간 값의 시퀀스일 수 있다.
- [0019] 다른 양태에 따르면, 원시 MS 데이터는 MS 데이터에서 피크 시작 및 중지 시간을 식별하기 위해 모델을 훈련할 목적으로 기계 학습 모델에 대한 훈련 세트를 생성하도록 특징화된다. 특징화는 MS 데이터의 아날로그 샘플을 포인트의 시퀀스로 이산화하는 것을 포함할 수 있으며, 여기서 아날로그 샘플의 피크 영역에 해당하는 머무름 시간은 포인트 시퀀스에서 제외되고, 피크 시작 및 피크 중지 시간에 해당하는 포인트 시퀀스 중 포인트에 라벨을 할당하여 라벨이 지정된 포인트 시퀀스를 생성한다. 라벨은 디스플레이 및 입력 디바이스를 사용하여 피크 시작 및 중지 시간을 선택하도록 훈련된 인간 주석자를 사용하여 찾을 수 있다.
- [0020] 다른 양태에 따르면, 방법은 주어진 m/z 비율 및 머무름 기간 동안 MS 데이터를 나타내는 포인트 시퀀스에 대한 확률 분포를 계산하며, 확률은 피크 시작 또는 피크 중지 머무름 시간을 나타낸다. 이러한 확률 분포에서 존재비에 대한 통계가 계산된다. 통계는 평균일 수 있다.
- [0021] 또 다른 양태에 따르면, 질병을 진단하는 방법은 질량 분석 전환을 정량화하는 단계-여기서 전환은 아미노산 시퀀스를 포함하는 글리코펩티드 또는 펩티드, 또는 이의 단편의 특징인 m/z 강도 값 피크임-, 및 샘플이 정량화를 기반으로 질병 분류에 속하거나 속하지 않을 확률을 결정하기 위해 모델을 사용하여 정량화를 분석하는 단계를 포함한다. 일부 예에서, 아미노산 시퀀스를 포함하는 글리코펩티드 또는 펩티드, 또는 그의 단편은 이들 시퀀스로부터 선택되며, 이는 국제특허출원 공개번호 WO 2019/046814-진단 및 치료 모니터링을 위한 바이오마커로서 글리코펩티드의 식별 및 사용; 국제특허출원 공개번호 WO2019079639A1-진단 및 치료 모니터링을 위한 생물학적 파라미터의 식별 및 사용; 미국 특허 출원 공개 번호 US20190101544A1-진단 및 치료 모니터링을 위한 바이오

마커로서 글리코펩티드의 식별 및 사용; 2020년 1월 31일자로 출원된 국제 PCT 특허 출원 번호 PCT/US2020/0162861; 및 또한 2020년 3월 13일에 출원된 미국 가특허출원 제62/989,510호에 설명되어 있다. 이들 특허 출원 각각의 전체 내용은 모든 목적을 위해 그 전체가 참고로 본원에 포함된다.

도면의 간단한 설명

도 1은 기계 학습 모델을 사용하여 질량 분석 데이터로부터 존재비를 계산하는 프로세스를 보여준다.

도 2는 도 1의 기계 학습 모델을 훈련하기 위한 프로세스를 도시한다. 테스트 세트에서 모델 성능. A: 테스트 세트의 경계 예측(boundary prediction)에서 절대 오차의 막대 그래프(Bar plot). B: 경계 예측 오류의 누적 분포(cumulative distribution). 19039개(28.1%) 샘플들은 규칙 기반 예측에서 범위(> 6초)를 벗어났고, 순차 NN에서는 4824개(7.1%) 샘플들, 참조 기반 NN에서는 1470개(2.2%) 샘플이다. C: 테스트 세트에 대한 피크 존재비 예측/주석의 산점도(scatter plot).

도 3은 도 2의 기계 학습 모델에 대한 입력으로 사용되는 MS 샘플로부터 포인트들의 시퀀스를 생성하기 위한 프로세스를 도시한다.

도 4는 MS 샘플에서 피크 시작 및 피크 정지 머무름 시간의 계산된 확률에 기초하여 존재비를 계산하기 위한 프로세스이다.

도 5a 및 도 5b는 장기 단기 메모리 유닛 및 주의 계층(attention layer)을 갖는 양방향 순환 신경망(recurrent neural network)인 제1 기계 학습 모델(모델 A)에 대한 모델 아키텍처(architecture)를 도시한다.

도 6a 및 6b는 장기 단기 메모리 유닛 및 주의 계층을 갖는 양방향 순환 신경망인 제2 기계 학습 모델(모델 B)에 대한 모델 아키텍처를 도시한다. 도 6c는 모델 B를 사용하는 샘플에 대한 주의 맵(attention map)을 도시한다. 도 6c에서, 샘플 피크 및 참조 기반 모델 예측은 모든 목적을 위해 그 전체가 참조로 본원에 포함되며, 2020년 1월 31일에 출원된 국제 PCT 특허 출원 번호 CT/US2020/0162861의 SEQ ID NO: 1에 대한 MRM 전환에 대한 것이며, 참조 피크는 오른쪽 하단 패널에, 주의 맵(모델에서 추출)은 왼쪽 하단 패널에 표시된다.

도 7a 및 7b는 각각 모델 A 및 모델 B에 대한 샘플 라벨링된 입력 및 계산된 확률(경계 예측)을 보여준다. 별표 및 삼각형 주석은 모델 학습을 위한 MS 데이터의 라벨링을 나타낸다. 녹색 곡선은 모델 B를 사용한 예측을 나타내고 파란색 곡선은 모델 A를 사용한 예측을 나타낸다.

도 8a, 8b 및 8c는 동일한 데이터(규칙 기반 방법 - 빨간색, 1)에 대해 훈련된 알려진 기계 학습 아키텍처에 대한 모델 A(파란색, 2) 및 모델 B(녹색, 3)의 성능을 보여준다. 도 10에서 세 막대들의 각 세트는 왼쪽에서 오른쪽으로 다음과 같이 표시된다; 규칙 기반, MAE 4.79s; 순차 NN, MAE 2.33s; 및 참조 기반 NN, MAE 1.56s.

도 9a 및 도 9b는 각각 모델 A 및 모델 B에 대한 입력 및 계산된 확률(경계 예측)로 라벨링된 추가 샘플을 도시한다. 별표 및 삼각형 주석은 모델 학습을 위한 MS 데이터의 라벨링을 나타낸다. 녹색 곡선은 모델 B를 사용한 예측을 나타내고 파란색 곡선은 모델 A를 사용한 예측을 나타낸다. 글리코펩티드 전환에 대한 참조 및 샘플 피크. A: 참조 피크에서, 회색 선은 입력 곡선을 나타내고; 빨간 별과 빨간 삼각형은 사람이 주석을 단(human annotated) 피크 시작과 끝을 나타내고; 파란색/녹색 실선 및 점선은 피크 시작 및 끝 확률의 순차/참조 기반 신경망 모델의 예측을 나타낸다. B: 입력 곡선이 다른 동일한 전환에 대한 샘플 피크. 참조 기반 모델들(녹색)은 훨씬 더 일관된 예측을 출력한다.

도 10 인간 주석자(회색, 1) 대 모델 A(파란색, 2) 및 모델 B(녹색, 3)에 대한 경계 포인트 선택의 상대 표준 편차를 도시한다. 막대 그래프들은 13명의 인간 주석자들의 상대 표준 편차(RSD)와 순차/참조 기반 신경망 예측의 상대 오차(RE)이다. 도 10에서 세 막대들의 각 세트는 왼쪽에서 오른쪽으로 다음과 같이 표시된다: 순차 NN, 평균(Mean) RE 0.034; 참조 기반 NN, 평균 R=0.019; 및 인간 주석자들, 평균 = 0.025.

도 11a 및 11b는 각각 모델 A와 모델 B 사이의 존재비 계산을 비교하는 산점도이다. 순차(A) 및 참조 기반(B) 신경망으로부터 피크 존재비 예측/주석의 산점도. 순차적 모델 예측에서 가중된 존재비(보라색 점)는 0.9969에서 더 높은 피어슨 상관 계수(Pearson's r)과 함께 argmax 존재비(파란색 포인트)보다 실측 자료(ground truth)와 더 높은 일관성과 더 적은 이상점(outlier)을 보여준다. 가중된 풍부함은 낮은 풍부함 범위에서 상향으로 편향된다. 참조 기반 예측에서는 결과가 매우 유사하다.

도 12는 모델 B에 대한 신뢰 점수와 신뢰 점수를 계산하는 데 사용되는 방정식을 보여준다. 신뢰 점수 예시: 파란색 점선(*) - 모델 예측 피크 종료 확률; 진한 빨간색 선(#) - argmax 예측; 밝은 빨간색 선(!) - 기타 로컬

최대값. 신뢰 점수(0.634)는 음영 영역의 영역을 검은색 상자의 영역으로 나눈 것이다.

도 13은 모델 A에 대한 예측 신뢰도의 히스토그램, 경계 예측의 평균 절대 오차 및 상이한 신뢰 점수 bin(bin)들에 대한 산점도를 나타낸다. A: 순차 신경망으로부터의 예측 신뢰도 히스토그램. 네 개의 bin들로부터의 샘플들의 존재비 예측은 다음 패널들에 설명되어 있다. B: 경계 예측의 평균 절대 오차(MAE)는 신뢰도가 높은 샘플들에서 꾸준히 감소한다. C, D, E, F: 다양한 신뢰 점수 bin들에서 피크 존재비 예측/주석의 산점도. 신뢰도가 높은 샘플은 더 나은 상관 점수를 얻었다. 보고된 log-존재비 피어슨 상관 계수 값은 서로 다른 전환들의 샘플들의 조합에서 계산된다.

도 14는 모델 B에 대한 예측 신뢰도의 히스토그램과 경계 예측의 평균 절대 오차를 보여준다. A: 참조 기반 신경망으로부터의 예측 신뢰도의 히스토그램. B: 경계 예측의 평균 절대 오차(MAE)는 신뢰도가 높은 샘플들에서 꾸준히 감소한다.

발명을 실시하기 위한 구체적인 내용

- [0023] 설명에서 도면에 나타나는 동일한 참조 번호 및 설명은 상이한 뷰(view)들 중에서 대응하거나 유사한 요소를 지칭한다.
- [0024] 정의
- [0025] 본 명세서에서 사용되는 하기 단어 및 구문은, 그들이 사용된 문맥에서 명백하게 달리 지시하는 경우를 제외하고, 다음과 같은 의미를 갖는 것으로 의도된다.
- [0026] "전환(transition)"은 질량 분석법에서 질량 대 전하 비율 강도의 시간적 시퀀스(temporal sequence)를 나타낸다. 본원에 사용된 바와 같이, "전환" 또는 "다중 반응 모니터링(MRM) 전환"라는 문구는 글리코펩티드 또는 이의 단편(fragment)이 MRM-MS에 의해 검출되는 경우 관찰된 질량 대 전하(m/z) 피크 또는 신호를 의미한다. MRM 전환은 전구체와 생성 이온의 전환으로서 검출된다. 전환 감지를 위한 시간 간격은 탠덤 액체 크로마토그래피 기기의 공급 속도와 MRM 분광기(spectrometer) 내의 유속에 따라 다르다.
- [0027] "생물학적 양(biologic quantity)"은 생물학적 샘플에 있는 분석물의 양 또는 탠덤 크로마토그래피 기기의 용질의 양을 의미하는 것으로 이해된다. 분석물 및 용질에는 글리칸, 펩티드, 글리코펩티드, 글리코실화된 펩티드, 프로틴, 글리코프로틴 및 이의 단편이 포함되나 이에 제한되지는 않는다.
- [0028] 본 명세서에 사용된 바와 같이, "생물학적 샘플"이라는 문구는 유기체로부터 유래, 획득, 생성, 제공, 취하거나 유기체로부터 제거된 샘플을 지칭하고; 또는 유기체의 체액 또는 조직으로부터의 샘플을 의미한다. 생물학적 샘플에는, 이에 제한되는 것은 아니지만, 혈액, 전혈, 혈청, 혈장, 소변, 가래, 조직, 타액, 눈물, 척수액, 생검으로 얻은 조직 절편; 조직 배양에 배치되거나 조직 배양에 적응된 세포(들); 땀, 점액, 대변, 위액, 복액, 양수, 낭종액, 복막액, 체장액, 모유, 폐세척, 골수, 위산, 담즙, 정액, 고름, 방수, 경질액, 및 전술한 것의 파생물, 부분 및 조합을 포함하는 유사물을 포함한다. 일부 예시들에서 생물학적 샘플에는 혈액 및/또는 혈장이 포함되지만 이에 제한되지는 않는다. 일부 예시들에서 생물학적 샘플에는 소변 또는 대변이 포함되지만 이에 제한되지는 않는다. 생물학적 샘플에는 타액이 포함되지만 이에 제한되지 않는다. 생물학적 샘플에는 조직 절개 및 조직 생검이 포함되지만 이에 제한되지는 않는다. 생물학적 샘플에는 앞서 언급한 생물학적 샘플의 파생물 또는 분획이 포함되지만 이에 제한되지는 않는다.
- [0029] 본원에 사용된 용어 "글리칸(glycan)"은 글리코펩티드, 글리코프로틴, 당지질 또는 프로테오글리칸의 탄수화물 부분과 같은 당접합체(glycoconjugate)의 탄수화물 잔기(carbohydrate residue)를 지칭한다.
- [0030] 본원에 사용된 용어 "글리코펩티드(glycopeptide)"는 그에 결합된 적어도 하나의 글리칸 잔기를 갖는 펩티드를 지칭한다.
- [0031] 본원에 사용된 "글리코실화된 펩티드(glycosylated peptide)"라는 문구는 글리칸 잔기에 결합된 펩티드를 지칭한다.
- [0032] 본 명세서에 사용된 바와 같이, "글리코펩티드 단편" 또는 "글리코실화된 펩티드 단편"이라는 문구는 이는 글리코실화된 펩티드가, 예를 들어 하나 이상의 프로테아제(들)로 분해 또는 단편화, 예를 들어 MRM-MS 기기 내 이온 단편화에 의해 수득되는 글리코실화된 단백질의 아미노산 시퀀스의 일부(전부는 아님)와 동일한 아미노산 시퀀스를 갖는 글리코실화된 펩티드(또는 글리코펩티드)를 지칭한다. MRM은 다중 반응 모니터링을 나타낸다.
- [0033] 본원에서 사용된 "글리코펩티드 단편화"라는 문구는 MRM-MS 기기에서 발생하는 이온 단편화 과정을 의미한다.

단편화는 동일한 질량을 갖지만 그들의 전하에 따라 다양한 단편들을 생성할 수 있다.

- [0034] 본 명세서에 사용된 바와 같이, "다중 반응 모니터링 질량 분석법(MRM-MS)"이라는 문구는 생물학적 샘플에서 글리칸 및 펩티드의 표적 정량화를 위한 매우 민감하고 선택적인 방법을 의미한다. 기존의 질량 분석법과 달리, MRM-MS는 고도로 선택성(표적화)이므로, 연구자가 특정 펩타이드 단편을 구체적으로 찾기 위해 기기를 미세 조정할 수 있다. MRM을 사용하면 잠재적인 바이오마커와 같은 관심 펩티드 단편의 감도, 특이성, 속도 및 정량을 높일 수 있다. MRM-MS는 삼중 사중극자(triple quadrupole)(QQQ) 질량 분석기 및 사중극자 비행 시간(qTOF) 질량 분석기 중 하나 이상을 사용한다.
- [0035] 본원에 사용된 "펩티드"는 달리 언급되지 않는 한 글리코펩티드를 포함하는 것을 의미한다.
- [0036] "머무름 시간(RT)"은 용질이 크로마토그래피 컬럼(column)을 통과하는 데 걸리는 시간을 의미하는 것으로 이해된다. 이는 주입에서 검출까지의 시간으로 계산된다. RT는 가스 유량; 오븐과 컬럼의 온도차; 컬럼 열화; 열 길이; 컬럼 유형 및 캐리어 가스를 포함하되 이에 제한되지 않는 여러 요인에 따라 달라진다.
- [0037] 질량 분석(MS) 데이터 또는 데이터세트는 전환 수의 함수로 전환의 강도를 표시하는 스펙트럼 모음을 의미하는 것으로 이해된다. 주어진 전환 값에 대한 MS 데이터는 머무름 시간(RT)인, 시간 기간(t)에 대한 질량(m) 대 전하(z) 강도 레벨(m/z)이다.
- [0038] "존재비(abundance)"는 질량 분석 데이터에 표시된 일부 생물학적 양의 계산된 양을 의미하는 것으로 이해된다. 존재비는 시작 및 중지 피크 경계 사이의 이산 시점들에서 값들을 합산하여 계산될 수 있다. 존재비는 시작 및 중지 피크 경계 사이의 이산 시간 단계들에 대한 통합을 통해 계산될 수도 있다. 통합은 모든 표준 방법으로 수행될 수 있으며, 예를 들어, $\sum(t_{i+1} - t_i) * \left(\frac{1}{2}\right) * [I(t_i) + I(t_{i+1})]$, $i = 1.. n-1$ (n은 경계들 사이의 이산 포인트들의 수, $I(t_i)$ 는 머무름 시간 t_i 에서 질량 대 전하 강도이다)이다. 달리 명시되지 않는 한, 존재비는 다른 MRM 피크에 대한 하나의 통합된 MRM 피크의 상대적 존재비를 나타낸다.
- [0039] 본 발명의 양태를 실행하기 위해 개발될 수 있는 수학적 모델은 비선형이 전역 최소값(global minimum value)을 찾는 데 기반한 최적화된 기능/모델 파라미터 세트가 없다는 점에서 비선형 수학적 모델로 제한되며; 오히려 최적화된 기능/모델 파라미터는 로컬 최소값에만 존재한다(글로벌 최소값은 존재하지 않음). 고려되는 모델 아키텍처에는 인공 신경망, 그래프 모델 및 지원 벡터 머신이 포함되지만 이에 제한되지는 않는다.
- [0040] "기계 학습 모델" 또는 "학습 모델"은 적어도 하나의 비선형 활성화 계층(하나 이상의 아핀 변환(affine transformation)에 추가하여)을 갖는 임의의 다양한 수학적 모델을 의미하는 것으로 이해된다. 또한 기계 학습 모델은 모델 자체와 별개인 하나 이상의 손실 함수(loss function)를 최소화(예를 들어, 교차 엔트로피 손실 또는 음의 로그 가능성 최소화)하여 훈련되거나 최적화된다. 훈련 또는 최적화 프로세스는 모델을 최적화하여 알려진 결과(낮은 바이어스)를 재현하고 모델이 보이지 않는 경험(낮은 분산)에서 정확한 예측을 할 수 있도록 한다. 모델의 출력은 작업과 관련된 다양한 항목, 예를 들어, 확률 또는 통계, 분류, 시퀀스 등이 될 수 있다. 한 특정 예에서, 모델을 훈련하고 최적화하는 데 사용되는 데이터는 목표 결과가 알려진 시계열(time-series) 데이터와 같은 시퀀스 데이터이며, 알려진 목표 출력을 갖는 시퀀스 데이터는 라벨링되거나 주석이 달린 시퀀스 데이터를 의미한다.
- [0041] 라벨링된 시퀀스 데이터를 기반으로 예측을 수행하는 데 적합한 잘 알려진 모델의 예로는 장기 단기 메모리(LSTM)유닛 또는 또는 GRU(gated recurrent unit) 및 선택적으로 주의 메커니즘이 있는 단방향 또는 양방향 RNN(recurrent neural network)이 있다. 주의 메커니즘(또는 주의 계층)은 과거 상태로부터 축적된 지식을 나타내는 양방향 RNN과 함께 시퀀스의 최종 LSTM 단위 또는 GRU, 또는 LSTM 유닛 또는 GRU의 조합에서 가능할 수 있는 것보다 미래 상태 예측에 영향을 미치기 위해 과거 상태 변수로부터 더 중요한 정보를 보유할 수 있다. 주의 메커니즘은 학습된 가중치를 각 입력 상태에 적용하여 출력 예측을 생성할 때 시퀀스에 나타나는 입력 MS 데이터 포인트의 덜 또는 더 중요한 특징을 강조한다. 이러한 학습된 가중치는 각각의 대응하는 출력 상태 또는 원하는 출력에 따라 최종 출력 상태에 더 또는 덜 직접적으로 연결된다.
- [0042] 라벨링된 데이터만 사용하여 훈련되고 최적화된 모델을 지도 모델(supervised model)이라고 한다. 학습 모델은 비지도(unsupervised) 또는 반지도(semi-supervised) 모델일 수도 있다. 이러한 모델은 라벨링되지 않은 데이터 또는 라벨링된 데이터와 라벨링되지 않은 데이터의 조합을 사용하여 훈련되고 최적화된다. 라벨링된 데이터와 라벨링되지 않은 데이터의 조합 또는 주로 라벨링되지 않은 데이터를 사용하여 훈련되고 최적화된 모델들을 각

각 반지도 및 비지도 모델이라고 한다. 학습 모델은 확률적 경사 하강법(SGD)의 잘 알려진 접근 방식에 의해 효율적으로 훈련되고 최적화될 수 있다. 학습 모델의 하이퍼파라미터는 훈련 및/또는 검증 동안 수동 조정으로 최적화될 수 있으며 또는 학습 모델의 전역 손실 함수의 최소화와 병행하여 하나 이상의 하이퍼파라미터에 대한 대리 손실 함수(surrogate loss function)를 최소화하는 방식으로 학습된다.

- [0043] "훈련 세트(training set)"는 알려진 목표 출력을 재현하기 위해 학습 모델에서 파라미터를 훈련하거나 조정하는 데 사용되는 데이터 세트이다.
- [0044] "검증 세트(validation set)"는 라벨링된 데이터 매칭(낮은 바이어스)과 보이지 않는 전환(낮은 분산)에 대한 시작/중지 위치들 예측 사이의 최적 균형에 도달하기 위해 새로운 경험에 대해 모델의 예측 기능을 조정하는 데 사용되는 데이터세트이다. 검증 세트는 훈련 세트를 기반으로 모델이 조정된 후에 적용된다.
- [0045] "테스트 세트"(또는 검증 세트 제외)는 최적화된 모델의 예측 기능을 평가하는 데 사용되는 데이터 세트이다. 테스트 세트에 대한 모델의 성능은 보이지 않는 새로운 전환을 정확하게 예측하는 모델의 능력을 평가하는 데 사용되는 메트릭(metric)이다.
- [0046] "특징화(featurization)"는 기계 학습 모델에 대한 입력으로 사용하기 위해 분포로부터의 샘플들을 나타내는 원시 데이터(raw data)를 변환하는 것과 관련된 프로세스를 의미하는 것으로 이해된다. 특징화의 예에는 아날로그 신호를 디지털 신호로 이산화, 데이터의 일부를 필터링 또는 노이즈 제거 또는 클리핑(예를 들어, 신호 대 잡음비 감소), 신호 정규화 및 신호 표준화(예를 들어, 윈도우 내에서 목표 피크를 중앙에 배치)가 포함된다. 특징화는 일반화를 위해 모델을 검증하거나 샘플들의 관련 증가(augmentation)를 검증하고 및/또는 새로운 (또는 보이지 않는) 경험으로부터 예측하는 모델의 능력을 개선하기 위한 목적으로 데이터 세트에 노이즈를 도입하는 것도 포함할 수 있다.
- [0047] 본 개시에 따르면 훈련, 검증 및 테스트 데이터세트가 샘플링된 시퀀스 데이터의 데이터세트는 전환을 위해 적어도 라벨링된 시작 및 중지 위치들을 포함할 수 있으며, 각 위치는 경계 포인트들 사이의 적분(integration)에 의해 존재비를 계산하기 위한 경계 포인트들을 나타낸다. 바람직한 실시 예에서 목표 출력은 전환의 시퀀스 데이터의 각 포인트에 대한 시작 및 중지 위치의 확률이다.
- [0048] 논의
- [0049] 앞서 언급 한 바와 같이, MRM 실험의 기존 분석에는 크로마토그래피 피크의 시작과 종지를 손으로 선택하고 이전에 데이터 종속 수집 또는 기타 발견 기술로 특징지어졌던 전환에서 작업하는 것이 포함된다. 애질런트(Agilent)의 메스헌터(MassHunter) 소프트웨어와 같은 소프트웨어 제품군은 주어진 전구체 및 제품 질량 대 전하 비율에 대해 사전 지정된 머무름 시간(RT) 범위에 걸쳐 신호 강도를 표시한다. 피크의 시작과 끝을 수동으로 선택하면 관찰된 RT, 피크 폭 및 통합된 존재비 값이 산출된다. 그러나 평가의 개인 간 변동성, 인적 오류 및 많은 시간 투자로 인해 이 방법은 연구 이외의 용도로는 부적절하다.
- [0050] 명확하게 정의된 단일 피크와 높은 신호 대 잡음비를 특징으로 하는 전환의 경우, 존재비를 계산하기 위해 경계 포인트를 선택하는 위의 수동 프로세스가 효과적일 수 있지만, 그러나 이 방법은 낮은 신호 대 잡음비를 특징으로 하는 생물학적 양의 존재비를 검출할 때 정확하지 않거나 일관된 결과를 제공하지 않을 수 있다.
- [0051] 도 1을 참조하면, 본 개시에 따르면 훈련된 기계 학습 모델은 대신 지정된 머무름 시간 윈도우 및 질량/전하 강도에 대한 질량 스펙트럼의 경계 포인트를 자동으로 선택하는 데 사용된다. 기계 학습 모델에 대한 입력은 포인트들의 시퀀스(MS 데이터의 샘플)이다. 각 포인트는 대응되는 머무름 시간에 대한 질량/전하 강도를 나타낸다.
- [0052] 도 2를 참조하면, 일 구현에 따르면 기계 학습 모델은 입력으로서 제공된 대응하는 포인트들의 각각에 대한 피크 시작 및 피크 중지 시간의 확률을 생성하는 신경망이다. 신경망은 라벨링된 목표 피크 시작 및 중지 머무름 시간을 갖는 질량 스펙트럼들을 사용하여 훈련된다. 훈련 및 검증 프로세스를 통해 최적의 모델 파라미터가 발견된다. 라벨링된 데이터는 그래픽 사용자 인터페이스 및 입력 디바이스를 사용하여 피크 시작/중지 시간을 선택하는 인간 주석가로부터 얻을 수 있다.
- [0053] 도 3을 참조하면 도 1의 기계 학습 모델에 대한 입력을 위한 원시 질량 스펙트럼 데이터를 특징화하기 위한 프로세스 흐름이 도시된다. 특징화 단계는 (1) 특정 임계값을 초과하는 질량/전하 강도를 제거하는 단계, 또는 대안적으로, MS 데이터의 신호 대 잡음 전환 영역 근처에 있고 대부분 잡음인 포인트들만 선택하는 단계; 및 (2) 시퀀스 데이터를 생성하기 위해 아날로그 신호를 포인트들의 세트로 이산화하는 단계(discretizing)를 포함한다.

- [0054] 도 4를 참조하면, 하나의 특정 구현에서 모델은 적분(integrals)의 가중 합에 기초하여 존재량(abundance amount)을 계산한다. 가장 직접적인 방법 중 하나는 기계 학습 모델의 예측 결과를 확률 분포로 해석하는 것이다. 다시 말해, 모델이 피크 시작/종료 마커 역할을 할 수 있는 시퀀스의 가능한 모든 포인트 쌍의 확률을 평가하도록 하는 것이다. N 포인트들의 시퀀스의 경우, 이는 순서 제약(order constraint)을 부과한 후 전체 $N \times (N-1)/2$ 쌍들을 포함한다. 각 쌍의 확률은 두 포인트 상에서 피크 시작/종료 작업의 예측 값들을 곱하여 쉽게 계산될 수 있다. 그런 다음 확률을 사용하여 대응하는 쌍들 사이의 적분에 대한 가중 합계를 계산하여 실제 경계가 될 가능성이 가장 높은 두 포인트들로부터 계산된 단일 존재비 값 대신 평균 추정치(mean estimate)를 제공한다. 존재비의 평균은 더 강력한 존재비 추정치를 제공할 수 있다.
- [0055] 실시 예
- [0056] 일부 예시에서, 질량 분석 피크들을 식별하기 위해 모델을 훈련하는 방법이 본원에 설명되며, 상기 방법은: 복수의 라벨링된 시퀀스 데이터를 사용하는 단계-상기 복수의 라벨링된 시퀀스 데이터의 각각은 글리코펩티드 또는 펩티드, 또는 이의 단편들의 특징적인 머무름 시간 값들을 나타내는 피크 시작 및 피크 중지 라벨을 포함 함 -, 및 상기 라벨링된 시퀀스 데이터를 사용하여 적어도 하나의 기계 학습 모델을 훈련시키는 단계를 포함하고, 여기서 상기 훈련된 모델은 생물학적 샘플에 존재하는 하나 이상의 글리코펩티드 또는 펩티드, 또는 이의 단편의 존재를 나타내는 질량 분석 시작 및 중지 피크들을 식별하도록 적응된다.
- [0057] 전술한 것을 포함하는 일부 예에서, 상기 방법은 상기 생물학적 샘플로부터의 상기 질량 분석 시작 및 중지 피크들 사이의 시간 경과에 따른 통합에 의해 상기 하나 이상의 글리코펩티드 또는 펩티드의 존재를 정량화하는 단계를 더 포함한다.
- [0058] 전술한 것을 포함하는 일부 예에서, 상기 기계 학습 모델은 순환 신경망 모델이다.
- [0059] 전술한 것을 포함하는 일부 예에서, 상기 순환 신경망은 관심 계층을 갖는 양방향 LSTM 순환 신경망 모델이다.
- [0060] 일부 예에서, 본원에 설명된 프로세스는 기계 학습 모델을 사용하여 글리코펩티드 또는 펩티드에 대한 존재비를 계산하는 것을 포함하는 방법이 본원에 설명되어 있다.
- [0061] 일부 예에서, 시스템이 본원에 설명되어 있으며, 상기 시스템은: 라벨링된 시퀀스 데이터의 데이터 저장소; 상기 저장소에 결합된 하나 이상의 프로세서들; 상기 하나 이상의 프로세서들에 의해 실행될 때 본원에 명시된 방법을 수행하는, 상기 하나 이상의 프로세서들에 액세스 가능한 비일시적 메모리에 상주하는, 기계 실행 가능 코드를 포함한다.
- [0062] 일부 예에서, 질량 분석 피크들을 식별하는 방법이 본원에 설명되어 있으며, 상기 방법은: 입력 파라미터들의 함수로서 출력을 생성하는 적어도 하나의 훈련된 신경망 모델을 제공하는 단계-여기서 상기 입력 파라미터들에 대한 값들은 m/z (질량 대 전하) 비율을 갖는 글리코펩티드에 대한 머무름 시간 값들의 시퀀스를 포함 함-; 및 생물학적 샘플의 질량 분석법으로부터의 상기 머무름 시간 값들의 시퀀스에 대한 질량 대 전하(m/z) 비율들을 포함하는 질량 분석(MS) 데이터를 상기 입력 파라미터들에 대한 값들로 제공하는 단계를 포함하고; 상기 출력은 상기 생물학적 샘플에서 하나 이상의 분석물 또는 이의 단편의 존재를 나타내는 하나 이상의 피크 시작 및 피크 중지 머무름 시간 값들에 대한 위치를 포함한다.
- [0063] 전술한 것을 포함하는 일부 예에서, 상기 방법은 상기 생물학적 샘플로부터의 상기 질량 분석 시작 및 중지 피크들 사이의 머무름 시간 값들을 합산하거나 통합함으로써 상기 하나 이상의 글리코펩티드 또는 펩티드의 존재를 정량화하는 단계를 더 포함한다.
- [0064] 전술한 것을 포함하는 일부 예에서, 상기 방법은 생물학적 샘플로부터 질량 분석 시작 및 중지 피크 사이의 머무름 시간 값들을 합산함으로써 하나 이상의 글리코펩티드 또는 펩티드의 존재를 정량화하는 단계를 더 포함한다. 전술한 것 중 임의의 것을 포함하는 일부 예에서, 상기 방법은 생물학적 샘플로부터의 질량 분석 시작 및 중지 피크 사이의 머무름 시간 값에 대한 통합에 의해 하나 이상의 글리코펩티드 또는 펩티드의 존재를 정량화하는 단계를 더 포함한다.
- [0065] 전술한 것을 포함하는 일부 예에서, 상기 정량화 단계는 상기 예측된 시작 및 중지 피크들에 대해 가중 평활 함수를 적용하는 단계를 포함한다.
- [0066] 일부 예에서 시스템이 본원에 설명되어 있으며, 상기 시스템은: 질량 분석(MS) 데이터의 데이터 저장소; 상기 저장소에 결합된 하나 이상의 프로세서들; 상기 하나 이상의 프로세서에 의해 실행될 때 본원에 명시된 방법을 수행하는, 상기 하나 이상의 프로세서들에 액세스 가능한 비일시적 메모리에 상주하는, 기계 실행 가능 코드를

포함한다.

- [0067] 일부 예에서 방법이 본원에 설명되어 있으며, 상기 방법은: 원시 MS 데이터로부터의 데이터를 특징화하는 단계를 포함하고, 상기 원시 MS 데이터는, 각각이 질량/전하 강도를 나타내는, 복수의 아날로그 샘플들을 포함하고, 상기 특징화 단계는, 상기 아날로그 샘플들의 각각에 대해, 머무름 시간 윈도우 내에서 상기 샘플을 중앙에 배치하는 단계, 상기 샘플을 강도 값들을 나타내는 포인트들의 시퀀스로 이산화하고, 상기 강도 값들을 표준화하는 단계, 및 라벨링된 포인트들의 시퀀스를 생성하기 위해 상기 포인트들의 시퀀스 중 피크 시작 및 피크 중지 시간들에 대응하는 포인트들에 라벨들을 할당하는 단계를 포함하고; 상기 라벨링된 포인트들의 시퀀스는 보이지 않는 MS 데이터의 존재비를 예측하기 위해 기계 학습 모델을 훈련하도록 구성된다.
- [0068] 전술한 것을 포함하는 일부 예에서, 상기 중앙에 배치하는 단계는 상기 머무름 윈도우의 중심에 곡선의 정점을 위치시키는 단계를 포함한다.
- [0069] 전술한 것을 포함하는 일부 예에서, 상기 강도 값들을 표준화하는 단계는 임계값보다 큰 강도 값을 제외한다. 전술한 것을 포함하는 일부 예에서, 상기 임계 값이 약 500이다.
- [0070] 일부 예에서 시스템이 본원에 설명되어 있으며, 상기 시스템은 복수의 아날로그 샘플들을 포함하는 원시 질량 분석(MS) 데이터의 데이터 저장소; 상기 저장소에 결합된 하나 이상의 프로세서들; 상기 하나 이상의 프로세서에 의해 실행될 때 본원에 설명된 방법을 수행하는, 상기 하나 이상의 프로세서들에 액세스 가능한 비일시적 메모리에 상주하는, 기계 실행 가능 코드를 포함한다. 일부 예에서 방법이 본원에 설명되어 있으며, 상기 방법은 주어진 m/z 비율 및 머무름 기간에 대한 질량 분석(MS) 데이터를 나타내는 포인트들의 시퀀스에 대한 확률들의 분포를 계산하는 단계-상기 확률들은 피크 시작 또는 피크 중지 머무름 시간을 나타냄-, 및 상기 확률들의 분포를 사용하여 존재비에 대한 통계를 계산하는 단계를 포함하고, 상기 통계는 상기 MS 데이터에서 생물학적 샘플의 존재를 나타낸다.
- [0071] 전술한 것을 포함하는 일부 예에서, 상기 통계는 적분을 평가하기 위한 시작 및 끝 포인트들의 가능한 모든 쌍들을 사용하여 적분들의 가중 합으로 계산된 평균 존재비이다.
- [0072] 전술한 것을 포함하는 일부 예에서, 상기 기계 학습 모델은 시퀀스 모델(모델 A) 또는 참조 기반 모델(모델 B)이다.
- [0073] 일부 예에서 방법이 본원에 설명되어 있으며, 상기 방법은: 복수의 라벨링된 시퀀스 데이터를 사용하는 단계, 및 상기 라벨링된 시퀀스 데이터를 사용하여 적어도 하나의 기계 학습 모델을 훈련시키는 단계를 포함하고, 상기 훈련된 모델은 보이지 않는 샘플에서 강도의 시작 및 중지 피크가 발생하는 시간 주기 또는 주파수 범위를 식별하도록 적응된다.
- [0074] 전술한 것을 포함하는 일부 예에서, 상기 시작 및 중지 피크는 잡음에서 신호로의 전환 기간에 포함된다.
- [0075] 전술한 것을 포함하는 일부 예에서, 상기 기계 학습 모델은 주의 계층을 갖는 순환 신경망이다.
- [0076] 전술한 것을 포함하는 일부 예에서, 상기 샘플은 상기 강도의 시작 피크에서 중지 피크까지의 시간 또는 주파수에 대한 합산 또는 적분이 물리적인 것(physical thing)을 정량화하도록 상기 물리적인 것을 나타낸다.
- [0077] 전술한 것을 포함하는 일부 예에서, 상기 샘플은 지정된 시간 기간 동안 질량 대 전하 강도들의 시퀀스 시리즈들이고 상기 합산 또는 적분은 상기 물리적인 것의 존재비를 나타낸다.
- [0078] 일부 예에서 방법이 본원에 설명되어 있으며, 상기 방법은: 입력 파라미터들의 함수로서 출력을 생성하는 적어도 하나의 훈련된 신경망 모델을 제공하는 단계-상기 입력 파라미터들에 대한 값들은 양을 나타내는 강도 값들의 시간적 또는 스펙트럼 시퀀스임-; 및 상기 입력 파라미터들의 각각에 대한 값들을 제공하는 단계를 포함하고, 상기 값들은 양에 대해 보이지 않는 강도들의 시퀀스를 나타내고; 상기 출력은 상기 양을 정량화하는데 사용되는 하나 이상의 피크 시작 및 피크 중지 시간들 또는 주파수들을 포함한다.
- [0079] 전술한 것을 포함하는 일부 예에서, 상기 양은 상기 피크 시작 및 상기 피크 중지 시간들 또는 주파수들 사이의 시간 또는 주파수에 대한 합산 또는 적분에 의해 각각 정량화된다.
- [0080] 전술한 것을 포함하는 일부 예에서, 여기서 상기 강도들은 질량 분석법, MRM, 액체 또는 기체 크로마토그래피, X선 회절, 라만 분광법, UV-VIS 분광법, 형광분석 분광법 또는 인광분석 분광법, 핵 자기 공명(NMR) 분광법, 또는 자기 공명 영상(MRI)의 세트로부터 선택된 프로세스로부터 얻어진다.

- [0081] 예시
- [0082] 다음 예시는 본 개시의 여러 양태들에 따라 수행된 연구를 설명한다.
- [0083] 이 연구에서 두 개의 학습 모델들이 훈련되고, 검증된 다음 피크(또는 더 정확하게는, 관련 전환에 대한 존재비를 더 정확하게 계산하기 위해 피크의 시작/중지 경계)의 위치 예측의 정확성을 평가하기 위해 테스트되었다. 학습 모델을 위해 선택된 특정 아키텍처는 장기 단기 메모리 유닛과 주의 계층(RNN-LSTM-AL)이 있는 양방향 순환 신경망이었다. 모델 아키텍처 및 훈련 방법은 도 5a, 5b, 6a, 6b 및 6c에 예시되어 있다. 모델의 입력 파라미터 또는 입력 상태 변수는 질량 대 전하 비율(m/z) 강도를 나타내는 시간 시퀀스에 대응한다.
- [0084] 이 연구는, 펩티드와 손상되지 않은(intact) 글리코펩티드를 모두 포함하여, 광범위한 동적 범위에 걸쳐 있는 질량 분석(MS) 전환으로부터 수동으로 라벨링된 피크의 대규모 데이터 세트에 액세스하였다. 이 데이터 세트에는 각 전환의 머무름 시간(RT) 내에서 강도들의 시간적 시퀀스가 포함되어 있다. 이러한 각 전환은 RNN-LSTM-AL에 대한 시퀀스 데이터 입력이며, 훈련, 검증 및 테스트 전환 각각에 포함된 관련 인간 주석이 달린(또는 라벨링된) 시작 및 중지 RT(존재비를 유도하기 위해 통합을 위한 경계 포인트를 나타냄)는 지도 학습, 검증 및 테스트에 각각 사용되는 라벨이다.
- [0085] **2.1 데이터 세트**
- [0086] 두 개의 질량 분석 데이터세트가 연구에 사용되었다. 제1 실험에 대한 로우 애질런트(Raw Agilent) QQQ MS 출력은 210개의 인간 혈청 샘플들에서 평가된 716개의 전환들로 구성되었으며, 이는 양성 존재비 및 RT 시프트(shift)를 위해 필터링한 후 106355개의 유효한 샘플 피크들을 생성하였다. 이 세트는 572개의 전환들이 있는 훈련 세트와 144개의 피크들이 있는 검증 세트로 분할되었다.
- [0087] 제2 실험은 상업용 생물리포지토리(biorepository)에서 구입한 135개의 인간 혈청 샘플들에서 생성된 MS 출력을 사용하였다. 503개의 전환들이 평가되었으며, 테스트용으로만 사용된 67672개의 고품질 피크들이 생성되었다. 313개의 테스트 세트 전환들이 학습/검증 세트에 존재하지 않았다. 313개의 테스트 세트 전환들을 사용하여 모델 출력 예측들을 비교하였다.
- [0088] **2.1 기선법(Baseline Method)**
- [0089] 훈련되고 검증된 RNN-LSTM-AL 모델의 성능을 두 가지 기준선 예측자(baseline predictor)들과 비교되었다. 이러한 기준선 예측자들 중 첫 번째는 다중 반응 모니터링을 포함한 질량 분석 어플리케이션의 데이터 분석에 사용되는 잘 알려진 오픈 소스 소프트웨어 도구인 스카이라인(Skyline)이다. 스카이라인(Skyline) 소프트웨어는 테스트 샘플들의 피크 영역(존재비)을 계산한다. RNN 모델과의 비교를 위한 두 번째 기준선 예측자는 결정론적(deterministic) 예측 방법이며 이하 "규칙 기반(rule-based)" 방법이라고 한다. 규칙 기반 방법의 아키텍처는 아오시마 켄(Ken Aoshima), 타카하시 켄타로(Kentaro Takahashi), 이카와 마사유키(Masayuki Ikawa), 기무라 타카유키(Takayuki Kimura), 후쿠다 미즈루(Mitsuru Fukuda), 다나카 사토시(Satoshi Tanaka), 하웰 E 패리(Howell E Parry), 후지타 유이치로(Yuichiro Fujita), C 요시자와 아키야스(Akiyasu C Yoshizawa), 우즈노미야 신이치(Shin-ichi Utsunomiya) 외, *크로마토그래피 질량 분석을 위한 간단한 피크 검출 및 라벨 프리 정량 알고리즘*, BMC 생물정보학, 15(1):376, 2014에 설명되어 있다. 이 연구를 위해 모델 파라미터를 조정하여 학습 및 검증 세트 성능을 최대화하여 규칙 기반 방법 모델이 개발되었다. 그리고 이러한 조정된 파라미터는 테스트 세트 평가 중에 적용되었다.
- [0090] **2.3 피크 통합을 위한 시퀀스 모델**
- [0091] RNN-LSTM-AL 아키텍처는 연구에서 MS 데이터 시퀀스에서 예측을 수행하기 위해 선택되었다.
- [0092] 연구에 사용된 두 개의 RNN-LSTM-AL 모델들 중 첫 번째 모델(이하 "모델 A" 또는 "시퀀스 모델")에 대한 훈련/예측 프로세스 및 흐름에 대한 시각적 및 블록 설명이 도 5a 및 도 5b에 도시되어 있다. 입력(시간에 따른 질량/전하 비율 강도)은 별도의 위치들을 연결하기 위한 주의 메커니즘과 바이(bi)-LSTM 계층을 통해 인코딩된다. 예측은 포인트별로 수행되며, 즉, 모든 입력에 대해 모델이 출력을 예측하는 것을 의미한다. 모델은 $N \times 2$ 차원의 출력 시퀀스를 생성하도록 훈련되었으며, 여기서 "N"은 모델 A에 대한 입력 파라미터들의 수(강도 값이 모델에 입력된 머무름 시간에 해당하는 포인트들, 샘플의 곡선 특성에 따라 50 내지 300 범위, 아래 섹션 5.3 참조), 시퀀스의 데이터 포인트들의 수 및 각 입력 값에 해당하는 출력 값들이다. N개의 포인트들의 각각에 대해 출력된 두 (2) 값들은 해당 포인트가 피크 시작 또는 피크 중지를 나타낼 확률(0에서 1 사이)이다. 주어진 전환에 대해 1에 가장 가까운 값(높은 확률)을 갖는 모델로부터 출력된 포인트 또는 이웃하는 포인트들은 존재비 적분

(시간이 지남에 따라 통합된)이 계산되는 간격에 대한 경계를 정의하는 데 사용된다.

[0093] 따라서 모델로부터의 출력 확률은 경계들의 분포를 형성한다. 모델은 예측된 경계와 실측 자료 사이의, 또는 수동으로 라벨링된 피크 시작 및 피크 중지 사이의 교차 엔트로피 손실에 대한 기울기 하강법(gradient descent)으로 종단 간(end-to-end) 학습된다. 최소 하이퍼파라미터 검색, 예를 들어, 모델의 하이퍼파라미터에 대해 최소화된 목적 함수(objective function)에 적용된 베이저안 최적화(Bayesian optimization)는 양호한 일반화를 달성하고 훈련 세트에 대한 과적합(overfitting)을 피하기 위해 생략 검증(leave-out validation) 또는 테스트 세트에 대한 성능을 최적화하기 위해 적용된다. 특성화 및 하이퍼 파라미터는 아래에서 더 자세히 설명된다.

[0094] 2.4 참조 기반 시퀀스 모델

[0095] 일부 어플리케이션에서는, 전환 피크의 형상과 폭이 매우 다양하기 때문에, 모델 A와 같은 통합된 예측 모델이 보이지 않는 전환으로 일반화되지 않을 수 있다. 동일한 전환에 있는 다른 환자(patient)들의 샘플이 형상이 일관된다는 관찰에 기초하여, 두 번째 RNN-LSTM-AL 모델(이하 "모델 B" 또는 "참조 기반 모델")도 개발되었다.

[0096] 모델 B에 대한 훈련/예측 프로세스 및 흐름의 시각적 및 블록 설명은 도 6a, 6b 및 6c에 예시되어 있다. 모델 B는 쿼리 샘플(query sample)과 참조를 각각 입력으로 취한다. 둘 모두에 대한 인코딩된 특징은 샘플 참조 주의 계층에서 비교 및 병합되어, 예측을 생성한다. 쿼리 피크 특징, 참조 피크 특징 및 참조 라벨이 제공되는 주의 계층은 쿼리와 참조를 포인트별로(point by point) 비교하여, 참조에서 라벨링된 시작/종료 포인트들과 매우 유사한 쿼리의 포인트들이 더 높은 예측 확률을 얻을 수 있다. 두 개의 일관된 샘플이 주어지면, 이 작업의 이상적인 주의 맵은 동일한 환경의 포인트들이 매우 유사한 것으로 간주되는 "항등 행렬(identity matrix)"이 된다(도 6c 참조). 주의 맵과 머무름 시간 매핑 매트릭스 사이의 발산 용어(divergence term)는 우수한 정규화를 달성하기 위해 최적화 프로세스에 포함되었다.

[0097] 참조 기반 예측(쿼리 기반 예측과 반대)의 설정에서, 작업은 더 이상 모델 A에서와 같이 잡음 신호 전환 포인트를 식별하려고 하지 않고, 오히려 유사성 매핑을 최적화하기 위해 시퀀스 및 그의 컨텍스트의 포인트에 대한 우수한 인코딩 메커니즘이라는 점에 주의한다. 어플리케이션에서, 참조 기반 모델은 퓨샷(few-shot) 학습 모델과 유사하게 작동한다. 이들은, 쿼리 샘플의 예측에 참조 라벨을 정확하게 매핑하는 것을 목표로, 쿼리/참조의 쌍에 대해 훈련된다. 이렇게 하면 올바르게 라벨링된 단일(또는 몇 개) 참조가 제공되는 한 자연스럽게 모델이 보이지 않는 전환 피크에서 더 잘 수행할 수 있다.

[0098] 모델 B에 대해 채택된 구조는 몇 가지 고품질 참조 피크들의 수동 주석이 합리적인 전체 조건인 동일한 전환 세트에 대한 고처리량 실험에 가장 적합하고 비용 효율적일 수 있다. 대부분의 발견 또는 소규모 실험에서 독립형 참조-프리 예측자가 선호될 수 있다.

[0099] 도 7a 및 도 7b를 참조하면, 각 도면은 두 개의 그래프들을 갖는다. 첫 번째 그래프는 각각 별과 삼각형으로 표시된 시작 및 중지 시간이 라벨링된 전환 그래프이다. 하단 그래프는 훈련되고 검증된 모델에 의해 예측된 시작/중지 시간의 확률 그래프이다.

[0100] 3.1 경계 및 존재비 예측 성능

[0101] 모델을 평가하기 위해 제외 또는 테스트 세트 샘플이 사용되었다. 해결 중인 문제에 대한 포괄적인 관점을 제시하기 위해 여러 메트릭들이 사용되었다. 예측은 두 가지 기준으로 평가된다: (1) 경계(피크 시작/종류) 예측의 정확성 및 (2) 존재비(피크 영역) 예측의 정확성. 전자는 문제의 설정과 직접적으로 관련되어 있지만, 후자는 분석을 위해 분석물 존재비를 활용하는 다운스트림 어플리케이션에 훨씬 더 중요하다.

[0102] 예측에서, 두 작업들, 즉 피크 시작 시간과 피크 종료 시간을 찾는 작업들에 대해 예측 확률이 최대인 곡선의 포인트가 경계에 대한 마커로 사용되며, 이는 다음 텍스트에서 "argmax 경계"라고 한다. 경계 예측이 실측 자료 주석(사람이 라벨링한 포인트들) 주변의 1.2초(시퀀스의 2 포인트들)의 오류 임계값 내에 있는 샘플의 비율로 정의된 정확도 점수와 함께 argmax 경계와 사람의 주석 또는 라벨 사이의 평균 절대 오차(MAE)가 계산되었다. 피크 지속 시간이 테스트 세트에서 20 +/- 8초라는 점을 감안할 때(시퀀스의 33 +/- 14 포인트들), 오류 임계값은 윈도우 내의 임의의 경계 예측이 존재비를 크게 변경하지 않을 만큼 충분히 작다.

[0103] 존재비의 평가를 위해, 먼저 각각 실측 자료(사람이 라벨링한) 및 모델 예측으로부터의 경계들 사이의 기준선 조정 적분(baseline-adjusted integral)이 계산되고 각 개별 전환에서 둘 사이의 스피어만 랭크 상관 계수(Spearman's r)와 피어슨의 상관 계수(Pearson's r)가 평가된다. 존재비의 넓은 범위로 인해($1 \sim 10^{15}$), 피어슨의 상관 계수는 많은 존재비의 샘플들에 의해 지배될 것이며, 여기에 보고된 것은 바이어스를 피하기 위해 대신

에 로그(log)-존재비에 대해 계산된 상관 관계이다. 모든 전환들에 대한 평균 점수가 보고된다. 또한, 실제 사용의 표시를 위해, 예측된 존재비가 주석이 달린 존재비의 $\pm 5\%$ 이내인 샘플들의 비율로 정의된 정확도 점수가 또한 보고된다. 이 정확도 메트릭에는 한 가지 주의 사항(caveat)이 있다: 시작/중지 경계 예측 오류들이 서로를 상쇄할 수 있기 때문에 주석이 달린 경계에서 벗어나는 경계 예측에 의해 정의된 간격에 대해 계산된 적분은 주석이 있는 존재비와 매우 유사한 존재비를 갖는 것으로 나타날 수 있다. 즉, 예측 피크 시작 시간의 % 오차는 강도 값들을 적분으로 합산할 때 예측 피크 중지 시간의 대응되는 % 오차를 상쇄할 수 있다.

표 1

[0104]

생략 테스트 세트에 대한 예측 성능 (67672 피크들)					
	모델 경계 예측		존재비 예측		
	MAE (초)	정확도*	로그-존재비 피어슨의 상관 계수	스피어만의 상관 계수	정확도**
스카이라인(Skyline)	N/A	N/A	0.6388	0.7372	0.4365
규칙 기반	4.79	0.285	0.9879	0.9708	0.5847
모델 A	2.33	0.432	0.9897	0.9835	0.6889
모델 B	1.56	0.584	0.9974	0.9917	0.7995

표 2

[0105]

훈련 세트에 대한 예측 성능 (86311 샘플들)					
모델	경계 예측		존재비 예측		
	MAE (초)	정확도*	로그-Abd 피어슨의 상관 계수	스피어만의 Rho	정확도**
규칙 기반	3.97	0.311	0.9097	0.9924	0.645
모델 A	1.79	0.601	0.9477	0.9965	0.743
모델 B	1.54	0.667	0.9592	0.9969	0.779

[0106]

* 경계 예측이 1.2초 오차 이내인 샘플의 비율

[0107]

** 존재비 예측이 5% 오차 이내인 샘플의 비율

표 3

[0108]

검증 세트에 대한 예측 성능 (20044 샘플들)					
모델	경계 예측		존재비 예측		
	MAE(초)	정확도	로그-Abd 피어슨의 상관 계수	스피어만의 Rho	정확도
규칙 기반	4.08	0.320	0.9141	0.9023	0.6593
모델 A	2.57	0.445	0.9332	0.9270	0.6884
모델 B	2.80	0.372	0.9541	0.9398	0.6563

[0109]

표 1은 세 가지 테스트 방법의 성능 점수와 스카이라인(Skyline) 예측을 나타낸다. 스카이라인(Skyline)은 존재비 예측만을 생성하였으며 경계 예측 작업에서는 평가 및 비교되지 않았다. 모델 A와 모델 B는 모두 규칙 기반 방법과 스카이라인(Skyline) 예측에 비해 모든 작업에서 상당한 성능 향상을 보여주었다. 모델 B는 추가 참조 데이터의 도움으로 최고 점수를 받았다. 경계 예측에서, 규칙 기반 방법이 훈련/검증 세트 성능을 최적화하도록 조정되기 때문에(표 3 참조), 강한 과적합이 관찰된다. 이에 비해, 모델 A와 B는, 더 나은 샘플 품질 덕분에, 테스트 세트에서 각각 2.33 초와 1.56 초의 MAE를 달성했으며(참조 유무에 관계없이), 이는 규칙 기반 방법보다 2배 더 작다. 마찬가지로, 정확도 점수 성능에서 모델 B는 규칙 기반 방법보다 30% 우위에 있었다. 도 8a 및 도 8b는 각각 막대 그래프 및 경계 예측 오차의 비정규화된 누적 분포를 나타내며, 여기서 모델 A 및 B는 더 낮은 오차뿐만 아니라 더 적은 이상점을 갖는 분포를 나타낸다.

- [0110] 존재비 예측은 유사한 결과를 나타냈다: 모델 A 및 B는 피크 존재비 평가에서 더 우수하고 강력한 성능을 달성했으며(도 8c), 정확도가 20% 이상 증가했다. 이 문제에서 경계 포인트가 일반적으로 전체 적분에 거의 기여하지 않기 때문에 경계 예측의 오류로 인한 영향이 크게 완화된다. 상관 점수는 모델 A와 B가, 0.9975와 0.9990에서 거의 완벽한 스피어만의 상관 계수와 함께, 현재의 기성품 질량 분석법 데이터 분석 소프트웨어보다 훨씬 더 나은 성능을 갖는다는 것을 나타낸다. 흥미롭게도, 규칙 기반 방법은 스카이라인(Skyline) 예측보다 훨씬 더 높은 상관 관계 점수를 달성했다. 어떤 이론에도 얽매이지 않고, 이 결과는 테스트 샘플의 복잡한 구성으로 인한 것일 수 있으며, 이는 머무름 시간이 가까운 동시 용리(co-eluting) 피크 또는 화합물을 생성하였다. RT 윈도우가 원하는 전환 주위에 가깝게 설정되면(이는 인간 주석자가 일반적으로 사용하는 설정이기도 함), 스카이라인(Skyline)이 실수로 다른 범위에서 피크를 선택하여 성능이 크게 저하될 수 있다.
- [0111] 모델 A와 모델 B를 비교하면, 모델 B에 대한 교육 및 검증 단계에서 추가 "참조 데이터"가 포함되었기 때문에, 이 모델은 모델 A보다 성능이 우수하여, 경계 및 존재비 예측의 정확도가 15% 및 10% 증가했다. 모델 B는 더 적은 이상점으로 더 나은 경계 오차 분포를 나타냈으며(2.2% 대 7.1%, 도 8b), 이는 존재비 상관 관계 그래프의 관찰 결과와 일치한다: 녹색 포인트들은 중심선 주위에 더 집중되어 있는 반면 파란색 포인트들은 몇몇 외부 예측(outlying prediction)을 포함한다(도 8c).
- [0112] 그러나 어플리케이션에 따라 더 적은 훈련 세트가 필요하기 때문에 모델 A가 선호될 수 있다. 모델 A는 결과가 보여주듯이 좋은 정확도/예측된 풍부함을 제공할 수 있기 때문에 모델 B와 함께 사용되는 추가 참조 데이터는 필요하지 않거나 소규모 어플리케이션이나 실험에 실용적일 수 있다. 모델 B는 더 큰 규모의 실험이나 더 많은 정확도가 필요한 경우 선호될 수 있다.
- [0113] 이러한 실험에서, 모델 B는 그래프에 표시되지 않은 또 다른 이점을 제공했다: 여러 샘플에서 동일한 전환들에 대한 예측의 더 높은 일관성. 특히 경계가 숄더(shoulder) 및 밸리(valley) 포인트로 인해 달라질 수 있는 전환의 경우, 모델 B는 그의 참조의 동일한 예측 형식을 따르는 경향이 있으며(도 9a), 이것이 모델 구조 설계의 주요 목적이다. 이는 매우 동일한 표준을 기반으로 존재비가 생성되어 보다 정확한 비교 및 정량화를 허용한다는 점에서 높은 처리량 실험에서 매우 바람직하다. 그러나 모델의 강력한 사전 가정(prior assumption)은 머무름 시간 이동 또는 강력한 배치 효과의 경우에도 성능을 저하시킬 수 있다.
- [0114] **3.2 인간 주석자와의 비교**
- [0115] 신경망 기반 모델의 실제 사용을 추가로 검증하기 위해, 모델 예측을 피크 존재비를 계산하는 인간 주석자 그룹과 비교했다.
- [0116] 테스트 세트 내에서, 135개의 모든 혈청 샘플들에 대한 12개의 전환으로 구성된 서브세트가 분석되었으며, 12명의 주석자들이 피크 시작/종료를 독립적으로 표시했다. 원래 테스트 세트 라벨과 결합하여, 이 서브세트에 대해 13개의 인간 주석 세트가 준비된다. 그런 다음 우리는 13개의 주석들에서 각 샘플 피크에 대한 상대 표준 편차를 서로 다른 주석자들 간의 변동에 대한 프록시(proxy)로 계산하였고, 모델 A 및 모델 B의 예측 상대 오차와 비교했다.
- [0117] 결과는 도 10에 요약되어 있다. 도 10의 회색 막대에 의해 알 수 있듯이, 대부분의 샘플들에서 주석자들에 대한 높은 일관성이 관찰된다. 인간 주석자들의 모든 샘플 피크들에 대한 평균 RSD는 2.5%이다. 동시에 모델 A와 모델 B는 서브세트에서 우수한 성능을 보여, 인간 주석자 평균과 비교하여 각각 3.5% 및 2.1%의 평균 상대 오차를 달성했다. 파란색 및 녹색 막대들에 표시된 대로 오류 분포는 유사하지만 주석자들 전체에서 RSD보다 약간 나쁘다. 윌콕슨 부호위수(Wilcoxon signed-rank) 테스트는 두 모델들에 대해 유의성을 나타내지 않아($p=0.11$, $0.50\$$), 차이가 유의하지 않음을 나타낸다.
- [0118] 각 개별 주석자들로부터의 예측들을 그룹으로 보고, 주석자들 사이의 페어와이즈(pairwise) 차이를 계산했다: 각 샘플 피크에 대한 상대적 존재비 차이의 평균. 결과는, RSD 값과 인 라인(in line)의, 0.025 ± 0.009 에서 주석자들 사이의 불일치를 보여준다. 동시에, 모델 B 예측과 13명의 주석자들 간의 차이는 0.026 ± 0.007 로, 주석자 간 차이와 겹친다. 모델 A에 대응하는 값은 0.039 ± 0.005 이다. 모델 대 인간의 차이가 참조 기반 사례에서 크게 더 크지 않거나 심지어 인접하다는 점을 감안할 때, 신경망 기반 모델이 이 작업에서 수동 주석을 대신할 수 있다는 것은 분명하다. 우리는 또한 인간 주석자가 그들의 판독에 대한 참조 역할을 하기 위해 선택된 시작 및 종지와 함께 피크를 가졌기 때문에, 그들의 작업이 모델 B와 직접적으로 유사하다는 점에 주목한다.
- [0119] **3.3 존재비 계산을 위한 가중 피크 시작/종료**

- [0120] 솔더와 뱀리 포인트가 있는 피크에서, 경계의 결정은 훨씬 더 높은 변동성을 겪을 것이다. 인간 주석자들에서도 서로 다른 주석자들 사이의 변동이 존재하므로, 모델 A 및 모델 B의 출력은 훈련 샘플의 앙상블(ensemble)에 대한 평균으로 인해 다중 모드 모양을 나타낼 것으로 예상된다(도 9b, 도 7b). 이것은 강력한 참조 사전(reference prior)이 존재하지 않기 때문에 참조가 없는 경우(모델 A)에 다소 심각하여, argmax 경계를 적용할 때 결과가 때때로 임의적일 수 있다. 이러한 경우 경계의 단일 포인트 추정 대신 전체 예측을 더 잘 수용할 수 있는 방법이 검토되었다.
- [0121] 가장 직접적인 방법 중 하나는 예측 결과를 확률 분포로 해석하는 것이다. 다시 말해, 모델이 피크 시작/종료 마커 역할을 할 수 있는 시퀀스에서 가능한 모든 포인트 쌍의 확률을 평가하도록 하는 것이다. N개의 포인트들의 시퀀스의 경우, 순서 제약을 부과한 후 이는 총 $N \times (N-1)/2$ 쌍을 포함한다. 각 쌍의 확률은 두 포인트들에서 피크 시작/종료 작업들의 예측 값들을 곱하여 쉽게 계산될 수 있다. 그런 다음 확률을 사용하여 대응되는 쌍들 사이의 적분에 대한 가중 합계를 계산하여, MAP 추정값 대신 평균 추정값을 제공한다(이러한 존재비를 계산하기 위한 대안적 방법은 이하 "가중 합 방법"이라 칭함). MAP 추정은 argmax 값을 끝점으로 사용하여 하나의 존재비 값을 계산한다.
- [0122] 존재비의 평균은 특히 모델 B보다 모델 A와 더 유사한 모델 아키텍처가 사용되는 경우 argmax 경계를 사용하는 것보다 더 강력한 존재비 추정치를 제공할 수 있다. 명시적 경계가 사용되지 않기 때문에, 이 방법은 피크 시작 예측보다 먼저(또는 매우 가깝게) 나타나는 피크 종료 예측 문제를 피할 수 있다. 모델 A 및 모델 B에 대한 존재비 상관의 비교는 각각 도 11a 및 11b에 제시되어 있다.
- [0123] 특히 모델 A의 경우, 가중 합 방법이 보다 정확한 존재비를 생성하는 데 도움이 되었으며, 존재비 상관관계를 0.9897에서 0.9969로 높였다. 모델 B에서 증가는 상대적으로 미미하다(0.9974에서 0.9977로). 후처리 단계는 argmax 경계가 인간 주석보다 훨씬 낮은 존재비를 유발하는 이상점을 피하는 데 주로 도움이 되었다. 이러한 샘플은 일반적으로 피크 범위 내에서 예측된 경계 확률이 비정상적으로 높으며(뱀리 또는 솔더 포인트에서), 이는 그런 다음 argmax 작업에 의해 선택된다. 동시에, 가중 합 방법은 잡음이 많은 신호와 예측으로 인한 낮은 존재비 피크에 대해 전체 예측이 위쪽으로 바이어스되도록 했다. 신호-잡음 비율이 낮은 경우, 예측 경계 확률은 피크 주변에서 더 긴 머무름 시간 범위에서 변동하여(도 7a 및 7b, 및 도 12) 더 높은 가중치 합으로 이어질 수 있다. 노이즈가 많은 피크는 다운스트림 어플리케이션에서 가치가 떨어지기 때문에 이 효과는 허용 가능한 것으로 간주했다.
- [0124] 이 규칙에서 신뢰 값은 다음과 같이 계산되었다: N개의 포인트들이 있는 샘플 곡선에 대한 예측을 위해 범위의 모든 로컬 최대값을 계산한다. 전역 최대값은 다음과 같이 라벨링된다
- [0125] $(\hat{x}^*, \hat{y}^*)$, 여기서 $\hat{x}^* \in \{1, 2, \dots, N\}, \hat{y}^* \in [0, 1]$.
- [0126] 그의 왼쪽에 있는 로컬 최대값(존재하는 경우)은 다음과 같이 표시되며
- [0127] $(\hat{x}_i, \hat{y}_i), i = 1, 2, \dots, p$, 여기서 $\hat{x}_1 < \dots < \hat{x}_p$, 오른쪽의 로컬 최대값(존재하는 경우)은 다음과 같이 라벨링되며 $(\tilde{x}_j, \tilde{y}_j), j = 1, 2, \dots, q$, 여기서 $\tilde{x}_1 > \tilde{x}_2 > \dots > \tilde{x}_q$. 다음과 같이 정의하여 $\hat{x}_0 = \hat{x}^*$ 및 $\tilde{x}_0 = \tilde{x}^*$, 신뢰도는 다음과 같이 계산된다:
- [0128]
$$\text{신뢰도} = \hat{y}^* - \frac{1}{N-1} \left(\sum_{i=1}^p (\hat{x}_i - \hat{x}_{i-1}) \times \max(\hat{y}_i, \hat{y}_{i+1}, \dots, \hat{y}_p) + \right.$$
- [0129]
$$\left. \sum_{j=1}^q (\tilde{y}_{j-1} - \tilde{y}_j) \times \max(\tilde{y}_j, \tilde{y}_{j+1}, \dots, \tilde{y}_q) \right)$$
- [0130] 도 12에 그래픽 예시가 도시되어 있다.
- [0131] **3.4 예측 신뢰도**
- [0132] 포인트-와이즈(point-wise) 예측 구조의 또 다른 핵심 이점은 예측 확실성을 쉽게 추정할 수 있다는 것이다. 신경망 예측에서 불확실성을 정량화하는 내부 방법은 여전히 복잡하고 제한이 따르지만, 이 작업에서는 모델 예측에서 경계를 검사하여 특정/불확실성을 추론하는 것이 상당히 간단하다. 대부분의 예측된 확률은 모델의 신뢰도

를 나타내는 양과 높이인 가우스 혼합 형태를 취한다. 이 연구에서는 아래 가이드라인에 따라 예측 신뢰도를 정량화하기 위해 휴리스틱 방법(heuristic method)을 채택했다:

- [0133] 아이템 더 낮은 예측 값 -> 더 낮은 신뢰도;
- [0134] 아이템 다중 로컬 최대치 -> 낮은 신뢰도;
- [0135] 아이템 더 많은 분산된 로컬 최대치 -> 낮은 신뢰도.
- [0136] 이 규칙으로부터, 신뢰 값이 계산되었다. 도 12는 그래픽 그림과 함께 신뢰 수준을 계산하는 데 사용되는 수학에 대한 설명을 제공한다. 신뢰 점수는 음영 영역에 비례한다. 하나는 예측 값이 높고(1에 가까움) 단봉형(unimodal)이기를 원하거나 또는 다중 로컬 최대치가 나타나는 경우 단기를 원한다.
- [0137] 모델 A 및 모델 B에 대한 모든 예측에 대해 신뢰 점수가 계산되었다(도 13a 및 14a). 다양한 신뢰 컷오프(cutoff)들을 통해, 성능 변화의 추세를 확인할 수 있다. 도 13b 및 14b에 도시된 바와 같이 경계에 대한 예측 오류는 신뢰도가 증가함에 따라 꾸준히 감소하며, 가장 신뢰도가 높은 샘플에서는 거의 1초로 떨어진다.
- [0138] 상이한 신뢰 구간들에 대한 존재비들의 상관관계는 도 13a 내지 도 13f에 예시되어 있다. 높은 존재비 샘플들에 대한 예측은 일반적으로 더 높은 신뢰도를 가지며 사람의 주석과 더 잘 일치한다. 상위 두 개 빈들 예측들은 0.999를 초과하는 피어슨의 상관 계수를 달성했다. 대조적으로, 예측 신뢰도가 낮은 샘플은 더 높은 경계 평균 절대 오차(MAE) 및 더 낮은 존재비 상관 관계로 더 나쁜 성능을 산출한다. 이는 부분적으로 낮은 존재비로 표시된 입력 샘플의 더 나쁜 신호-잡음 비율 때문이다. 전반적으로 신뢰도 측정은 노이즈 스케일과 모델 성능의 좋은 지표로 사용되었다.
- [0139] 이 후처리 단계는 워크플로 품질 관리를 구현하는 데 다용성을 제공할 수 있다. 실제 어플리케이션에서는, 직접 사용은 임계값 신뢰 점수를 설정하는 것이며, 따라서 값 이하의 예측은 제외되거나 재평가를 위해 인간 전문가에게 제시되어야 한다. 이것은 또한 실험 데이터의 이상을 감지하는 데 도움이 될 것이다.

[0140] 5.0 실험적

[0141] 실험적 파라미터

[0142] 화학물질 및 시약: 인간 혈청, 디티오프레이톨(DTT) 및 요오도아세트아미드(IAA)는 시그마-알드리치(Sigma-Aldrich)(미주리주 세인트루이스)에서 구입했다. 시퀀싱 등급 트립신은 프로메가(Promega)(매디슨, 위스콘신)에서 구입했다. 아세트ونی트릴(LC-MS 등급)은 허니웰(Honeywell)(미시건주 머스키건)에서 구입했다.

[0143] 샘플 준비: 혈청 샘플을 DTT로 환원시키고 IAA로 알킬화시킨 후 18시간 동안 수조에서 트립신으로 소화시켰다. 소화를 퀸칭(quenching)하기 위해 인큐베이션(incubation) 후 포름산을 1%(v/v)의 최종 농도로 각 샘플에 첨가했다.

[0144] 5.1 LC-MS 분석

[0145] 소화된 혈청 샘플을 애질런트(Agilent) 1290 인피니티(infinity) UHPLC 시스템과 애질런트(Agilent) ZORBAX 이클립스 플러스(Eclipse Plus) C18 컬럼(2.1 mmx 150 mm i.d., 1.8 μ m 입자 크기)이 장착된 애질런트(Agilent) 6490 삼중 사중극자 질량 분석기에 주입했다. 펩티드와 당펩티드의 분리는 70분 바이너리 구배를 사용하여 수행되었다. 수성 이동상 A는 3% 아세트ونی트릴, 물 중 0.1% 포름산(v/v)이었고 그리고 유기 이동상 B는 물(v/v) 중 90% 아세트ونی트릴 0.1% 포름산이었다. 유속은 0.5mL/분으로 설정되었다. ESI(Electrospray ionization)는 이온화 소스로 사용되었으며 양이온 모드에서 작동되었다. 삼중 사중극자는 동적 다중 반응 모니터링(dMRM) 모드에서 작동되었다. 훈련 및 테스트에 사용된 펩티드 및 글리코펩티드 전환은 리(Li), 외. 개선된 질병 바이오마커 발견을 위한 예측적 당펩티도믹스로 강화된 50개 혈청 당단백질의 부위 특이적 당화 정량화, 분석(Anal.) 화학 2019, 91, 5433-5445; DOI: 10.1021/acs.분석화학(analchem).9b00776에 발표된 것 중 일부이며, 그 전체 내용은 모든 목적을 위해 그 전체가 참조로 본원에 포함된다.

[0146] 5.2 데이터세트 준비

[0147] 훈련 및 검증 세트 피크와 테스트 세트 피크는 두 세트의 질량 스펙트럼 실험에서 수집되었다. 세트 1은 2017년 말 UC 데이비스(Davis)에 있는 칼리토 레브리라(Carlito Lebrilla)의 연구실에서 측정된 인간 혈청 샘플과 전이로 구성되었으며, 세트 2는 2018년 중반에 인디뮴드(Indivumed) GmbH에서 구입한 난소암 및 양성 종괴 환자로부터 시판되는 혈청이었다. 두 세트의 샘플 모두 위에서 설명한 실험 프로토콜을 거쳤다. 실행 후 애질런트

(Agilent).D 형식 파일은 생물정보학 처리를 위해 프로테오위자드(Proteowizard) 3.0 소프트웨어 제품군 내에서 ms전환(msConvert)를 통해 mzML로 변환되었다.

[0148] 인간 주석을 얻기 위해, 질량 스펙트럼 전문가가 하나의 참조 혈청 샘플(풀링 시그마 혈청(pooled Sigma serum))에 대해 모든 전이들에 라벨링하였으며, 이러한 라벨들이 "참조" 피크들을 구성한다. 나머지 샘플의 경우, 애질런트(Agilent) 메스헌터(MassHunter) 정량 분석 B.5.0에서 피크 시작 및 피크 종료 머무름 시간 표시 라벨링된 주석자는 참조 피크를 기반으로 하고 머무름 시간의 변화를 설명한다.

[0149] 두 실험에서, 양의 존재비로 표시된 피크만 선택되었다. 머무름 시간 시프트가 큰(> 0.2 분) 샘플도 제외되었다. 전체 106355개(훈련/검증) 및 67672개(테스트) 피크들이 온전성 테스트(sanity test)를 통과했으며 이 연구에 사용되었다. 각 전환에 대해 참조 머무름 시간 시작 - 0.2분에서 참조 머무름 시간 중지 + 0.2분까지 머무름 시간 윈도우 내의 모든 신호를 수집하여, 원하는 전구체 및 제품 m/z 주위에 +/- 0.1의 질량 전하 비율 윈도우를 허용한다. 신호들은 질량 전하 비율 윈도우를 따라 합산되어 추출된 이온 크로마토그램(XIC)을 생성한다. 순차 모델과 인간 주석자 모두 타겟 전환 주위에서 잘린 XIC 곡선으로 표시된다(도 7a, 7b 및 9b는 몇 가지 예시적인 XIC 입력을 보여준다).

[0150] "인간 주석자와의 비교"라는 제목의 위 섹션에서 사용된 테스트 세트에는 1619개의 피크들이 포함되어 있다. 모든 피크들은 일관성을 위해 독립적으로 12 명의 주석자들에 의해 추가로 추가되었지만, 양의 존재비를 갖는 주석들만이 전이별 상대 표준 편차 및 주석자들 간의 페어-와이즈 불일치를 계산하는 데 사용되었다.

[0151] 5.3 특성화

[0152] 입력은, 고정된 시작/중지 또는 길이 없이, 강도 머무름 시간 평면에서 포인트들의 시퀀스로 공식화된다. 균일한 표현을 생성하기 위해, 각 샘플 곡선에 특징화를 적용했다. 단일 샘플의 모든 포인트들에 대한 머무름 시간은 먼저 샘플 곡선의 정점 위치에 중심을 둔 다음 -1분에서 1분 범위의 128개의 동일한 이격된 가우스 빈(Gaussian bin)들로 확장되어, 시점을 나타내는 256개의 실수로 구성된 벡터 F_{RT} 가 생성되었다.

[0153] 질량 대 전하 강도 값은 넓은 범위를 가지며, 강도 값은 0에서 500까지의 범위에 256 가우시안 함수들을 적용하여 유사하게 이산화되어, 아날로그 값을 강도를 나타내는 256개의 실제 값의 벡터 I_{RT} 로 변환하였다. 강도가 상단 임계값보다 큰 임의의 포인트는 피크의 일부로 간주되어야 한다는 가정을 채택하여, 노이즈 또는 노이즈 신호 전환 단계에 있는 포인트만 특성화하였다. F_{RT} 및 I_{RT} 가 형성된 후 각 샘플 곡선에 대한 전체 입력 A(즉, RNN 모델에 대한 입력 파라미터 또는 상태)는 크기 $N \times 384$ 의 행렬이다. 변수 N 은 곡선의 길이를 나타내며 50에서 300 사이이다.

[0154] 모델 B의 경우 각 샘플은 해당 참조와 쌍을 이룬다. 입력들의 쌍은 위와 동일한 특징화 프로세스를 가지며 $N \times 384$ 및 $N' \times 384$ 차원의 두 행렬 X 및 X_{ref} 로 형성된다. N 과 N' 은 가까워야 하지만 반드시 같을 필요는 없다.

[0155] 샘플의 라벨링은 동일한 시퀀스 구조로 공식화되었으며, 여기서 피크 시작/종료의 인간 주석은 대응되는 작업에서 1의 값을 갖는다. 안정성과 견고성을 높이기 위해 추가 라벨 평활화(label smoothing)가 적용되었다: 사람이 주석을 추가한 라벨/실측 자료 시작/종류 주석에 가까운 샘플 포인트도 양의 라벨 값을 가지며 다음과 같이 계산된다:

$$[0156] \tilde{y} = \exp\left(-\frac{d_{RT}^2}{\sigma_{RT}} - \frac{d_{ABC}^2}{\sigma_{ABC}}\right)$$

[0157] d_{RT} 와 d_{ABC} 는 머무름 시간(RT)의 차이와 풍부도(ABD)의 차이이고 σ_{RT} 와 σ_{ABC} 는 훈련 중 각각의 감쇠율을 제어하는 하이퍼파라미터이다. 평활화의 목적은 노이즈 라벨에 대한 모델 적응성을 허용하고 단일 포인트를 지정하지 않는 사람의 라벨링/주석을 설명하는 것이다. 그러나 라벨 근처의 이웃 포인트는 노이즈에서 신호로의 전환을 나타내는 합리적인 포인트를 나타낼 수도 있다.

[0158] 5.4 훈련 설정

[0159] 도 5a 및 도 5b(모델 A) 및 도 6a 및 도 6b(모델 B)를 다시 참조하면, 모델 A 및 모델 B는 파이토치(Pytorch)에서 구현되었다.

- [0160] 모델 A에는 은닉(hidden) 상태를 위한 128개의 특징이 있는 두 개의 바이(bi)-LSTM 계층이 포함되어 있으며 LSTM 출력은 8개의 헤드와 각 헤드에 32개의 특징이 있는 다중 헤드 주의 계층에 적용된다. 헤드들로부터의 출력들은 연결되어 32개의 유닛이 있는 완전히 연결된 은닉 계층을 거친 다음, 아핀 계층(affine layer)을 통해 각각 두 개의 클래스들이 있는 두 개의 작업들에 매핑된다. 소프트맥스(Softmax)는 피크 시작/종료 확률을 생성하기 위해 시퀀스의 각 작업 및 각 포인트에 적용된다. 모델은 20 시대(epoch)에 대한 86311개의 샘플들을 포함하고 배치 크기가 8이고 학습률이 0.001인 훈련 서브세트에 대한 부드러운 라벨과 소프트맥스(softmax) 예측들 간의 교차 엔트로피 손실에 의해 훈련된다. 검증 세트 실적에 따라 조기 중지가 적용된다.
- [0161] 모델 B는 샘플과 참조를 각각 인코딩하는 데 사용되는, 완전 연결 계층을 제외하고, 위와 동일한 두 세트의 시퀀스 네트워크 구조를 포함한다. 가중치는 두 구조들 사이에 공유되지 않는다. 샘플 참조 주의 계층은 샘플 인코딩(쿼리로서), 참조 인코딩(키로서) 및 참조 레이블(값으로서)에 적용되며, 이의 출력은 샘플 인코딩과 길이가 같고 작업 차원에서 2가 된다. 이것은 피크 시작/종료 확률에 대한 모델 예측으로 직접 사용된다. 또 다른 손실 텀(term)은 샘플 참조 주의 맵과 샘플 및 참조에서 동일한 머무름 시간을 가진 포인트를 연결하는 표준 매핑 간의 KL 발산으로 계산된다(단위 행렬과 유사). 모델은 교차 엔트로피 손실과 동일한 설정에서 동일한 훈련 세트에 대한 추가 발산 텀으로 훈련된다.
- [0162] 추가 정의
- [0163] 본원에서 사용된 "프로세서"는 SAS 프로그램 코드 실행, GUI 용 데이터 처리, 메모리 관독 또는 기록과 같은 작업을 수행하는 데 사용되는 하나 또는 복수의 프로세서들을 포함하는 것으로 간주되어야 하며, 범용 컴퓨터를 포함한다.
- [0164] 본원에서 사용된 "사용자"는 어플리케이션을 로컬에서 사용하거나 원격에 위치한 컴퓨터를 사용하여 웹 브라우저를 사용하여 어플리케이션에 원격으로 액세스하는 실제 사람을 포함하는 것으로 간주되어야 한다.
- [0165] 본원에서 사용된 "컴퓨터"는 하드 디스크에서 관독하고 기록하기 위한 하드 디스크 드라이브, 이동식 자기 디스크에서 관독하고 기록하기 위한 자기 디스크 드라이브, 및 CD ROM, DVD 또는 기타 광학 미디어와 같은 이동식 광 디스크에서 관독하고 기록하기 위한 광 디스크 드라이브를 갖는 컴퓨터를 포함하는 것으로 간주되어야 한다. 하드 디스크 드라이브, 자기 디스크 드라이브 및 광 디스크 드라이브는 각각 하드 디스크 드라이브 인터페이스, 자기 디스크 드라이브 인터페이스 및 광 드라이브 인터페이스에 의해 시스템 버스에 연결된다. 드라이브 및 그의 관련 컴퓨터 관독 가능 매체는 컴퓨터 관독 가능 명령어, 데이터 구조, 프로그램 및 컴퓨터 시스템에 대한 기타 데이터의 비휘발성(비일시적) 저장소를 제공한다.
- [0166] 본원에서 사용된 "데이터베이스"는 중앙 집중식 또는 분산형 데이터베이스를 포함하는 것으로 간주되어야 하며 SQL 데이터베이스를 포함한다.
- [0167] 본원에서 사용된 바와 같이, "데이터"라는 용어는 파일 시스템의 기계 관독 가능 메모리에 포함되거나 데이터베이스 또는 기타 구조화된 데이터에 사용되는 모든 정보, 또는 일시적 또는 비일시적 메모리의 구조화되지 않은 데이터 표현을 포함하는 것으로 간주되어야 한다.
- [0168] 본원에서 사용된 바와 같이, "입력 디바이스"는 키보드, 마우스, 포인터, 터치 패드, 음성 인식기 또는 메모리에 저장될 수 있는 전자 콘텐츠를 수정, 탐색, 액세스 또는 생성하기 위한 컴퓨터(디스플레이 유무에 관계없이)에 연결된 기타 형태의 입력 디바이스를 포함하는 것으로 간주되어야 한다.
- [0169] 본원에서 사용된 바와 같이, "디스플레이"는 LCD/LED 유형 디스플레이 또는 컴퓨터에 연결되어 GUI를 사용하여 입력을 렌더링 및 수신하거나 기계 실행 코드를 실행하는 컴퓨터에서 실행되는 응용 프로그램에서 생성된 기타 정보를 표시할 수 있는 기타 유형의 디스플레이 디바이스이다.
- [0170] 본원에서 사용된 바와 같이, "모듈" 또는 "구성 요소"는 소프트웨어 또는 소스 코드 파일, 또는 소스 코드 파일에 포함되거나 포함되지 않고 비일시적 또는 일시적 메모리 매체에 존재하는 기계 실행 가능한 명령어 세트를 포함하는 것으로 간주되어야 한다. 모듈은 독립 실행형 프로그램으로 존재하거나 다른 프로그램에서 호출한 프로그램을 나타낼 수 있다. "구성 요소"는 모듈이 구성 요소를 호출하여 모듈의 작업 또는 작업(들) 부분을 수행한다는 의미에서 "모듈"의 일부이다.
- [0171] 사용 및 해석되는 "데이터 리포지토리" 또는 "리포지토리"에는 데이터베이스, 데이터 저장소, 데이터 웨어하우스, 데이터 레이크 또는 이들의 조합이 포함될 수 있다.
- [0172] 본원에서 사용된 "메모리"는 일시적인 컴퓨터 관독 가능한 저장 매체 또는 비일시적인 컴퓨터 관독 가능한 저장

매체인 기계 판독 가능한 매체를 포함하는 것으로 간주되어야 한다. 다양한 실시 예들에서, 메모리의 휘발성(일시적) 부분은 정적 랜덤 액세스 메모리(SRAM), 동기식 동적 RAM 또는 임의의 다른 유형의 메모리와 같은 임의의 적절한 메모리 기술을 사용하여 구현될 수 있다. 도시되고 설명된 메모리의 예시는 단일 기계 판독 가능한 매체를 참조할 수 있지만, 메모리는 하나 이상의 기계 실행 코드 또는 데이터 세트를 저장하는 단일 매체 또는 다중 매체(예를 들어, 중앙 집중식 또는 분산형 데이터베이스, 및/또는 관련 캐시 및 서버, 및 또는 프로세서 레지스터, 메모리 및 저장 장치와 같은 다양한 저장 매체)를 포함하는 것으로 간주되어야 한다. "기계 판독 가능 매체"라는 용어는 기계에 의한 실행을 위한 명령 세트를 저장, 인코딩 또는 전달할 수 있고 기계가 본 발명의 방법론 중 하나 이상을 수행하게 하고 또는 그러한 명령어 세트에 의해 이용되거나 이와 관련된 데이터 구조를 저장, 인코딩 또는 운반할 수 있는 모든 매체를 포함하는 것으로 간주된다. 따라서 메모리라는 용어는 솔리드 스테이트 메모리, 광학 및 자기 매체와 같은 유형 매체를 포함하는 것으로 간주되어야 한다.

[0173] 본원에서 사용된 "모듈"은 다른 모듈에 정보를 제공하고 다른 모듈로부터 정보를 수신할 수 있는 고유한 기능의 단위를 포함하는 것으로 간주되어야 한다. 따라서, 설명된 모듈들은 통신적으로 결합된 것으로 간주될 수 있다. 모듈은 또한 입력 또는 출력 장치와 통신을 시작할 수 있으며 리소스에서 작동할 수 있다(예를 들어, 데이터베이스와 같은 정보 모음). 모듈은 다양한 실시 예의 특정 구현에 적절한 하드웨어 회로, 광학 구성요소, 단일 또는 다중 프로세서 회로, 메모리 회로, 소프트웨어 프로그램 모듈 및 객체, 펌웨어, 및 이들의 조합을 포함할 수 있다. "모듈"이라는 용어는 특정 기능, 작업, 처리 또는 절차를 달성하기 위한 코드, 데이터 또는 계산 개체의 식별 가능한 부분을 포함할 수 있다.

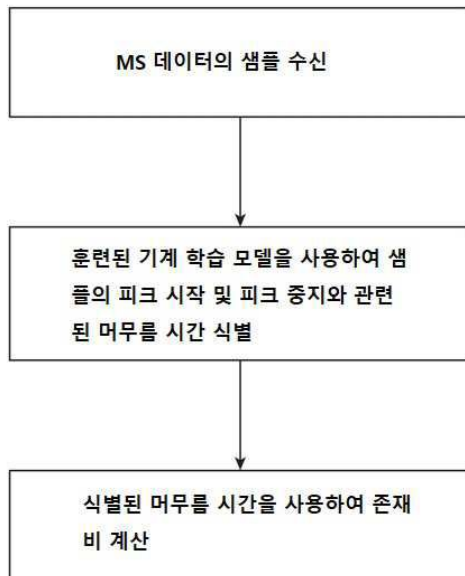
[0174] 일부 예에서, 질량 분석 전환을 정량화하는 단계-여기서 전환은 아미노산 서열을 포함하는 글리코펩티드 또는 펩티드의 특징적인 m/z 강도 값 피크 임- 및 샘플이 정량화를 기반으로 질병 분류에 속하거나 속하지 않을 확률을 결정하기 위해 모델을 사용하여 정량화를 분석하는 단계를 포함하는 질병 진단 방법이 본원에 제시되어 있다. 이들 예 중 일부에서, 전이, 당펩티드, 펩티드, 아미노산 서열, 또는 이들의 임의의 조합은 2020년 1월 31일에 출원된 국제 PCT 특허 출원 PCT/US2020/0162861호에 기재된 것일 수 있다.

[0175] 요약에 설명된 것을 포함하여 본 발명의 예시된 실시 예에 대한 상기 설명은 본 발명을 개시된 정확한 형태로 완전하게 한정하거나 제한하려는 의도가 아니다. 본 발명의 특정 실시 예 및 예가 예시 목적으로 본 명세서에 기재되어 있지만, 관련 기술분야의 통상의 기술자가 인식하는 바와 같이 본 발명의 범위 내에서 다양한 변형이 가능하다.

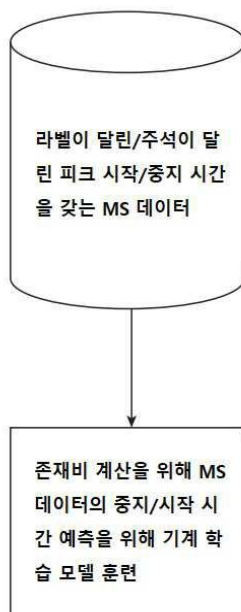
[0176] 이러한 수정은 상기 상세한 설명에 비추어 본 발명에 만들어질 수 있다. 청구범위에 사용된 용어는 본 발명을 명세서에 개시된 특정 실시 예로 제한하는 것으로 해석되어서는 안 된다.

도면

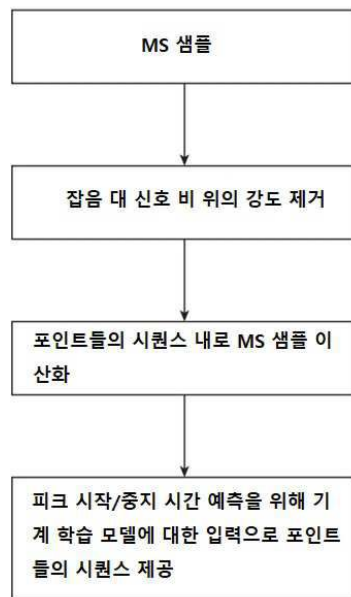
도면1



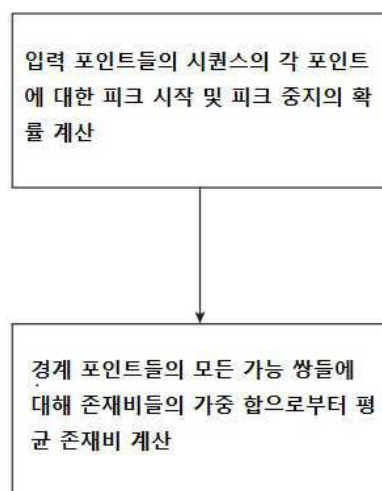
도면2



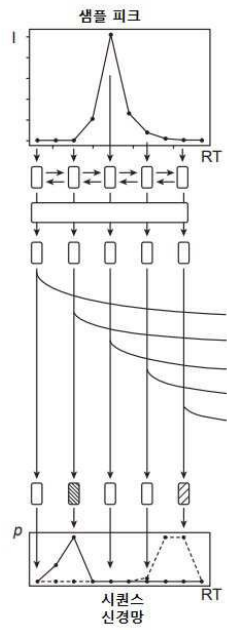
도면3



도면4



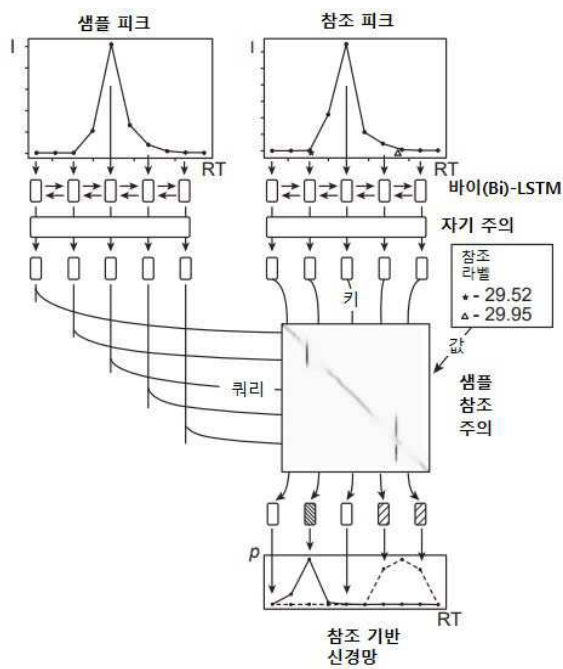
도면5a



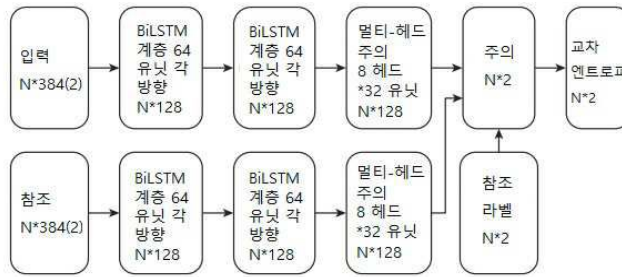
도면5b



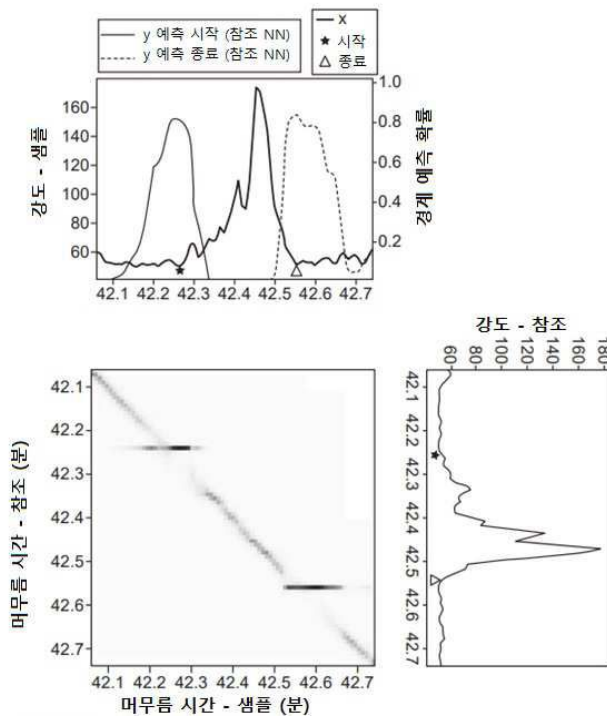
도면6a



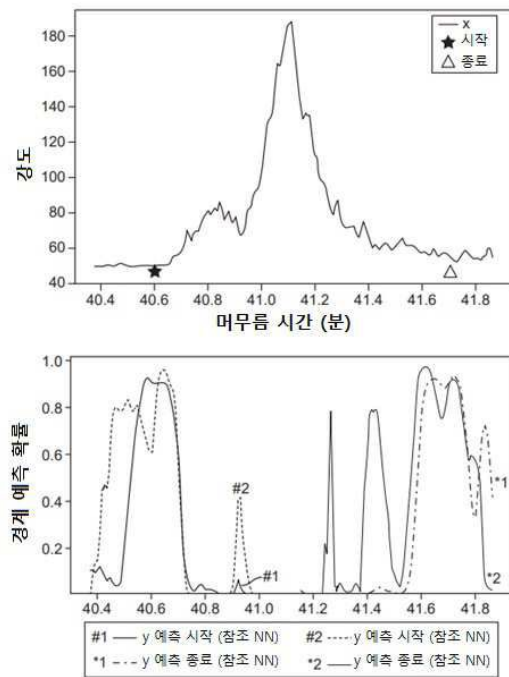
도면6b



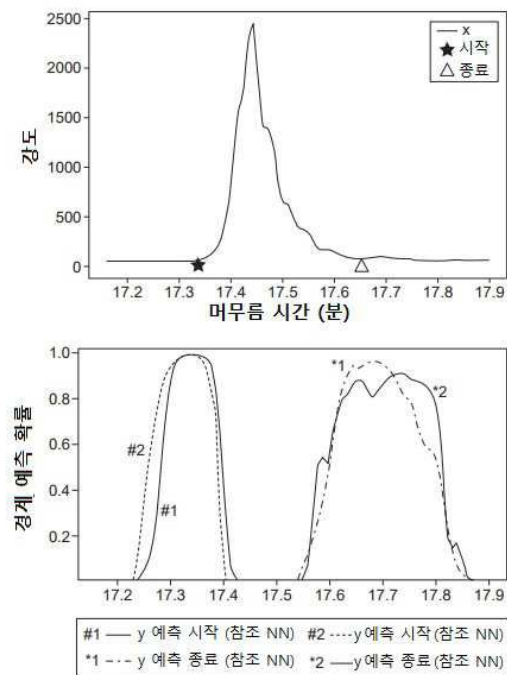
도면6c



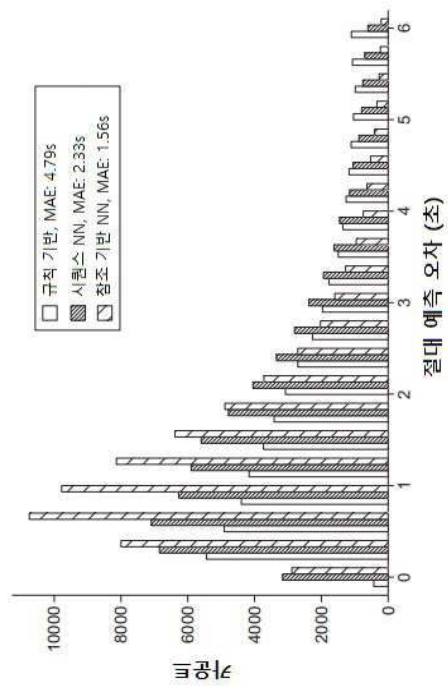
도면7a



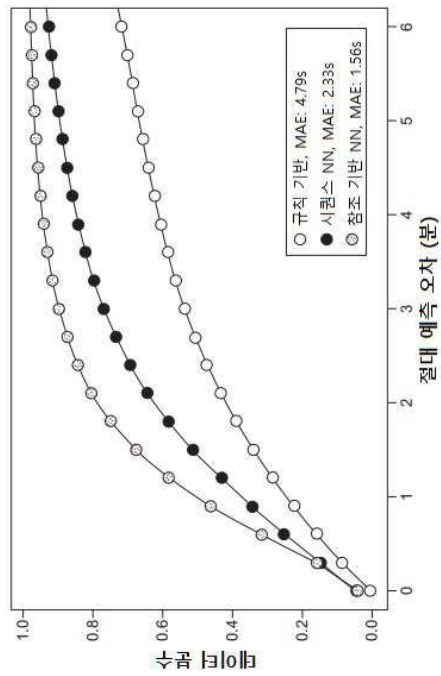
도면7b



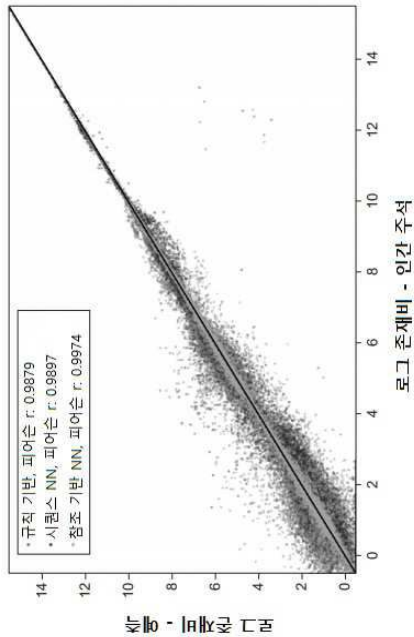
도면8a



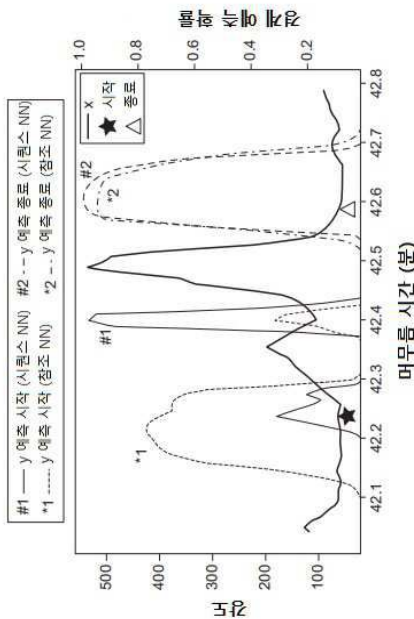
도면8b



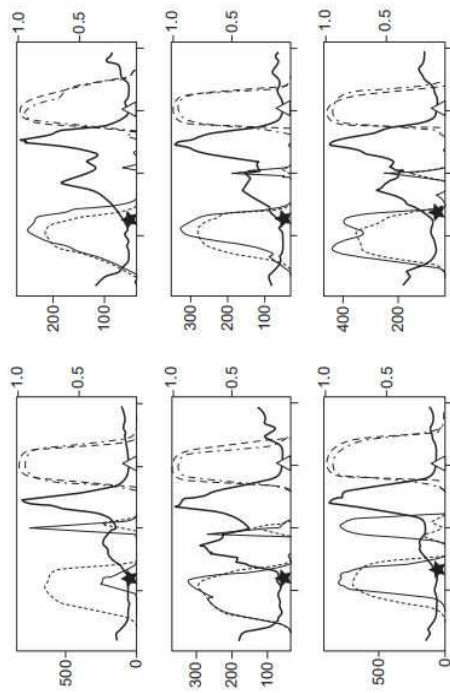
도면8c



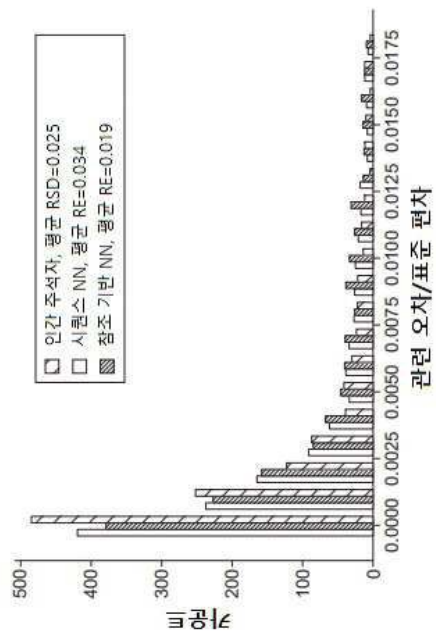
도면9a



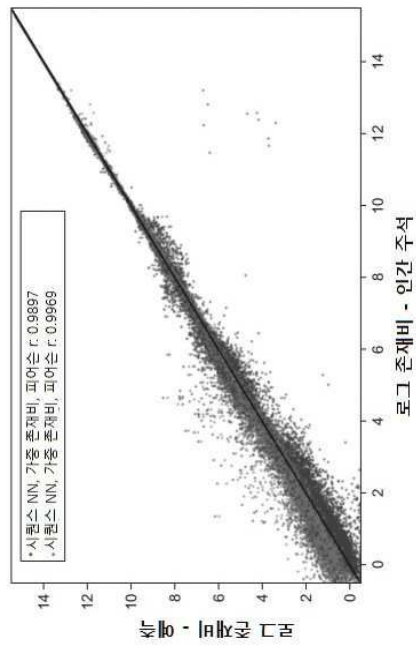
도면9b



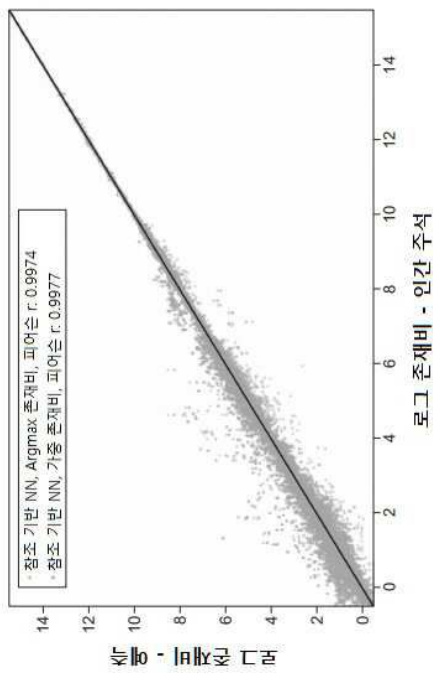
도면10



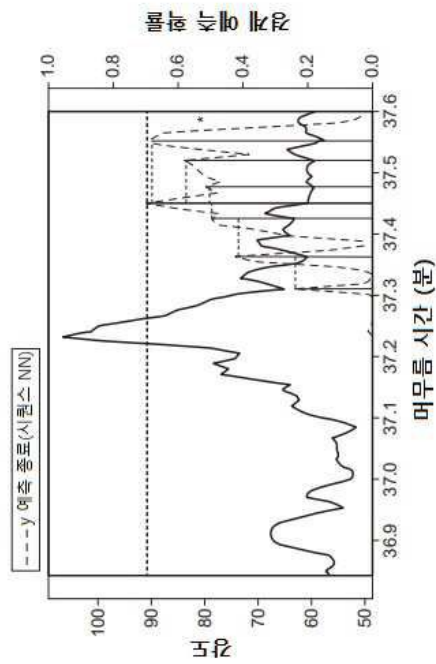
도면11a



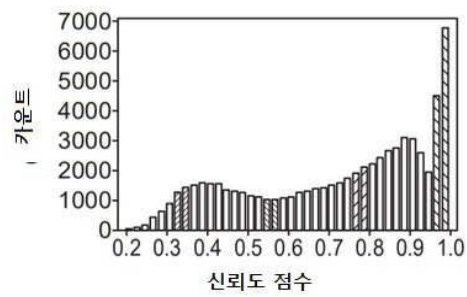
도면11b



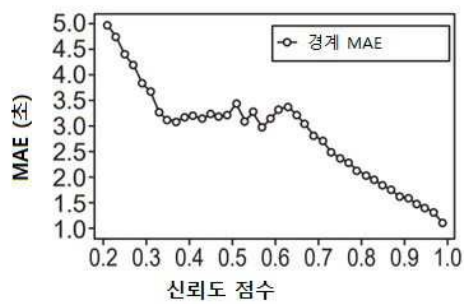
도면12



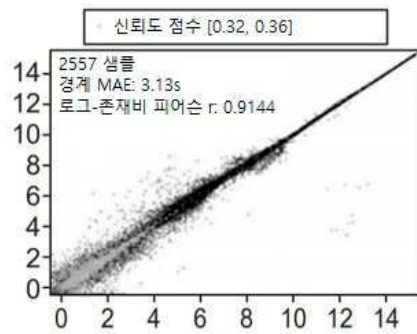
도면13a



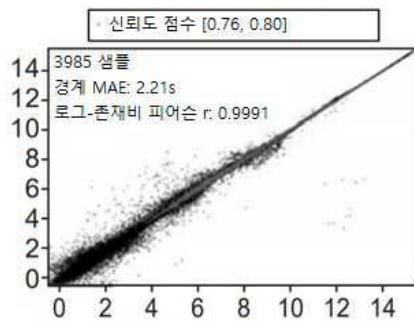
도면13b



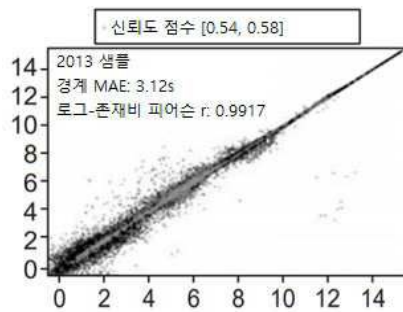
도면13c



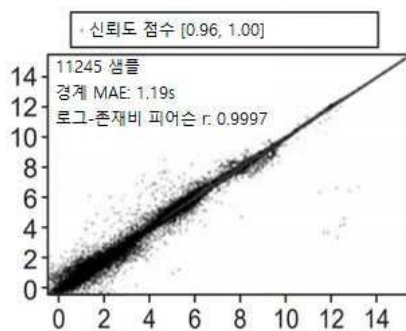
도면13d



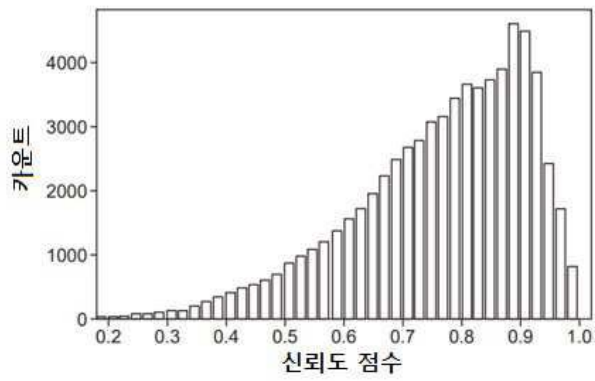
도면13e



도면13f



도면14a



도면14b

