(12) 특허협력조약에 의하여 공개된 국제출원

(19) 세계지식재산권기구
국제사무국

(43) 국제공개일
2024 년 5 월 30 일 (30.05.2024)

WIPO | PCT

(10) 국제공개번호
WO 2024/112153 A1

(51) 국제특허분류:
G16B 40/00 (2019.01) *G16B 45/00* (2019.01)
G16B 30/10 (2019.01) *G16B 5/00* (2019.01)

(21) 국제출원번호: PCT/KR2023/019095

(22) 국제출원일: 2023 년 11 월 24 일 (24.11.2023)

(25) 출원언어: 한국어

(26) 공개언어: 한국어

(30) 우선권정보:
10-2022-0160091 2022 년 11 월 25 일 (25.11.2022) KR

(71) 출원인: 주식회사 씨젠 (**SEEGENE, INC.**) [KR/KR]; 05548 서울특별시 송파구 오금로 91, 지하 1층,3층,4층,5층,6층,7층,8층,9 층,10층,11층,12층, Seoul (KR).

(72) 발명자: 정하늘 (**JEONG, Ha Neul**); 05548 서울특별시 송파구 오금로 91, 지하1층,3층,4층,5층,6층,7층,8층,9층,10층,11층,12층, Seoul (KR). 장종하 (**JANG, Jong Ha**); 05548 서울특별시 송파구 오금로 91, 지하 1층,3층,4층,5층,6층,7층,8층,9 층,10층,11층,12층, Seoul (KR). 김현호 (**KIM, Hyun Ho**); 05548 서울특별시 송파구 오금로 91, 지하1층,3층,4층,5층,6층,7층,8층,9층,10층,11층,12층, Seoul (KR). 이광호 (**LEE, Gwang Ho**); 05548 서울특별시 송파구 오금로 91, 지하 1층,3층,4층,5층,6층,7층,8층,9 층,10층,11층,12층, Seoul (KR).

(74) 대리인: 제일특허법인(유) (**FIRSTLAW P.C.**); 06775 서울특별시 서초구 마방로 60, Seoul (KR).

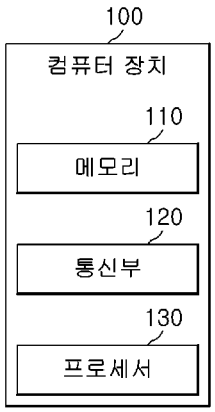
(81) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 국내 권리의 보호를 위하여): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 역내 권리의 보호를 위하여): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 유라시아 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 유럽 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

공개:
— 국제조사보고서와 함께 (조약 제21조(3))

(54) Title: METHOD FOR ESTIMATING ORGANISM OR HOST, METHOD FOR ACQUIRING MODEL FOR ESTIMATING ORGANISM OR HOST, AND COMPUTER DEVICE FOR PERFORMING SAME

(54) 발명의 명칭: 유기체 또는 HOST 추정 방법, 유기체 또는 HOST를 추정하는 모델의 획득 방법 및 이를 수행하는 컴퓨터 장치



100 ... Computer device
110 ... Memory
120 ... Communication unit
130 ... Processor

(57) Abstract: Provided according to an embodiment is a computer-implemented method that is performed by a computer device using a memory, a processor, and one or more programs stored in memory and executed by a processor, the method comprising the steps of: accessing an estimation model obtained by fine-tuning a pre-trained model; providing a nucleic acid sequence to the estimation model; and estimating an organism containing the nucleic acid sequence or a host of the organism from the estimation model.

(57) 요약서: 일 실시예에 따라, 메모리, 프로세서 및 상기 메모리에 저장되고 상기 프로세서에 의해 실행되도록 구성된 하나 이상의 프로그램을 사용하는 컴퓨터 장치에 의해 수행되는 컴퓨터 구현 방법에 있어서, 사전 학습된 모델을 fine-tuning 하여서 획득된 추정 모델에 접근하는 단계; 상기 추정 모델에 핵산 서열을 제공하는 단계; 및 상기 추정 모델로부터, 상기 핵산 서열을 포함하는 유기체 또는 상기 유기체의 host 를 추정하는 단계를 포함하는, 컴퓨터 구현 방법이 제공된다.

WO 2024/112153 A1

명세서

발명의 명칭: 유기체 또는 HOST 추정 방법, 유기체 또는 HOST 를 추정하는 모델의 획득 방법 및 이를 수행하는 컴퓨터 장치

기술분야

[1] 본 발명은 유기체 또는 host 추정 방법, 유기체 또는 host를 추정하는 모델의 획득 방법 및 이를 수행하는 컴퓨터 장치에 관한 것이다.

배경기술

[2] 분자 진단은 현재 질병을 조기 진단하기 위한 체외 진단시장에서 빠르게 성장하고 있는 분야이다. 그 중에서 핵산(nucleic acid)을 이용한 방법들이 높은 특이성과 민감성을 바탕으로 바이러스, 박테리아에 의한 감염 등에 의한 원인 유전 인자를 진단하는데 유용하게 사용되고 있다.

[3] 핵산을 이용한 진단 방법들 대부분에서는 타겟 핵산(예컨대, 바이러스 또는 박테리아 핵산)을 증폭하는 핵산 증폭 반응(nucleic acid amplification reaction)을 이용된다. 대표적인 예로서, 핵산 증폭 반응 중 중합효소 연쇄반응(Polymerase chain reaction: PCR)에서는 이중가닥 DNA의 변성, DNA 주형에로의 올리고뉴클레오타이드 프라이머의 어닐링 및 DNA 중합효소에 의한 프라이머 연장의 반복된 사이클 과정이 수행된다(Mullis 등, 미국 특허 제4,683,195호, 제4,683,202호 및 제4,800,159호; Saiki et al., Science 230:1350-1354(1985)).

[4] PCR-기반 기술들은 타겟 DNA 서열의 증폭 뿐만 아니라 생물학 및 의학 연구 분야에서 과학적 응용 또는 방법에 널리 이용되고 있다. 이러한 PCR-기반 기술에는 예컨대, 역전사 효소 PCR(RT-PCR), 분별 디스플레이 PCR(DD-PCR), PCR에 의한 공지 또는 미지의 유전자의 클로닝, cDNA 말단의 고속 증폭(RACE), 임의적 프라이밍 PCR(AP-PCR), 멀티플렉스 PCR, SNP 지놈 타이핑, 및 PCR-기반 지놈 분석이 있다(McPherson and Moller, (2000) PCR. BIOS Scientific Publishers, Springer-Verlag New York Berlin Heidelberg, NY).

[5] 핵산을 증폭하기 위한 다른 방법으로 LCR(Ligase Chain Reaction), SDA(Strand Displacement Amplification), NASBA(Nucleic Acid Sequence-Based Amplification), TMA(Transcription Mediated Amplification), RPA(Recombinase polymerase amplification), LAMP(Loop-mediated isothermal amplification) 및 RCA(Rolling-Circle Amplification)와 같은 다양한 방법들이 제시되었다.

[6] 이처럼 핵산 증폭 반응을 기반으로 하나의 튜브 내에서 타겟 핵산을 검출하기 위한 분자 진단 기술이 사용되고 있으며, 최근에는 하나의 튜브 내에서 복수개의 타겟 핵산을 검출하기 위한 멀티플렉스(multiplex) 진단 기술이 사용되고 있다. 대표적인 예로서, PCR-기반 기술 중 멀티플렉스 PCR은 하나의 튜브 내에서 복수개의 올리고뉴클레오타이드 세트(전방향 및 역방향 프라이머, 및 프로브)의 조합

을 이용하여 하나의 타겟 핵산분자 또는 복수개의 타겟 핵산분자의 복수개의 영역을 동시에 증폭 및 검출하는 것을 의미한다.

[7] 이러한 분자 진단 기술에 이용되는 올리고뉴클레오타이드 세트는 타겟 핵산분자의 복수개의 핵산서열을 최대한의 커버리지로 검출할 수 있는 성능(performance)를 가지도록 디자인되어야 한다. 상기 올리고뉴클레오타이드 세트에 포함되는 올리고뉴클레오타이드(프라이머 및 프로브)가 타겟을 잘 검출하기 위한 성능을 갖도록 디자인되기 위해서는 해당 타겟 핵산분자의 핵산서열에 대한 충분한 정보가 제공될 필요가 있다.

[8] 이에 타겟 핵산분자의 핵산서열에 대한 정보들을 다양한 방식으로 수집하여 올리고뉴클레오타이드의 디자인에 적용하는 기술이 활용되고 있다.

[9] 대표적인 예로서, 공개 서열 데이터베이스에 타겟 핵산분자의 핵산서열로서 등재되어 있는 서열 정보들을 가져와 해당 타겟 핵산분자의 검출을 위한 올리고뉴클레오타이드의 디자인에 활용하는 기술이 있다. 공개 서열 데이터베이스에는 연구기관이나 기업체, 개인 등에 의해 수득된 실험결과로서 특정 생명체의 유전체 서열이나 유전자 서열, 해당 생명체의 숙주(host) 등에 대한 정보가 등재될 수 있다. 이처럼 공개 서열 데이터베이스에 등재된 서열 정보들은 특정한 숙주를 대상으로 감염을 일으키는 타겟 병원체의 검출에 이용되는 올리고뉴클레오타이드의 디자인 과정에 활용될 수 있다.

[10] 일 예로서, 설사의 원인이 되는 Rotavirus를 검출하기 위한 올리고뉴클레오타이드를 디자인하는 과정에서, Rotavirus 과에 속하는 Rotavirus A종 내지 H종 중에서 Homo sapiens 대상으로 감염을 일으키는 Rotavirus A종 내지 C종의 서열 정보들이 선택적으로 이용될 수 있다. 이러한 정보들을 이용해서 디자인된 올리고뉴클레오타이드는 사람에게서 채취된 샘플 내에 Rotavirus의 유전 물질이 존재하는지 여부를 검출하기 위한 분자 진단 시약으로서 효율적으로 기능할 수 있다.

[11] 그러나, 이처럼 올리고뉴클레오타이드의 디자인을 위해 타겟 핵산분자의 핵산서열에 대한 충분한 정보가 제공되더라도, 그 중에 부정확한 핵산서열 정보가 포함되어 있는 경우, 올리고뉴클레오타이드의 디자인이 상대적으로 부정확하게 이루어질 수 있으며, 이로 인해 타겟 핵산분자의 정확한 검출에 실패할 가능성이 커지게 된다.

[12] 공개 서열 데이터베이스에는 실험 오류나 기재 오류와 같은 human error 등으로 인해 부정확한 정보가 등록될 가능성이 상존하는 반면, 이러한 오류를 검증하기 위한 기술적 수단이 미비한 문제점이 있다. 특히, 핵산 서열과 함께 등재되는 숙주 개체 정보의 경우, 오류 여부를 검증하기 어려워 데이터 신뢰성의 저하로 이어지는 문제점이 있다. 또한, 핵산 서열과 함께 등재되는 병원체 개체 정보에 오류가 있을 경우, 올리고뉴클레오타이드 디자인의 정확성이 크게 저하됨으로 인해 타겟 핵산분자의 정확한 검출에 실패할 가능성이 더욱 커지게 되는 문제점이 있다.

[13] 상술한 예시에서, Rotavirus A종 내지 C종의 핵산서열인 것으로 등재되어 있는 정보들 중에, 만일 등록자의 기재 오류로 인해 Sus scrofa를 대상으로 감염을 일으키는 Rotavirus E종의 서열이 포함되어 있다면, 불필요한 타겟 검출을 위해 더 많은 올리고뉴클레오타이드가 소모되며, 이로 인해 분자 진단 시약의 정확성이 감소될 뿐만 아니라, 올리고뉴클레오타이드의 효율성이 감소하는 문제점이 있다.

발명의 상세한 설명

기술적 과제

[14] 일 실시예에 따라 해결하고자 하는 과제는, 상술한 문제점을 해결하기 위한 것으로, 핵산 서열이 주어졌을 때 해당 핵산 서열을 포함하는 유기체가 어떤 개체인지 또는 해당 유기체의 host가 어떤 개체인지 추정할 수 있는 방법을 제공하는 것을 포함한다.

[15] 또한, 이러한 유기체 또는 host 추정을 통해 유기체 또는 host 정보가 검증된 핵산 서열이 분자 진단 시약의 개발에서 이용되도록 하기 위한 방법이 이러한 과제에 포함될 수 있다.

[16] 다만, 본 발명이 해결하고자 하는 과제는 이상에서 언급한 것으로 제한되지 않으며, 언급되지 않은 또 다른 해결하고자 하는 과제는 아래의 기재로부터 본 발명이 속하는 통상의 지식을 가진 자에게 명확하게 이해될 수 있을 것이다.

과제 해결 수단

[17] 본 개시의 일 실시예에 따른 컴퓨터 구현 방법은, 메모리, 프로세서 및 상기 메모리에 저장되고 상기 프로세서에 의해 실행되도록 구성된 하나 이상의 프로그램을 사용하는 컴퓨터 장치에 의해 수행되며, 방법은, 사전 학습된 모델을 fine-tuning하여서 획득된 추정 모델에 접근하는 단계; 상기 추정 모델에 핵산 서열을 제공하는 단계; 및 상기 추정 모델로부터, 상기 핵산 서열을 포함하는 유기체 또는 상기 유기체의 host를 추정하는 단계를 포함할 수 있다.

[18] 일 실시예에서, 상기 추정에서는, 상기 유기체 또는 상기 host의 생물학적 카테고리로서, 생물학적 분류 체계를 구성하는 복수개의 hierarchical level 중 어느 하나의 hierarchical level에 위치하는 카테고리가 추정될 수 있다.

[19] 일 실시예에서, 상기 어느 하나의 hierarchical level은, 상기 생물학적 분류 체계에서의 종(species) 레벨일 수 있다.

[20] 일 실시예에서, 상기 사전 학습된 모델은, 학습용 데이터로서 복수개의 핵산 서열이 이용된 것을 특징으로 할 수 있다.

[21] 일 실시예에서, 상기 사전 학습된 모델은 상기 복수개의 핵산 서열 각각에 포함된 베이스들 중 일부의 베이스에 마스크(mask)를 적용한 뒤, 마스킹된(masked) 베이스를 맞추는 semi-supervised learning 방식에 의해 학습된 것일 수 있다.

[22] 일 실시예에서, 상기 사전 학습된 모델은, 각각 두 개 이상의 베이스들을 갖는 토큰들로 토큰화되는 핵산 서열을 이용해서 학습된 것일 수 있다.

[23] 일 실시예에서, 상기 토큰들은 (i) 상기 핵산 서열을 k 개(상기 k는 자연수)씩 분할하거나 (ii) 상기 핵산 서열을 기능 단위로 분할하여서 토큰화되는 베이스들을 각각 포함할 수 있다.

[24] 일 실시예에서, 상기 fine-tuning은 복수의 학습용 데이터 세트를 이용해서 수행되되, 각 학습용 데이터 세트는 (i) 핵산 서열을 포함하는 학습용 입력 데이터 및 (ii) 해당 핵산 서열을 포함하는 유기체 또는 해당 유기체의 host에 대한 라벨 데이터를 포함하는 학습용 정답 데이터를 포함할 수 있다.

[25] 일 실시예에서, 상기 fine-tuning은, (i) 상기 학습용 입력 데이터에 포함된 핵산 서열을 토큰화(tokenization)하여 복수의 토큰들을 획득하는 과정, (ii) 상기 복수의 토큰들로부터 생성되는 컨텍스트 벡터를 이용해서 상기 학습용 입력 데이터에 포함된 핵산 서열의 유기체 또는 host에 대해 추정하는 과정 및 (iii) 상기 추정된 결과와 상기 학습용 정답 데이터 간의 차이가 줄어들도록 상기 사전 학습된 모델을 훈련시키는 과정을 포함할 수 있다.

[26] 일 실시예에서, 상기 추정 모델에 제공되는 상기 핵산 서열은, 베이스들의 개수가 사전설정된 제1 cutoff 이상 또는 제2 cutoff 이하일 수 있다.

[27] 일 실시예에서, 상기 방법은, 상기 핵산 서열에 포함된 베이스들의 개수가 소정의 제2 cutoff를 초과하는 경우, 상기 제2 cutoff 개수 이하의 베이스들이 포함되도록 상기 핵산 서열로부터 하나 이상의 부분 서열을 획득하는 단계를 더 포함하며, 상기 핵산 서열을 제공하는 단계에서 상기 추정 모델에는, 상기 하나 이상의 부분 서열이 제공될 수 있다.

[28] 일 실시예에서, 상기 핵산 서열이 제공되는 단계에서 상기 추정 모델에는, 사전설정된 시작 지점을 기준으로 상기 제2 cutoff 개수 이하의 베이스들을 포함하는 하나의 부분 서열이 제공되고 상기 하나의 부분 서열을 제외한 나머지 서열은 제공되지 않거나, 또는 상기 제2 cutoff 개수 이하의 베이스들을 포함하는 복수개의 부분 서열 각각이 제공될 수 있다.

[29] 일 실시예에서, 상기 추정하는 단계에서는, 상기 부분 서열이 복수개인 경우, 상기 복수개의 부분 서열 각각에 대한 유기체 또는 host가 추정되고, 상기 방법은, 상기 복수개의 부분 서열 각각에 대한 유기체 또는 host를 통계 처리하는 단계를 더 포함하며, 상기 추정된 유기체 또는 상기 추정된 host는, 상기 통계 처리한 결과를 이용해서 획득된 것일 수 있다.

[30] 일 실시예에서, 상기 통계 처리에는, Majority vote 방식, 평균 방식 및 표준 편차 방식 중 적어도 하나가 포함될 수 있다.

[31] 일 실시예에서, 상기 host를 대상으로 하는 분자 진단용 시약 개발에는, (i) 상기 추정된 유기체 또는 상기 추정된 host 및 (ii) 상기 핵산 서열이 이용될 수 있다.

[32] 일 실시예에서, 상기 분자 진단용 시약의 개발에서는, 상기 유기체의 검출에 이용되는 프라이머 및 프로브 중 적어도 하나가 개발될 수 있다.

[33] 일 실시예에서, 상기 핵산 서열을 제공하는 단계에서는, 상기 핵산 서열과 상기 유기체 또는 상기 host에 대한 정보를 포함하는 서열 관련 정보로부터 상기 핵산

서열을 획득하고, 상기 방법은, 상기 유기체에 대한 정보와 상기 추정된 유기체 간의 비교 결과 또는 상기 host에 대한 정보와 상기 추정된 host 간의 비교 결과를 획득하는 단계를 더 포함할 수 있다.

[34] 일 실시예에서, 상기 핵산 서열을 제공하는 단계에서는, 상기 핵산 서열과 상기 유기체 또는 상기 host에 대한 정보를 포함하는 서열 관련 정보로부터 상기 핵산 서열을 획득하고, 상기 방법은, 상기 유기체 또는 상기 host에 대한 정보가 상기 추정된 유기체 또는 host와 상이하면, 상기 유기체 또는 상기 host에 대한 정보가 상기 추정된 유기체 또는 host로 수정되도록 제어하는 단계를 더 포함할 수 있다.

[35] 일 실시예에서, 상기 추정 모델에 제공되는 상기 핵산 서열은, 베이스들의 개수가 사전설정된 제1 cutoff 이상 또는 제2 cutoff 이하이고, 상기 방법은, 상기 추정된 유기체 또는 host의 개수가 사전설정된 제1 개수 이상 또는 제2 개수 이하인 경우, 상기 제1 cutoff 또는 상기 제2 cutoff를 상이한 값으로 갱신하는 단계를 더 포함할 수 있다.

[36] 일 실시예에서, 상기 갱신하는 단계에서는, 상기 추정된 유기체 또는 host의 개수가 상기 제1 개수 이상인 경우에는 상기 제1 cutoff를 더 큰 값으로 갱신하고, 상기 제2 개수 이하인 경우에는 상기 제2 cutoff를 더 작은 값으로 갱신하는 것을 특징으로 할 수 있다.

[37] 일 실시예에서, 상기 방법은, 상기 추정된 유기체 또는 host의 개수가 기 설정된 기준 개수 이상인 경우, 상기 핵산 서열이 상기 유기체 또는 상기 host를 대상으로 하는 분자 진단용 시약 개발에 이용되지 않도록 제어하는 단계를 더 포함하는 것을 특징으로 할 수 있다.

[38] 본 개시의 다른 일 실시예에 따른 컴퓨터 판독 가능한 기록 매체에 저장되어 있는 컴퓨터 프로그램은, 상기 방법에 포함된 각 단계를 수행하도록 프로그램될 수 있다.

[39] 본 개시의 또 다른 일 실시예에 따른 컴퓨터 판독 가능한 기록 매체는, 상기 방법에 포함된 각 단계를 수행하도록 프로그램된 컴퓨터 프로그램이 저장되어 있을 수 있다.

[40] 본 개시의 또 다른 일 실시예에 따른 컴퓨터 장치는, 적어도 하나의 명령어를 저장하는 메모리; 및 프로세서를 포함하며, 상기 프로세서는 상기 적어도 하나의 명령어를 실행시킴으로써, 사전 학습된 모델을 fine-tuning하여서 획득된 추정 모델에 접근하고, 상기 추정 모델에 핵산 서열을 제공하고, 상기 추정 모델로부터, 상기 핵산 서열을 포함하는 유기체 또는 상기 유기체의 host를 추정할 수 있다.

[41] 본 개시의 도 다른 일 실시예에 따른 컴퓨터 구현 방법은, 메모리, 프로세서 및 상기 메모리에 저장되고 상기 프로세서에 의해 실행되도록 구성된 하나 이상의 프로그램을 사용하는 컴퓨터 장치에 의해 수행되며, 상기 컴퓨터 구현 방법은, 사전 학습된 모델을 fine-tuning하여서 획득된 추정 모델에 접근하는 단계; 상기 추정 모델에 핵산 서열을 제공하는 단계; 및 상기 추정 모델로부터, 상기 핵산 서열을 포함하는 유기체 또는 상기 유기체의 host를 추정하는 단계를 포함하고, 상

기 사전 학습된 모델은 학습용 데이터로서 복수개의 핵산 서열을 이용하고, 상기 fine-tuning은 복수의 학습용 데이터 세트를 이용해서 수행되되, 각 학습용 데이터 세트는 (i) 핵산 서열을 포함하는 학습용 입력 데이터 및 (ii) 해당 핵산 서열의 유기체 또는 host에 대한 라벨 데이터를 포함하는 학습용 정답 데이터를 포함할 수 있다.

[42] 일 실시예에서, 상기 fine-tuning은, (i) 상기 학습용 입력 데이터에 포함된 핵산 서열을 토큰화(tokenization)하여 복수의 토큰들을 획득하는 과정, (ii) 상기 복수의 토큰들로부터 생성되는 컨텍스트 벡터를 이용해서 상기 학습용 입력 데이터에 포함된 핵산 서열의 유기체 또는 host에 대해 추정하는 과정 및 (iii) 상기 추정된 결과와 상기 학습용 정답 데이터 간의 차이가 줄어들도록 상기 사전 학습된 모델을 훈련시키는 과정을 포함할 수 있다.

[43] 본 개시의 또 다른 일 실시예에 따른 컴퓨터 구현 방법은, 메모리, 프로세서 및 상기 메모리에 저장되고 상기 프로세서에 의해 실행되도록 구성된 하나 이상의 프로그램을 사용하는 컴퓨터 장치에 의해 수행되며, 상기 컴퓨터 구현 방법은, 사전 학습된 모델을 획득하는 단계; 및 상기 사전 학습된 모델을 fine-tuning하여서, 핵산 서열이 제공되면 상기 핵산 서열을 포함하는 유기체 또는 상기 유기체의 host를 추정하도록 학습된 추정 모델을 획득하는 단계를 포함하고, 상기 fine-tuning은 복수의 학습용 데이터 세트를 이용해서 수행되되, 각 학습용 데이터 세트는 (i) 핵산 서열을 포함하는 학습용 입력 데이터 및 (ii) 해당 핵산 서열의 유기체 또는 host에 대한 라벨 데이터를 포함하는 학습용 정답 데이터를 포함할 수 있다.

[44] 일 실시예에서, 상기 라벨 데이터는, 상기 유기체 또는 상기 host의 생물학적 카테고리로서, 생물학적 분류 체계를 구성하는 복수개의 hierarchical level 중 어느 하나의 hierarchical level에 위치하는 카테고리에 대한 라벨 데이터일 수 있다.

[45] 일 실시예에서, 상기 사전 학습된 모델은, 학습용 데이터로서 복수개의 핵산 서열이 이용된 것을 특징으로 할 수 있다.

[46] 일 실시예에서, 상기 사전 학습된 모델은 상기 복수개의 핵산 서열 각각에 포함된 베이스들 중 일부의 베이스에 마스크(mask)를 적용한 뒤, 마스킹된(masked) 베이스를 맞추는 semi-supervised learning 방식에 의해 학습된 것일 수 있다.

[47] 일 실시예에서, 상기 사전 학습된 모델은, 각각 두 개 이상의 베이스들을 갖는 토큰들로 토큰화되는 핵산 서열을 이용해서 학습된 것일 수 있다.

[48] 일 실시예에서, 상기 토큰들은 (i) 상기 핵산 서열을 k 개(상기 k는 자연수)씩 분할하거나 (ii) 상기 핵산 서열을 기능 단위로 분할하여서 토큰화되는 베이스들을 각각 포함할 수 있다.

[49] 일 실시예에서, 상기 fine-tuning은, (i) 상기 학습용 입력 데이터에 포함된 핵산 서열을 토큰화(tokenization)하여 복수의 토큰들을 획득하는 과정, (ii) 상기 복수의 토큰들로부터 생성되는 컨텍스트 벡터를 이용해서 상기 학습용 입력 데이터에 포함된 핵산 서열의 유기체 또는 host에 대해 추정하는 과정 및 (iii) 상기 추정

된 결과와 상기 학습용 정답 데이터 간의 차이가 줄어들도록 상기 사전 학습된 모델을 훈련시키는 과정을 포함할 수 있다.

[50] 일 실시예에서, 상기 추정 모델을 획득하는 단계에서, 상기 추정 모델은, 상기 핵산 서열에 포함된 베이스들의 개수가 소정의 제2 cutoff를 초과하는 경우, 상기 제2 cutoff 개수 이하의 베이스들이 포함되도록 상기 핵산 서열로부터 획득되는 하나 이상의 부분 서열이 제공되면 상기 하나 이상의 부분 서열의 유기체 또는 host를 추정하도록 학습될 수 있다.

[51] 일 실시예에서, 상기 추정 모델을 획득하는 단계에서, 상기 추정 모델은, 상기 부분 서열이 복수개인 경우, 상기 복수개의 부분 서열 각각의 유기체 또는 host를 추정하고, 상기 복수개의 부분 서열 각각의 유기체 또는 host를 통계 처리하고, 상기 통계 처리한 결과를 이용해서 상기 핵산 서열의 유기체 또는 host를 추정하도록 학습될 수 있다.

발명의 효과

[52] 본 개시의 일 실시예에 따르면, 핵산 서열을 이용해서 해당 핵산 서열을 포함하는 유기체 또는 해당 유기체의 host가 어떤 개체인지 정확하게 추정될 수 있다.

[53] 아울러, 전이학습 방식으로 학습된 추정 모델을 이용함에 따라, 적은 양의 라벨링되어 있는 학습용 데이터를 이용하여 학습하더라도 충분한 추정 정확성이 확보될 수 있다. 또한, 타겟으로 하는 유기체나 해당 유기체의 host에 관한 다른 추가적인 정보가 없더라도, 핵산 서열만을 이용하여 해당 핵산 서열을 포함하는 유기체 또는 해당 유기체의 host가 어떤 개체인지 정확하게 추정될 수 있다.

[54] 뿐만 아니라, 유기체 또는 host 추정을 통해 유기체 또는 host 정보가 검증된 핵산 서열이 분자 진단 시약의 개발에 활용될 수 있다. 예를 들면, 특정 host를 대상으로 하는 분자 진단용 시약 개발에, 해당 host에 대한 정보가 올바르게 기재되어 있는 핵산 서열이 이용될 수 있다. 다른 예를 들면, 특정 유기체의 검출을 위한 분자 진단용 시약 개발에, 해당 유기체에 대한 정보가 올바르게 기재되어 있는 핵산 서열이 이용될 수 있다. 이와 같이, 등재된 유기체 또는 host가 부정확한 경우에는 해당 핵산 서열이 분자 진단 시약의 개발에서 배제됨으로써, 올리고뉴클레오타이드 디자인의 효율성이 향상될 수 있다. 이는 타겟 핵산에 대한 검출 정확성 향상으로 이어질 수 있다.

[55] 본 개시의 효과는 상기한 효과로 한정되는 것은 아니며, 본 개시의 상세한 설명 또는 특허청구범위에 기재된 발명의 구성으로부터 추론 가능한 모든 효과를 포함하는 것으로 이해되어야 한다.

도면의 간단한 설명

[56] 도 1은 일 실시예에 따른 컴퓨터 장치의 블록 구성도를 개략적으로 도시한다.

[57] 도 2는 일 실시예에 따른 사전 학습의 과정에 대한 개념도를 예시적으로 도시한다.

[58] 도 3은 일 실시예에 따른 사전 학습에 이용되는 BERT 기반의 언어 모델의 구조 및 동작을 예시적으로 도시한다.

[59] 도 4는 일 실시예에 따른 컴퓨터 장치(100)에서 사전 학습된 모델에 의해 마스킹된 베이스의 베이스별 확률값이 추정되는 예시적인 방식을 도시한다.

[60] 도 5는 일 실시예에 따른 fine-tuning의 과정에 대한 개념도를 예시적으로 도시한다.

[61] 도 6은 일 실시예에 따른 생물학적 분류 체계와 생물학적 카테고리에 대한 개념도를 예시적으로 도시한다.

[62] 도 7은 일 실시예에 따른 fine-tuning의 과정에서 BERT 기반의 추정 모델의 구조 및 동작을 예시적으로 도시한다.

[63] 도 8은 일 실시예에 따라 핵산 서열이 부분 서열로 전처리되는 과정을 예시적으로 도시한다.

[64] 도 9는 일 실시예에 따라 추정 모델을 획득하기 위한 예시적인 흐름도를 도시한다.

[65] 도 10은 일 실시예에 따라 추정 모델의 추론(inference) 동작에 대한 개념도를 예시적으로 도시한다.

[66] 도 11은 일 실시예에 따라 추정 모델에서 카테고리별 확률값을 추정하는 예시적인 방식을 도시한다.

[67] 도 12는 일 실시예에 따라 유기체 또는 host를 추정하기 위한 예시적인 흐름도를 도시한다.

[68] 도 13은 일 실시예에 따른 컴퓨터 장치가 추정 모델에 의한 추정 결과가 분자 진단용 시약 개발에 이용되도록 하는 과정을 예시하는 순서도이다.

[69] 도 14는 일 실시예에 따른 컴퓨터 장치가 추정 모델에 의한 추정 결과가 분자 진단용 시약 개발에 이용되도록 제어하는 예시적인 방식을 도시한다.

발명의 실시를 위한 형태

[70] 다양한 실시예들이 도면을 참조하여 설명된다. 본 명세서에서, 다양한 설명들이 본 개시내용의 이해를 제공하기 위해서 제시된다. 본 개시내용의 실시를 위한 구체적인 내용을 설명하기에 앞서, 본 개시내용의 기술적 요지와 직접적 관련이 없는 구성에 대해서는 본 발명의 기술적 요지를 흩뜨리지 않는 범위 내에서 생략하였음에 유의하여야 할 것이다. 또한, 본 명세서 및 청구범위에 사용된 용어 또는 단어는 발명자가 자신의 발명을 최선의 방법으로 설명하기 위해 적절한 용어의 개념을 정의할 수 있다는 원칙에 입각하여 본 발명의 기술적 사상에 부합하는 의미와 개념으로 해석되어야 할 것이다.

[71] 본 명세서에서 사용되는 용어 "또는"은 배타적 "또는"이 아니라 내포적 "또는"을 의미하는 것으로 의도된다. 즉, 달리 특정되지 않거나 문맥상 명확하지 않은 경우에, "X는 A 또는 B를 이용한다"는 자연적인 내포적 치환 중 하나를 의미하는 것으로 의도된다. 또한, 본 명세서에 사용된 "및/또는"이라는 용어는 열거된

관련 아이템들 중 하나 이상의 아이템의 가능한 모든 조합을 지칭하고 포함하는 것으로 이해되어야 한다.

[72] 또한, "포함한다" 및/또는 "포함하는"이라는 용어는, 해당 특징 및/또는 구성요소가 존재함을 의미하는 것으로 이해되어야 한다. 다만, "포함한다" 및/또는 "포함하는"이라는 용어는, 하나 이상의 다른 특징, 구성요소 및/또는 이들의 그룹의 존재 또는 추가를 배제하지 않는 것으로 이해되어야 한다. 또한, 달리 특정되지 않거나 단수 형태를 지시하는 것으로 문맥상 명확하지 않은 경우에, 본 명세서와 청구범위에서 단수는 일반적으로 "하나 또는 그 이상"을 의미하는 것으로 해석되어야 한다.

[73] 또한, 용어 "제공"은 대상이 특정 정보를 획득하거나 직간접적으로 특정 대상에게 송수신하는 과정을 포함하며 이러한 과정에서 요구되는 관련 동작의 수행을 포괄적으로 포함하는 것으로 해석될 수 있다.

[74] 제시된 실시예들에 대한 설명은 본 개시의 기술 분야에서 통상의 지식을 가진 자가 본 발명을 이용하거나 또는 실시할 수 있도록 제공된다. 이러한 실시예들에 대한 다양한 변형들은 본 개시의 기술 분야에서 통상의 지식을 가진 자에게 명백할 것이다. 여기에 정의된 일반적인 원리들은 본 개시의 범위를 벗어남이 없이 다른 실시예들에 적용될 수 있다. 그리하여, 본 발명은 여기에 제시된 실시예들로 한정되는 것이 아니다. 본 발명은 여기에 제시된 원리들 및 신규한 특징들과 일관되는 최광의의 범위에서 해석되어야 할 것이다.

[75] 도 1을 설명하기에 앞서, 본원에서 사용된 일부 용어들에 대해 살펴보기로 한다.

[76] 본 명세서에서 용어 "핵산 서열(nucleic acid sequence)"은 타겟 분석물(예컨대, 타겟 핵산분자)을 특정 핵산 서열로 나타낸 것이다. 또한, 핵산 서열은 뉴클레오타이드의 구성 성분 중 하나인 베이스(base)들을 순서대로 나열해 놓은 것을 의미한다. 일례로, 본 개시내용에서의 핵산 서열은 베이스 서열과 상호 교환가능하게 사용될 수 있다. 핵산 서열을 구성하는 개별적인 베이스들 각각은 예를 들어, A, G, C 및 T의 4종류의 베이스들 중 하나와 대응될 수 있다.

[77] 본 명세서에서 용어 "분석물(analyte)"은 다양한 물질(예컨대, 생물학적 물질 및 비생물학적 물질)을 지칭할 수 있다. 이러한 타겟 분석물은 구체적으로 생물학적 물질, 보다 구체적으로 핵산분자(예컨대, DNA 및 RNA), 단백질, 펩타이드, 탄수화물, 지질, 아미노산, 생물학적 화합물, 호르몬, 항체, 항원, 대사물질 및 세포 중 적어도 하나를 포함할 수 있다.

[78] 본 명세서에서 용어 "타겟 분석물"(target analyte) 또는 "유기체"(organism)는 분석, 획득 또는 검출하고자 하는 임의의 형태의 생명체 또는 유기체를 포함할 수 있다. 예를 들어, 유기체는 하나의 속, 종, 아종, 서브타입, 지노타입, 시로타입, 스트레인, 분리종(isolate) 또는 재배종(cultivar)에 속한 생명체를 의미할 수 있다. 본 개시내용에서의 유기체와 타겟 분석물은 서로 상호 교환가능하게 사용될 수 있다.

[79] 유기체는 원핵세포(예컨대, Mycoplasma pneumoniae, Chlamydophila 등), 진핵세포(예컨대, 원생동물과 기생동물, 균류, 효모, 고등 식물, 하등 동물 및 포유동물과 인간(human)을 포함하는 고등동물), 바이러스 또는 비로이드를 포함할 수 있다. 상기 진핵세포 중 기생충(parasite)은 예를 들어, Giardia lamblia, Entamoeba histolytica 등을 포함할 수 있다. 바이러스는 예를 들어, 호흡기 질환을 유발하는 인플루엔자 A 바이러스(Flu A), 인플루엔자 B 바이러스(Flu B), 호흡 씬시티얼 바이러스 A(Respiratory syncytial virus A), 호흡 씬시티얼 바이러스 B(Respiratory syncytial virus B), 코비드(Covid)-19 바이러스, 파라인플루엔자 바이러스 1(PIV 1) 내지 4(PIV 4), 인간 라이노바이러스(HRV), 코로나바이러스 및 아데노바이러스 등을 포함할 수 있다. 상기 바이러스는 그 밖의 다양한 공지된 바이러스들을 포함할 수 있으며, 전술한 예시들로 제한되지 않는다.

[80] 본 개시내용에서의 유기체는 전술한 바이러스뿐만 아니라 박테리아, human 등의 다양한 분석 대상물들을 포함할 수 있으며, 전술한 예시들로 유기체에 대한 범위는 제한되지 않는다.

[81] 본 개시내용에서의 "핵산 서열을 포함하는 유기체"는 핵산 서열이 주어져있을 때, 해당 핵산 서열이 나타내는 유기체를 의미한다. 일 예로, 특정 유기체의 핵산 분자에 포함된 뉴클레오타이드들의 구성 성분 중 하나인 베이스들을 순서대로 나열한 것과 주어진 핵산 서열에 따라 베이스들을 나열한 것이 대응되는 경우, 해당 특정 유기체는 해당 핵산 서열을 포함하는 유기체인 것으로 볼 수 있다. 다른 예로서, 핵산 서열을 포함하는 유기체는 해당 핵산 서열에 대응되는 유전 물질을 갖는 유기체(예: 병원균)를 의미한다. 이러한 핵산 서열을 포함하는 유기체의 종류는 실시예에 따라서 하나 이상일 수 있으며, 예컨대, 종류에 대응되는 생물학적 카테고리가 생물학적 분류 체계 상에서의 어떤 hierarchical level에 위치하도록 설정되는지에 따라서 상이한 명칭 및/또는 개수로 표현될 수 있다.

[82] 본 개시내용에서의 "생물학적 분류 체계(biological taxonomy hierarchy)"는 유기체가 속하는 범위를 구분하기 위한 분류 체계이다. 일 실시예에 따른 생물학적 분류 체계는 복수개의 hierarchical level로 구성될 수 있다. 예를 들어, 종(Species), 속(Genus), 과(Family), 목(Order), 강(Class), 문(Phylum), 계(Kingdom) 또는 역(Domain) 등으로 표현되는 복수개의 hierarchical level이 본 개시내용에서의 생물학적 분류 체계의 범위 내에 포함될 수 있다. 일례로, 본 개시내용의 일 실시예에 따른 생물학적 분류 체계는 계층 구조(hierarchical structure)를 가질 수 있다. 여기서, 계층 구조는 상위 hierarchical level이 하위 hierarchical level을 포괄하는 형태의 생물학적 계통 구조인 계층 구조를 의미할 수 있다.

[83] 계층 구조에서 상위 hierarchical level은 하위 hierarchical level에 포함된 구성요소들을 모두 포괄할 수 있다. 예를 들어, A, B, 및 C의 구성요소들을 포함하는 하위 hierarchical level에 대한 상위 hierarchical level은, 적어도 A, B, 및 C의 구성요소들을 포함하되 추가적인 구성요소를 더 포함할 수도 있다.

[84] 본 개시내용에서의 "분자 진단용 시약"는 타겟 분석물의 검출에 이용되는 시약을 의미한다. 예컨대, 분자 진단용 시약은 타겟 분석물 또는 타겟 분석물의 존재를 나타내는 신호를 증폭시키기 위한 증폭 반응에 이용되는 하나 이상의 올리고뉴클레오타이드를 포함할 수 있다.

[85] 본 명세서에서 올리고뉴클레오타이드(oligonucleotide)는 자연의 또는 변형된 모노머 또는 연쇄(linkages)의 선형 올리고머를 의미하며, 데옥시리보뉴클레오타이드 및 리보뉴클레오타이드를 포함하고 타겟 핵산서열에 특이적으로 혼성화할 수 있으며, 자연적으로 존재하거나 또는 인위적으로 합성될 수 있다. 본 발명의 올리고뉴클레오타이드는 자연(naturally occurring) dNMP(즉, dAMP, dGMP, dCMP 및 dTMP), 뉴클레오타이드 유사체 또는 유도체를 포함할 수 있다. 특히, 올리고뉴클레오타이드는 데옥시리보뉴클레오타이드로 이루어진 단일 가닥이다. 올리고뉴클레오타이드는 타겟 핵산 서열에 의존적으로 발생하는 절단 단편과 혼성화되는 올리고뉴클레오타이드를 포함한다. 구체적으로, 상기 올리고뉴클레오타이드는 프라이머 및/또는 프로브를 포함한다.

[86] 본 명세서에 용어 "프라이머"는 핵산 가닥(주형)에 상보적인 프라이머 연장 산물의 합성이 유도되는 조건, 즉, 뉴클레오타이드와 DNA 중합효소와 같은 중합제의 존재, 그리고 적합한 온도와 pH의 조건에서 합성의 개시점으로 작용할 수 있는 올리고뉴클레오타이드를 의미한다.

[87] 본 명세서에서 용어"프로브(probe)"는 타겟 핵산서열에 상보적인 부위 또는 부위들을 포함하는 단일-가닥 핵산 분자를 의미한다. 또한, 상기 프로브는 타겟 검출을 위한 신호를 발생시킬 수 있는 표지를 포함할 수 있다.

[88] 한편, 일 실시예에 따른 분자 진단용 시약은 타겟 분석물의 검출에 이용되되, 정해진 host(숙주)를 대상으로 하는 시약일 수 있다. 예컨대, 분자 진단용 시약은 host로부터 채취된 샘플(예: 세포, 혈액, 타액, 스왑 등)과 함께 반응용기에 수용됨에 따라, 이후 과정에서 타겟 분석물 또는 타겟 분석물의 존재를 나타내는 신호의 증폭 반응이 진행되도록 하여 타겟 분석물의 검출에 이용될 수 있다. 일 예로, 설사의 원인이 되는 Rotavirus를 검출하기 위한 분자 진단용 시약은 host인 사람에게서 채취된 샘플 내에 Rotavirus의 유전 물질이 존재하는지 여부를 검출하는 데 이용되는 프라이머 및/또는 프로브를 포함할 수 있다.

[89] 본 명세서에서 용어 "host"는 유기체의 host에 해당하는 개체를 의미한다. 일 실시예에 따른 host가 의미하는 바는 핵산 서열과 연관되어 정의될 수 있으며, 예컨대, 어떠한 핵산 서열의 host는 해당 핵산 서열을 포함하는 유기체의 host를 의미한다.

[90] 일 실시예에 따른 host는 해당 핵산 서열이 채취된 객체를 의미할 수 있다. 일 예로, host는 타겟 분석물의 핵산 서열이 포함된 샘플을 제공한 대상을 의미하며, 연구자에 의해 샘플이 분석되어 해당 핵산 서열에 대한 정보가 획득됨에 따라 해당 타겟 분석물의 핵산 서열, 해당 타겟 분석물(유기체)에 대한 정보 및 해당 host에 대한 정보가 함께 기록될 수 있다. 예컨대, 사람의 비인두로부터 채취된 샘플에

서 Rotavirus A종의 핵산 서열이 검출된 경우, 해당 핵산 서열의 타겟 분석물(유기체)는 Ratavirus이고, 해당 핵산 서열의 host는 사람(예: Homo sapiens)일 것이다.

[91] 다른 일 실시예에 따른 host는 해당 핵산 서열을 포함하는 유기체의 host로 알려진 개체를 포괄적으로 의미할 수 있다. 일 예로, host는 해당 핵산 서열을 포함하는 유기체와 기생 또는 공생 관계에 있을 수 있는 상대의 생명체로서 공지된 개체일 수 있다. 다른 일 예로, host는 해당 핵산 서열을 포함하는 유기체의 host에 해당하는 것으로 등록자에 의해 기록되어 있는 개체일 수 있다.

[92] 본 개시의 일 실시예에서의 host는 유기체가 기생 또는 공생할 수 있는 임의의 생명체로서, 예컨대, 사람, 식물 및 동물 등일 수 있으나, 이에 제한되지 않는다. 실시예에 따라서, host가 될 수 있는 생명체는 상술한 유기체의 다양한 예시들 중 일부를 포함할 수 있다.

[93] 도 1은 일 실시예에 따른 컴퓨터 장치(100)의 블록 구성도를 개략적으로 도시한다.

[94] 도 1을 참조하면, 본 개시의 일 실시예에 따른 컴퓨터 장치(100)는 메모리(110), 통신부(120)및 프로세서(130)를 포함할 수 있다.

[95] 도 1에 도시된 컴퓨터 장치(100)의 구성은 간략화하여 나타낸 예시일 뿐이다. 일 실시예에서, 컴퓨터 장치(100)는 컴퓨터 장치(100)의 컴퓨팅 환경을 수행하기 위한 다른 구성들이 포함될 수 있고, 개시된 구성들 중 일부만이 컴퓨터 장치(100)를 구성할 수도 있다.

[96] 컴퓨터 장치(100)는 본 개시내용의 실시예들을 구현하기 위한 시스템을 구성하는 노드를 의미할 수 있다. 컴퓨터 장치(100)는 임의의 형태의 사용자 단말 또는 임의의 형태의 서버를 의미할 수 있다. 전술한 컴퓨터 장치(100)의 컴포넌트들은 예시적인 것으로 일부가 제외될 수 있거나 또는 추가 컴포넌트가 포함될 수도 있다. 일례로, 전술한 컴퓨터 장치(100)가 단말을 포함하는 경우, 출력부(미도시) 및 입력부(미도시)가 그 범위 내에 포함될 수 있다.

[97] 메모리(110)는 프로세서(130)에 의해 실행될 수 있는 적어도 하나의 명령어를 저장할 수 있다. 일 실시예에서, 메모리(110)는 프로세서(130)가 생성하거나 결정한 임의의 형태의 정보 및 컴퓨터 장치(100)가 수신한 임의의 형태의 정보를 저장할 수 있다. 일 실시예에서, 메모리(110)는 프로세서(130)가 본 개시의 실시예들에 따른 동작을 수행하도록 하는 컴퓨터 소프트웨어를 저장하는 저장매체 일 수 있다. 따라서, 메모리(110)는 본 개시내용의 실시예들을 수행하는 데 필요한 소프트웨어 코드, 코드의 실행 대상이 되는 데이터, 코드의 실행 결과를 저장하기 위한 컴퓨터 판독 매체들을 의미할 수 있다.

[98] 일 실시예에서, 메모리(110)는 임의의 타입의 저장 매체를 의미할 수 있다 예를 들어, 메모리(110)는 플래시 메모리 타입(flash memory type), 하드디스크 타입(hard disk type), 멀티미디어 카드 마이크로 타입(multimedia card micro type), 카드 타입의 메모리(예: SD 또는 XD 메모리 등), RAM(Random Access Memory), SRAM(Static Random Access Memory), ROM(Read-Only

Memory), EEPROM(Electrically Erasable Programmable Read-Only Memory), PROM(Programmable Read-Only Memory), 자기 메모리, 자기 디스크, 광디스크 중 적어도 하나의 타입의 저장매체를 포함할 수 있다. 컴퓨터 장치(100)는 인터넷 상에서 상기 메모리(110)의 저장 기능을 수행하는 웹 스토리지와 관련되어 동작할 수도 있다. 전술한 메모리에 대한 기재는 예시일 뿐, 본 개시내용에서의 메모리(110)는 전술한 예시들로 제한되지 않는다.

[99] 통신부(120)는 유선 및 무선 등과 같은 그 통신 양태를 가리지 않고 구성될 수 있으며, 단거리 통신망(PAN: Personal Area Network), 근거리 통신망(WAN: Wide Area Network) 등 다양한 통신망으로 구성될 수 있다. 또한, 통신부(120)는 공지의 월드와이드웹(WWW: World Wide Web) 기반으로 동작할 수 있으며, 적외선(IrDA: Infrared Data Association) 또는 블루투스(Bluetooth)와 같이 단거리 통신에 이용되는 무선 전송 기술을 이용할 수도 있다. 일례로, 통신부(120)는 본 개시의 일 실시예에 따른 기법을 수행하는데 필요한 데이터에 대한 송수신을 담당할 수 있다.

[100] 프로세서(130)는 메모리(110)에 저장된 적어도 하나의 명령어를 실행시킴으로써, 후술될 본 개시내용의 실시예들에 따른 기술적 특징들을 수행할 수 있다. 일 실시예에서, 프로세서(130)는 적어도 하나의 코어로 구성될 수 있으며, 컴퓨터 장치(100)의 중앙 처리 장치(CPU: central processing unit), 범용 그래픽 처리 장치 (GPGPU: general purpose graphics processing unit), 텐서 처리 장치(TPU: tensor processing unit) 등의 데이터 분석 및/또는 처리를 위한 프로세서를 포함할 수 있다.

[101] 프로세서(130)는 메모리(110)에 저장된 컴퓨터 프로그램을 판독하여 본 개시내용의 일 실시예에 따라, 추정 모델을 이용해서, 주어진 핵산 서열을 포함하는 유기체 및/또는 해당 유기체의 host를 추정할 수 있다. 여기서, 추정 모델은 핵산 서열을 이용해서 해당 핵산 서열의 유기체 및/또는 host를 추정하도록 학습된 인공지능 기반의 모델을 나타낸다.

[102] 일 실시예에 따른 프로세서(130)는 신경망의 학습을 위한 연산을 수행할 수 있다. 프로세서(130)는 딥러닝(DL: deep learning)에서 학습을 위한 입력 데이터의 처리, 입력 데이터에서의 피처 추출, 오차 계산, 역전파(backpropagation)를 이용한 신경망의 가중치 업데이트 등의 신경망의 학습을 위한 계산을 수행할 수 있다. 프로세서(130)의 CPU, GPGPU, 및 TPU 중 적어도 하나가 네트워크 함수의 학습을 처리할 수 있다. 예를 들어, CPU 와 GPGPU가 함께 네트워크 함수의 학습, 네트워크 함수를 이용한 데이터 분류를 처리할 수 있다. 또한, 본 개시의 일 실시예에서 복수개의 컴퓨팅 장치들의 프로세서들을 함께 사용하여 네트워크 함수의 학습, 네트워크 함수를 이용한 데이터 분류를 처리할 수도 있다. 또한, 본 개시의 일 실시예에 따른 컴퓨팅 장치에서 수행되는 컴퓨터 프로그램은 CPU, GPGPU 또는 TPU 실행가능 프로그램일 수 있다.

[103] 본 개시내용에서의 컴퓨터 장치(100)는 임의의 형태의 사용자 단말 및/또는 임의의 형태의 서버를 포함할 수 있다. 사용자 단말은 서버 또는 다른 컴퓨팅 장치와 상호작용 가능한 임의의 형태의 단말을 포함할 수 있다. 사용자 단말은 예를 들어, 휴대폰, 스마트 폰(smart phone), 노트북 컴퓨터(laptop computer), PDA(personal digital assistants), 슬레이트 PC(slate PC), 태블릿 PC(tablet PC) 및 울트라북(ultrabook)을 포함할 수 있다. 서버는 예를 들어, 마이크로프로세서, 메인프레임 컴퓨터, 디지털 프로세서, 휴대용 디바이스 및 디바이스 제어기 등과 같은 임의의 타입의 컴퓨팅 시스템 또는 컴퓨팅 장치를 포함할 수 있다.

[104] 이하에서는 먼저 인공지능 기반의 모델로서 추정 모델이 가질 수 있는 기본적인 구조의 실시예들에 대해 개략적으로 기술하였으며, 추정 모델의 구체적인 구조와 학습 방법 등에 대해서는 뒤에서 후술하도록 한다.

[105] 본 명세서에서의 추정 모델은 네트워크 함수, 인공신경망 및/또는 뉴럴 네트워크에 기반하여 동작하는 임의의 형태의 컴퓨터 프로그램을 의미할 수 있다. 본 명세서에 걸쳐, 모델, 신경망, 네트워크 함수, 뉴럴 네트워크(neural network)는 상호 교환 가능한 의미로 사용될 수 있다. 신경망은 하나 이상의 노드들이 하나 이상의 링크를 통해 상호 연결되어 신경망 내에서 입력 노드 및 출력 노드 관계를 형성한다. 신경망 내에서 노드들과 링크들의 개수 및 노드들과 링크들 사이의 연관관계, 링크들 각각에 부여된 가중치의 값에 따라, 신경망의 특성이 결정될 수 있다. 신경망은 하나 이상의 노드들의 집합으로 구성될 수 있다. 신경망을 구성하는 노드들의 부분 집합은 레이어(layer)를 구성할 수 있다.

[106] 딥 뉴럴 네트워크(DNN: deep neural network, 심층신경망)는 입력 레이어와 출력 레이어 외에 복수개의 히든 레이어를 포함하는 신경망을 의미할 수 있다. 딥 뉴럴 네트워크는 컨볼루션 뉴럴 네트워크(CNN: convolutional neural network), 리커런트 뉴럴 네트워크(RNN: recurrent neural network), 오토 인코더(auto encoder), GAN(Generative Adversarial Networks), 제한 볼츠만 머신(RBM: restricted boltzmann machine), 심층 신뢰 네트워크(DBN: deep belief network), Q 네트워크, U 네트워크, 샴 네트워크, 적대적 생성 네트워크(GAN: Generative Adversarial Network), 트랜스포머(transformer) 등을 포함할 수 있다. 전술한 딥 뉴럴 네트워크의 기재는 예시일 뿐이며 본 개시는 이에 제한되지 않는다.

[107] 뉴럴 네트워크는 지도학습(supervised learning), 비지도학습(unsupervised learning), 준지도학습(semi supervised learning), 자가학습(self-supervised learning) 또는 강화학습(reinforcement learning) 중 적어도 하나의 방식으로 학습될 수 있다. 뉴럴 네트워크의 학습은 뉴럴 네트워크가 특정한 동작을 수행하기 위한 지식을 뉴럴 네트워크에 적용하는 과정일 수 있다.

[108] 뉴럴 네트워크는 출력의 오류를 최소화하는 방향으로 학습될 수 있다. 뉴럴 네트워크의 학습에서 반복적으로 학습 데이터를 뉴럴 네트워크에 입력시키고 학습 데이터에 대한 뉴럴 네트워크의 출력과 타겟의 에러를 계산하고, 에러를 줄이기 위한 방향으로 뉴럴 네트워크의 에러를 뉴럴 네트워크의 출력 레이어에서부

터 입력 레이어 방향으로 역전파(backpropagation)하여 뉴럴 네트워크의 각 노드의 가중치를 업데이트 하는 과정이다. 지도학습의 경우 각각의 학습 데이터에 정답이 라벨링 되어있는 데이터(labelled data)를 사용하며, 비지도학습의 경우는 각각의 학습 데이터에 정답이 라벨링되어 있지 않은 데이터(unlabeled data)를 사용할 수 있다. 업데이트 되는 각 노드의 연결 가중치는 학습률(learning rate)에 따라 변화량이 결정될 수 있다. 입력 데이터에 대한 뉴럴 네트워크의 계산과 에러의 역전파는 학습 사이클(epoch)을 구성할 수 있다. 학습률은 뉴럴 네트워크의 학습 사이클의 반복 횟수에 따라 상이하게 적용될 수 있다. 또한, 과적합(overfitting)을 막기 위해서 학습 데이터의 증가, 레귤러화(regularization), 노드 일부를 비활성화하는 드롭아웃(dropout), 배치 정규화 레이어(batch normalization layer) 등의 방법이 적용될 수 있다.

[109] 일 실시예에서, 추정 모델은 트랜스포머의 적어도 일부분을 차용할 수 있다. 트랜스포머는 임베딩된 데이터들을 인코딩하는 인코더 및 인코딩된 데이터들을 디코딩하는 디코더로 구성될 수 있다. 트랜스포머는 일련의 데이터들을 수신하여, 인코딩 및 디코딩 단계를 거쳐 상이한 타입의 일련의 데이터들을 출력하는 구조를 지닐 수 있다. 일 실시예에서, 일련의 데이터들은 트랜스포머가 연산가능한 형태로 가공될 수 있다. 일련의 데이터들을 트랜스포머가 연산가능한 형태로 가공하는 과정은 임베딩 과정을 포함할 수 있다. 데이터 토큰, 임베딩 벡터, 임베딩 토큰 등과 같은 표현들은, 트랜스포머가 처리할 수 있는 형태로 임베딩된 데이터들을 지칭하는 것일 수 있다.

[110] 트랜스포머가 일련의 데이터들을 인코딩 및 디코딩하기 위하여, 트랜스포머 내의 인코더 및 디코더들을 어텐션(attention) 알고리즘을 활용하여 처리할 수 있다. 어텐션 알고리즘이란 주어진 쿼리(Query)에 대해, 하나 이상의 키(Key)에 대한 유사도를 구하고, 이렇게 주어진 유사도를, 각각의 키(Key)와 대응하는 값(Value)에 반영한 후, 유사도가 반영된 값(Value)들을 가중합하여 어텐션 값을 계산하는 알고리즘을 의미할 수 있다.

[111] 쿼리, 키 및 값을 어떻게 설정하느냐에 따라, 다양한 종류의 어텐션 알고리즘이 분류될 수 있다. 예를 들어, 쿼리, 키 및 값을 모두 동일하게 설정하여 어텐션을 구하는 경우, 이는 셀프-어텐션 알고리즘을 의미할 수 있다. 입력된 일련의 데이터들을 병렬로 처리하기 위해, 임베딩 벡터를 차원을 축소하여, 각 분할된 임베딩 벡터에 대해 개별적인 어텐션 헤드를 구하여 어텐션을 구하는 경우, 이는 멀티-헤드(multi-head) 어텐션 알고리즘을 의미할 수 있다.

[112] 일 실시예에서, 트랜스포머는 복수개의 멀티-헤드 셀프 어텐션 알고리즘 또는 멀티-헤드 인코더-디코더 알고리즘을 수행하는 모듈들로 구성될 수 있다. 일 실시예에서, 트랜스포머는 임베딩, 정규화, 소프트맥스(softmax) 등 어텐션 알고리즘이 아닌 부가적인 구성요소들 또한 포함할 수 있다. 어텐션 알고리즘을 이용하여 트랜스포머를 구성하는 방법은 Vaswani et al., Attention Is All You Need, 2017 NIPS에 개시된 방법을 포함할 수 있으며, 이는 본 명세서에 참조로 통합된다.

[113] 트랜스포머는 임베딩된 자연어, 분할된 이미지 데이터, 오디오 파형 등 다양한 데이터 도메인에 적용하여, 일련의 입력 데이터를 일련의 출력 데이터로 변환할 수 있다. 다양한 데이터 도메인을 가진 데이터들을 트랜스포머에 입력가능한 일련의 데이터들로 변환하기 위해, 트랜스포머는 데이터들을 임베딩할 수 있다. 트랜스포머는 일련의 입력 데이터 사이의 상대적 위치관계 또는 위상관계를 표현하는 추가적인 데이터를 처리할 수 있다. 또는 일련의 입력 데이터에 입력 데이터들 사이의 상대적인 위치관계 또는 위상관계를 표현하는 벡터들이 추가적으로 반영되어 일련의 입력 데이터가 임베딩될 수 있다. 일 예에서, 일련의 입력 데이터 사이의 상대적 위치관계는, 자연어 문장 내에서의 어순, 각각의 분할된 이미지의 상대적 위치 관계, 분할된 오디오 파형의 시간 순서 등을 포함할 수 있으나, 이에 제한되지 않는다. 일련의 입력 데이터들 사이의 상대적인 위치관계 또는 위상관계를 표현하는 정보를 추가하는 과정은 위치 인코딩(positional encoding)으로 지칭될 수 있다.

[114] 일 실시예에서, 추정 모델은 RNN(Recurrent Neural Network), LSTM(Long Short Term Memory) 네트워크, BERT(Bidirectional Encoder Representations from Transformers), 또는 GPT(Generative Pre-trained Transformer)를 포함할 수 있다.

[115] 일 실시예에서, 추정 모델은 전이학습(transfer learning) 방식으로 학습된 모델일 수 있다. 여기서, 전이학습은 대용량의 라벨링되어 있지 않은 학습용 데이터를 준지도학습 또는 자가학습 방식으로 사전 학습(pre-training)하여 제1 태스크를 갖는 사전 학습된(pre-trained) 모델을 얻고, 사전 학습된 모델을 제2 태스크에 적합하도록 fine-tuning하여 라벨링된 학습용 데이터를 지도학습 방식으로 학습해 타겟으로 하는 모델을 구현하는 학습 방식을 나타낸다.

[116] 일 실시예에서, 사전 학습의 제1 태스크와 fine-tuning의 제2 태스크는 상이할 수 있다. 구체적인 실시예로서, 제1 태스크는 타겟 분석물(예: 바이러스)의 핵산 서열에서 나타나는 일종의 언어와 유사한 서열 패턴이라는 현상을 언어 모델링(language modeling)하기 위한 것일 수 있다. 또는, 제1 태스크는 복수개의 핵산 서열을 이용한 범용의 태스크일 수 있다. 또한, 제2 태스크는 제1 태스크의 하위 태스크로서, 핵산 서열을 이용해서 해당 핵산 서열을 포함하는 유기체를 추정하기 위한 것 및/또는 해당 유기체의 host를 추정하기 위한 것일 수 있다. 예컨대, 바이러스의 서열 패턴을 언어 모델링하기 위한 태스크에 적합하도록 언어 모델을 이용해서 다양한 바이러스들의 핵산 서열들이 포함된 학습용 데이터를 트레이닝하여 사전 학습된 언어 모델을 얻은 후, 사전 학습된 언어 모델의 구조와 가중치를 유기체 추정 또는 host 추정을 위한 태스크에 적합하도록 fine-tuning하여 추정 모델을 얻을 수 있다.

[117] 일 실시예에서, 전이 학습 방식으로 학습된 추정 모델은 베이스들의 종류 및 순서 정보를 사용하여 사전 학습된 모델에, 복수개의 핵산 서열 각각에서의 유기체의 추정값 또는 host의 추정값의 출력이 fine-tuning으로서 적용된 모델을 의미할 수 있다. 일 실시예에서, 사전 학습된 모델은 특정한 태스크(예: 분류, 검출, 세그

먼테이션 등) 또는 범용 태스크에 따라 학습된 딥러닝 언어 모델을 의미할 수 있다. 일례로, 사전 학습된 모델은 베이스들의 종류(예컨대, A, G, C 및 T) 및 베이스들의 정렬 순서에 기초하여 사전 학습이 이루어질 수 있다.

[118] 일 실시예에서, fine-tuning은 사전 학습된 모델을, 유기체 추정 또는 host 추정을 위한 특정한 태스크로 전이(transfer)하여 해당 모델을 사후적으로 학습하는 개념을 포괄할 수 있다. 구체적으로, fine-tuning은 사전 학습된 모델을 가져와 유기체 또는 host를 추정하기 위한 태스크에 맞게 변형하고, 사전 학습된 모델의 가중치들로부터 학습을 업데이트하는 방법을 의미할 수 있다. 일례로, 이러한 fine-tuning은 사전 학습된 모델에 유기체 추정 또는 host 추정을 위한 특정한 데이터셋을 추가로 학습시킴으로써 사전 학습된 모델의 파라미터들을 업데이트하는 과정을 포함할 수 있다.

[119] 이하에서는, 위에서 제시한 추정 모델이 전이학습 방식으로 학습되는 과정의 다양한 실시예들에 대해 보다 상세하게 서술하도록 한다.

[120]

[121] 사전 학습 예시

[122] 도 2는 일 실시예에 따른 사전 학습의 과정에 대한 개념도를 예시적으로 도시한다. 일 실시예에서, 도 2는 언어 모델(language model)(210)을 이용하여 사전 학습이 수행되는 방식을 예시적으로 도시한다. 사전 학습이 수행된 언어 모델(210)은 본 개시에서의 사전 학습된 모델에 대응될 수 있다.

[123] 도 2를 더 참조하면, 컴퓨터 장치(100)는 언어 모델(210)을 이용하여 사전 학습의 과정을 수행할 수 있다. 일 실시예에서, 언어 모델(210)은 인공 신경망 언어 모델로서, 상술한 트랜스포머의 적어도 일부분을 차용할 수 있다. 구체적인 예로서, 언어 모델(210)은 트랜스포머 계열의 언어 모델인 BERT 또는 GPT를 포함할 수 있다.

[124] 일 실시예에서, 사전 학습의 과정에서 복수개의 핵산 서열(220)이 학습용 데이터로서 이용될 수 있다. 여기서, 핵산 서열은 유기체의 핵산 서열 중 적어도 일부에 대한 정보를 의미하고, 예컨대, 병원체의 유전체(genome) 또는 그 중 일부의 유전자(gene)에 포함된 베이스들의 종류 및 순서 정보 또는 베이스 쌍(base pair)의 배열 정보를 의미할 수 있다.

[125] 일 실시예에서, 사전 학습 과정에서 이용되는 핵산 서열(220)은 해당 핵산 서열을 포함하는 유기체의 종류가 제한되어 있지 않을 수 있다. 일 예로, 생물학적 카테고리가 제한되지 않은 다양한 유기체들의 핵산 서열들이 광범위하게 수집되어 학습용 데이터로서 이용될 수 있다. 다른 일 예로, 수집된 핵산 서열들 중에서 적어도 일부의 핵산 서열들이 랜덤하게 또는 사전설정된 규칙(예: 카테고리별로 정해진 개수의 정보들 선택)에 따라 선택되어서 학습용 데이터로서 이용될 수 있다. 여기서, 생물학적 카테고리는 각 유기체의 생물학적 분류 체계 상에서의 위치 또는 범위를 표현하기 위한 것으로, 뒤에서 다시 설명하도록 한다.

[126] 다른 일 실시예에서, 사전 학습 과정에서 이용되는 핵산 서열(220)은 해당 핵산 서열을 포함하는 유기체의 종류가 일부 제한되어 있을 수 있다. 구체적으로, 복수개의 핵산 서열(220)은 하나 이상의 사전설정된 생물학적 카테고리에 속하는 유기체로부터 획득된 것일 수 있다. 예컨대, 생물학적 분류 체계 상에서 목 레벨에 속하는 바이러스들(예: 바이러스 목)의 핵산 서열들이 대용량으로 수집될 수 있고, 사전 학습의 과정에서 학습용 데이터로서 이용될 수 있다.

[127] 일 실시예에서, 복수개의 핵산 서열(220)은 공개 데이터베이스(public database)로부터 획득된 데이터이거나 이로부터 가공, 변형 또는 분리된 데이터일 수 있다. 예컨대, 컴퓨터 장치(100)는 NCBI(National Center for Biotechnology Information)나 GISAID(Global Initiative for Sharing All Influenza Data) 등의 공개 데이터베이스에 액세스하여 공개 데이터베이스에 등재되어 있는 대용량의 바이러스 서열들을 수집하고, 수집된 바이러스 서열들에 대한 텍스트 전처리(text preprocessing)를 수행하여 사전 학습을 위한 학습용 데이터로서 가공할 수 있다.

[128] 일 실시예에서, 언어 모델(210)에는 복수개의 핵산 서열(220) 각각이 입력되고, 언어 모델(210)은 각 핵산 서열에 포함된 베이스들의 종류 및 순서에 기초하여, 마스킹된 베이스의 베이스별 확률값(230)을 추정하도록 사전 학습될 수 있다. 이러한 사전 학습의 과정에서 언어 모델(210)은 복수개의 핵산 서열(220) 각각에 포함된 베이스들 중 일부의 베이스에 마스크(mask)를 적용한 뒤, 마스킹된(masked) 베이스를 맞추는 semi supervised learning 또는 self-supervised learning 방식에 의해 학습될 수 있다. 예컨대, 핵산 서열 내에서 임의의 베이스를 정하여 마스킹하는 과정을 통해 라벨링되어 있지 않은 학습용 데이터를 라벨링된 학습용 데이터로 변환하고, 라벨링된 학습용 데이터를 이용해서 마스킹된 베이스를 맞추도록 태스크를 부여하는 self-supervision 학습 방식으로 언어 모델(210)에 대한 사전 학습(240)이 이루어질 수 있다.

[129] 일 예로, 언어 모델(210)은 핵산 서열에 포함된 베이스들의 시퀀스에 확률을 할당(assign)하고, 마스킹된 베이스의 전후로 어떤 베이스들이 출현하였는지 고려하여, 마스킹된 베이스의 후보가 될 수 있는 여러 베이스들에 대해서 출현 확률을 추정해 마스킹된 베이스의 베이스별 확률값(230)을 출력할 수 있다. 또한, 언어 모델(210)은 출력 데이터인 마스킹된 베이스의 베이스별 확률값(230)과 해당 핵산 서열에 포함된 베이스들의 종류 및 순서를 비교함으로써 에러를 계산할 수 있으며, 에러를 줄이기 위한 역전파에 따라 언어 모델(210)의 파라미터들을 업데이트할 수 있다.

[130] 도 3은 일 실시예에 따른 사전 학습에 이용되는 BERT 기반의 언어 모델(210)의 구조 및 동작을 예시적으로 도시한다. 구체적인 일 실시예로서, 언어 모델(210)을 이용한 사전 학습에는 임베딩된 데이터들을 인코딩하는 인코더가 복수개 연결된 구조를 이용하는 BERT의 적어도 일부가 차용될 수 있다.

[131] 도 3을 더 참조하면, 언어 모델(210)은 복수개의 핵산 서열(220) 각각에 대응되는, 마스킹된 토큰들 및 비-마스킹된 토큰들을 사용하여 상기 마스킹된 토큰들

각각에 대한 복수개의 추정값들을 출력하는 분류(classification) 모델을 의미할 수 있다. 여기서, 하나의 추정값은 언어 모델(210)의 하나의 클래스에 대응될 수 있다. 일례로, 언어 모델(210)은 각 핵산 서열로서 마스킹된 토큰들 및 비-마스킹된 토큰들을 입력 받을 수 있다. 다른 예시로, 언어 모델(210)은 각 핵산 서열을 입력 받고 그리고 입력된 각 핵산 서열에 대한 전처리를 수행하여 마스킹된 토큰들 및 비-마스킹된 토큰들을 생성할 수 있다.

[132] 도 3에 도시된 것처럼, 언어 모델(210)은 입력 임베딩 레이어(input embedding layer)(310), 인코더 레이어(encoder layer)(320) 및 제1 분류 레이어(classifier layer)(330) 중 적어도 하나를 포함할 수 있다.

[133] 일 실시예에서, 입력 임베딩 레이어(310)는 일련의 입력 데이터인 핵산 서열(220)을 인코더가 연산 가능한 형태로 변환할 수 있다. 일 실시예에서, 입력 임베딩 레이어(310)는 핵산 서열(220)에 포함된 베이스들을 토큰화(tokenization)하는 토큰 임베딩 레이어(token embedding layer) 및 벡터들에 위치 정보를 인가하는 포지셔널(또는 포지션) 임베딩 레이어(position(al) embedding layer) 중 적어도 하나를 포함할 수 있다. 또한, 실시예에 따라서, 입력 임베딩 레이어(310)는 세그먼트 임베딩 레이어(segment embedding layer) 등의 추가 임베딩 레이어를 더 포함할 수 있다.

[134] 일 실시예에서, 토큰 임베딩 레이어는 핵산 서열(220)을 각각 두 개 이상의 베이스들을 갖는 토큰들로 토큰화하는 토큰화 과정을 수행할 수 있다. 일 실시예에서, 토큰화 과정은 핵산 서열에 포함된 복수개의 베이스들을 그룹화하는 작업을 의미할 수 있다. 토큰화 과정에서 생성되는 토큰들 중 하나의 토큰은 하나 이상의 베이스들을 포함할 수 있다.

[135] 일 실시예에서, 이러한 토큰들은 (i) 핵산 서열(220)을 k 개(k는 자연수)씩 분할하거나 (ii) 핵산 서열(220)을 기능 단위로 분할하여서 토큰화되는 베이스들을 각각 포함할 수 있다. 전자의 예를 들면, 토큰화 과정에서 베이스들을 k개씩 자르는 k-mer 기법이 이용될 수 있다. k-mer 기법을 이용하는 예시에서 만약 k가 3이면 각 토큰에 포함되는 베이스들의 개수는 총 3개일 수 있다. 후자의 예를 들면, 토큰화 과정에서 핵산 서열을 기능에 의해 splicing하는 유전자 추정(gene prediction) 기법이 이용될 수 있다. 구체적인 일 예시로서, 기능 단위로 분할하는 것은, 아미노산 1개의 코딩이 가능한 코돈(codon) 단위로 분할하는 것, 및 염기 서열 내에서 유전자 발현(예: 전사, 해독)이나 발현 양상 등과 연관된 구간(예: TATA box, Homeo box 등) 단위로 분할하는 것 중 적어도 하나를 포함할 수 있다.

[136] 추가적인 실시예에서, 토큰화 과정은 다양한 기법들에 기초하여 수행될 수도 있다. 예를 들어, Bite Pair Encoding 알고리즘에 기초하여 핵산 서열에 대한 토큰화가 이루어질 수 있다. 다른 예시로, 최적 k-mer 길이(optimal k-mer size)에 기초하여 핵산 서열(220)에 대한 토큰화가 이루어질 수 있다. 이러한 예시에서는 각각의 유기체들마다 Genome을 가장 잘 표현할 수 있는 특정 k-mer size을 가질 수 있기 때문에, 이러한 유기체들 단위로 결정된 특정 k-mer size에 기초하여 토큰화가

이루어질 수 있다. 또 다른 예시로, DNA motif에 기초하여 토큰화가 이루어질 수도 있다. 또 다른 예시로, 고등 생물에서 유전자(gene)내에서의 RNA로 전사하는 코딩 영역(coding region) 단위의 단위인 엑손(exon)에 기초하여 핵산 서열(220)에 대한 토큰화가 수행될 수도 있다.

[137] 일 실시예에서, 토큰 임베딩 레이어는 토큰화 과정을 통해 생성된 복수개의 토큰들에 대한 전처리를 수행하여 마스킹된 토큰들 및 비-마스킹된 토큰들을 생성할 수 있다. 예컨대, 토큰 임베딩 레이어는 핵산 서열로부터 토큰화된 복수개의 비-마스킹된 토큰들 중 적어도 하나를 특수 토큰인 [MASK] 토큰으로 마스킹하여 마스킹된 토큰을 생성할 수 있다.

[138] 일 실시예에서, 토큰 임베딩 레이어는 각 토큰을 벡터로 표현할 수 있으며, 예컨대, 토큰화된 베이스들을 워드 임베딩하여 dense vector의 형태의 임베딩 벡터로 변환할 수 있다.

[139] 일 실시예에서, 포지셔널(또는 포지션) 임베딩 레이어는 임베딩 벡터들이 인코더의 입력으로 사용되기 전에 임베딩 벡터들에 위치 정보를 인가할 수 있다. 예컨대, 위치 정보를 학습을 통해서 얻는 포지션 임베딩이 이용될 수 있으며, 일 예로, 핵산 서열의 길이에 대응되는 복수개의 포지션 임베딩 벡터를 학습시키고, 각 임베딩 벡터마다 해당 포지션 임베딩 벡터를 더하는 방식이 이용될 수 있다.

[140] 상술한 입력 임베딩 레이어(310)를 거치면서 인코더가 연산가능한 형태로 가공된 임베딩된 데이터들은 복수개의 인코더가 쌓인 구조의 인코더 레이어(320)에 입력으로서 제공될 수 있다. 이에 따라, 인코더 레이어(320) 내의 첫 번째 인코더에서 임베딩된 데이터들을 연산한 결과는 다음 인코더를 향해 출력되며, 마지막 인코더에서는 입력된 임베딩된 데이터들을 종합적으로 고려한 context vector가 출력될 수 있다.

[141] 일 실시예에서, 인코더 레이어(320)는 N 개(예: 12, 24 등)의 인코더 블록을 포함할 수 있다. N 개의 인코더 블록이 쌓인 구조는 입력 시퀀스 전체의 의미를 N번 반복적으로 구축하는 것을 의미하며, 인코더 블록의 개수가 많을수록 핵산 서열 내 베이스들 간의 의미 관계가 보다 잘 반영될 수 있다. N 개의 인코더 블럭은 입력 시퀀스 전체가 재귀적으로(recursive) 반복 처리되는 형태로 구성될 수 있다.

[142] 일 실시예에서, 각 인코더 블록은 제공되는 입력에 대해 멀티-헤드 어텐션 알고리즘을 이용하여 가중치 기반의 연산 결과를 출력할 수 있다. 예컨대, 각 인코더 블록은 서로 다른 가중치 행렬을 이용해 어텐션을 h번 계산한 후 이를 서로 연결한 결과인 순차(concatenation)를 출력할 수 있다. 이에 따라 작은 차이에도 결과의 차이가 두드러지게 나타날 수 있다. 또한, 각 인코더 블록 내에서 입력과 처리 결과는 정규화(normalization), 잔차 연결(residual connection) 및 피드 포워드 신경망 등을 거쳐 연산됨에 따라 학습 효과가 향상될 수 있다.

[143] 일 실시예에서, 제1 분류 레이어(330)는 인코더 레이어(320)에서 출력된 결과를 제1 태스크에 적합하도록 사용자에게 유의미한 형태로 변환할 수 있다. 예컨대, 제1 분류 레이어(330)는 인코더 레이어(320)의 마지막 인코더 블록에서 출력되는

임베딩 벡터(예: context vector)를 입력으로 사용하여 제1 태스크를 수행하기 위한 분류 기능을 수행하는 classifier를 포함할 수 있다. 일 예로, classifier는 마스킹된 토큰의 위치에 대응되는 출력 임베딩 벡터를 입력으로 이용하여 마스킹된 토큰의 베이스별 확률값(230)을 출력하기 위한 소프트맥스 함수를 포함할 수 있다.

[144] 이러한 제1 분류 레이어(330)의 출력과 학습용 데이터의 핵산 서열(220)에 포함된 베이스들을 비교하는 방식으로 에러 연산 및 가중치 업데이트가 이루어질 수 있으며, 이와 같은 사전 학습이 진행됨에 따라 사전 학습된 모델이 획득될 수 있다. 이처럼 마스킹된 베이스를 맞추는 MLM(Masked Language Modeling) 방식의 사전 학습은 주어진 핵산 서열 내에서 양방향에 위치한 베이스들을 고려하여 마스킹된 베이스를 추정하기 때문에, 추정 정확성이 높으며, 일종의 언어와도 유사한 특성을 갖는 핵산 서열들의 패턴에 대해 보다 잘 이해하는 사전 학습된 언어 모델이 구현될 수 있다.

[145] 일 실시예에서, 사전 학습의 과정에서는 마스킹된 베이스를 맞추도록 학습하는 방식과 일부의 베이스를 다른 베이스로 대체시킨 후 틀린 베이스를 수정하도록 학습하는 방식이 함께 사용될 수도 있다. 또한, 사전 학습의 과정에서 NSP(Next Sentence Prediction) 방식이 추가적으로 이용될 수도 있으며, 예컨대, MLM 방식과 NSP 방식을 함께 이용하여 사전 학습이 이루어질 수도 있다.

[146] 도 4는 일 실시예에 따른 컴퓨터 장치(100)에서 사전 학습된 모델에 의해 마스킹된 베이스의 베이스별 확률값(230)이 추정되는 예시적인 방식을 도시한다.

[147] 도 4에서는 하나의 토큰 또는 하나의 마스킹된 토큰에 포함되는 베이스들의 개수가 3개인 것을 예시로 들고 있다. 구현 양태에 따라, 토큰에 포함되는 베이스들의 개수가 다양한 개수를 포함할 수 있다는 점이 당업자에게 명백할 것이다.

[148] 일 실시예에서, 사전 학습된 모델은 핵산 서열(410) 중 특정 베이스(440)에 대한 베이스별 확률값(230)을 연산할 수 있다. 핵산 서열(410)이 사전 학습된 모델에 제공되면, 사전 학습된 모델은 핵산 서열(410) 중에서 하나 이상의 특정 베이스(440)를 결정하고, 특정 베이스(440)에 대한 베이스별 확률값(230)을 출력할 수 있다. 핵산 서열(410)은 핵산 서열(220)에 대응될 수 있다.

[149] 도 4의 예시에서, 핵산 서열(410) 중 베이스별 확률값(230)을 연산하기 위한 베이스(440)는 A이다. 사전 학습된 모델은 복수개의 베이스들로 이루어진 핵산 서열(410)을 복수개의 토큰들(420)로 토큰화할 수 있다.

[150] 도 4에서는 복수개의 토큰들(420) 각각이 3-mer 기법에 따라 생성되는 것을 예로 들고 있다. 복수개의 토큰들(420) 각각은 3개의 베이스들을 포함할 수 있다. 도 4에서의 예시에서, 핵산 서열(410)에서의 첫번째 베이스인 A는 ATT를 포함하는 토큰과 대응될 수 있다. 핵산 서열(410)에서의 두번째 베이스인 T는 ATT를 포함하는 토큰 및 TTG를 포함하는 토큰과 대응될 수 있다. 핵산 서열(410)에서의 세번째 베이스인 T는 ATT를 포함하는 토큰, TTG를 포함하는 토큰 및 TGA를 포함하는 토큰과 대응될 수 있다. 이처럼 도 4에서는 핵산 서열(410)에서의 복수개의 베이스들의 배치 순서 대로 하나의 베이스를 이동해가면서 토큰들이 생성되는

것이 도시된다. 이러한 예시에서는 인접한 토큰들에 대해서는 서로 2개의 베이스들을 서로 공유할 수 있다.

[151] 일 실시예에서, 생성되는 토큰들 각각은 핵산 서열(410)에서의 베이스들을 k 개씩 자른 결과인 k-mer에 대응될 수 있다. 여기서의 k는 자연수일 수 있으며, 일례로 k는 3 이상이고 20 이하인 자연수를 의미할 수 있다. 이러한 예시에서, 토큰들 각각이 포함하고 있는 베이스들의 개수는 k와 대응될 수 있다. 즉, k가 3인 경우, 하나의 토큰에 3개의 베이스들이 포함될 수 있다.

[152] 일 실시예에서, 마스킹된 토큰들 각각은 k개의 베이스들을 포함하고 그리고 핵산 서열을 구성하는 n개의 베이스들 중 k번째 내지 n-k 번째의 범위에 있는 베이스들 각각에 대응하여 생성되는 마스킹된 토큰들의 개수는 k개에 해당할 수 있다. 여기서 k 및 n은 각각 자연수이며, 예컨대, $k \geq 2$ 이고 $n \geq 2k$이다.

[153] 예컨대, 복수개의 토큰들(420) 중에서 핵산 서열(410) 중 제 1 베이스(440: A)에 대응되는 토큰들(450)은 TGA, GAC 및 ACG를 포함할 수 있다. 해당 토큰들(450)은 상기 제 1 베이스(440) A를 제 1 위치에 포함하는 제 1 토큰(TGA), 제 1 베이스(440) A를 제 2 위치에 포함하는 제 2 토큰(GAC) 및 제 1 베이스(440) A를 제 3 위치에 포함하는 제 3 토큰(ACG)을 포함할 수 있다. 전술한 예시에서, 3-mer 기법에 기초하여 토큰들이 생성되었기 때문에, 하나의 토큰은 3개의 베이스들을 포함할 수 있으며 하나의 베이스에 총 3개의 토큰들이 대응될 수 있다.

[154] 일 실시예에서, 복수개의 베이스들(410) 중 제 1 베이스(440)에 대하여, 서로 상이한 위치에서 상기 제 1 베이스(440)를 포함하는 제 1 세트의 토큰들(450) 및 상기 제 1 세트의 토큰들(450)에 대응되는 제 1 세트의 마스킹된 토큰들(460: 460a, 460b 및 460c)이 생성될 수 있다. 일 실시예에서, 제 1 베이스(440)에 대한 베이스별 확률값(230)은, 제 1 세트의 마스킹된 토큰들(460: 460a, 460b 및 460c) 각각에 대하여 언어 모델(210)로부터 출력되는 추정값들(480a, 480b, 480c)에 기초하여 결정될 수 있다.

[155] 사전 학습된 모델은 복수개의 토큰들(420) 중 적어도 일부의 토큰들인 제 1 세트의 토큰들(450)에 마스크를 적용함으로써 마스킹된 토큰들(460)을 획득할 수 있다. 예를 들어, 사전 학습된 모델은 제 1 베이스(440)에 대응하는 3개의 토큰들(450) 각각에 마스크를 적용하여 마스킹된 토큰들(460)을 생성할 수 있다. 전술한 예시에서, 3-mer에 기초하여 토큰들이 생성되었기 때문에, 하나의 베이스에 대해 마스킹된 토큰들은 3개의 베이스들과 대응될 수 있으며, 마스μg된 토큰들의 개수도 3개와 대응될 수 있다.

[156] 사전 학습된 모델은 토큰들(420)로부터 마스킹된 토큰들(460) 및 비-마스킹된 토큰들을 포함하는 중간 입력 데이터(430)를 생성할 수 있다.

[157] 일 실시예에서, 사전 학습된 모델은 마스킹된 토큰들(460) 각각(460a, 460b 및 460c)에 대하여 클래스들(470a, 470b 및 470c)에 대한 추정값들(480a, 480b 및 480c)을 획득할 수 있다. 사전 학습된 모델은 획득된 추정값들(480a, 480b 및

480c)에 기초하여 제 1 베이스(440)에 대한 베이스별 확률값(230)을 연산할 수 있다.

[158] 일 실시예에서, 컴퓨터 장치(100)는 클래스들(470a, 470b 및 470c)에 대한 추정값들(480a, 480b 및 480c)의 평균을 통해 제 1 베이스(440)에 대한 베이스별 확률값(230)을 연산할 수 있다. 다른 일 실시예에서, 컴퓨터 장치(100)는 클래스들(470a, 470b 및 470c) 각각에 부여한 가중치에 기반하여 클래스들(470a, 470b 및 470c) 각각에 대한 추정값들(480a, 480b 및 480c)에 대한 가중 평균을 통해 제 1 베이스(440)에 대한 베이스별 확률값(230)을 연산할 수 있다.

[159] 일 실시예에서, 사전 학습된 모델의 파라미터들은 이러한 특정 베이스(440)에 대한 베이스별 확률값(230)과 핵산 서열(410) 내의 특정 베이스(440)를 비교함으로써, 또는 마스킹된 토큰들(460: 460a, 460b 및 460c) 각각에 대한 예측값들(480a, 480b, 480c)과 특정 베이스(440)에 대응되는 토큰들(450)을 비교함으로써, 에러가 최소화되도록 업데이트될 수 있다. 이와 같은 방식으로, 사전 학습된 모델은 마스킹된 베이스를 보다 잘 맞추도록 학습될 수 있다.

[160]

[161] Fine-tuning 예시

[162] 도 5는 일 실시예에 따른 fine-tuning의 과정에 대한 개념도를 예시적으로 도시한다. 일 실시예에서, 도 5에 도시된 사전 학습된 모델(510)은 사전 학습이 수행된 언어 모델(210)에 대응될 수 있다.

[163] 도 5를 참조하면, 사전 학습된 모델(510)을 fine-tuning하여서 추정 모델이 획득될 수 있다. 일 실시예에서, fine-tuning의 과정은 사전 학습된 모델(510)을 이용하여 추정 모델의 구조를 정하는 과정과, host 추정 또는 유기체 추정을 위한 학습용 데이터로 추정 모델을 학습시키는 과정을 포함할 수 있다. 도 5에는 일 실시예에 따른 host 추정을 위한 fine-tuning의 과정이 도시되어 있으나, 이와 유사한 방식으로 다른 일 실시예에 따른 유기체 추정을 위한 fine-tuning의 과정이 수행될 수 있다.

[164] 일 실시예에서, 사전 학습된 모델(510)을 이용하여 추정 모델의 구조를 정하는 과정은, 사전 학습된 모델(510)에 host 추정 또는 유기체 추정을 위한 레이어를 적용하는 방식으로 수행될 수 있다. 예컨대, host 추정을 위한 fine-tuning의 경우, 사전 학습으로 가중치가 기 연산된 사전 학습된 모델(510) 중에서 입력 임베딩 레이어(310) 및 인코더 레이어(320) 중 하나 이상을 가져오고, 인코더 레이어(320)의 마지막 인코더 블록에 host 추정용 레이어(520)를 추가할 수 있고, 유기체 추정을 위한 fine-tuning의 경우, 같은 방식으로 유기체 추정용 레이어를 추가할 수 있다. 일 실시예에서, 유기체 추정용 레이어는 host 추정용 레이어(520)와 적어도 부분적으로 동일한 구조를 가질 수 있으며, 예컨대, host 추정과 비교하여 모델의 구조는 동일하되 상이한 학습용 데이터 세트를 이용하는 방식으로 fine-tuning이 이루어질 수 있다.

[165] 일 실시예에서, fine-tuning은 복수의 학습용 데이터 세트를 이용해서 수행될 수 있다. 각 학습용 데이터 세트는 (i) 핵산 서열을 포함하는 학습용 입력 데이터 및 (ii) 해당 핵산 서열을 포함하는 유기체 또는 해당 유기체의 host에 대한 라벨 데이터를 포함하는 학습용 정답 데이터를 포함할 수 있다. 즉, host 추정을 위한 fine-tuning에서 각 학습용 데이터 세트는 핵산 서열 및 해당 핵산 서열의 host에 대한 라벨 데이터(530)를 포함하고, 유기체 추정을 위한 fine-tuning에서 각 학습용 데이터 세트는 핵산 서열 및 해당 핵산 서열의 유기체에 대한 라벨 데이터를 포함할 수 있다.

[166] 일 실시예에서, fine-tuning의 과정에서 학습용 입력 데이터로서 이용되는 핵산 서열은, 해당 핵산 서열을 포함하는 유기체의 종류가 제한되어 있을 수 있다. 구체적으로, fine-tuning의 과정에서의 핵산 서열은 하나 이상의 사전설정된 생물학적 카테고리에 속하는 유기체들로부터 획득된 것일 수 있다. 일 예로, 생물학적 분류 체계 상에서 목 레벨에 속하는 바이러스들의 핵산 서열들이 학습용 입력 데이터로서 이용될 수 있다. 다른 일 예로, 생물학적 분류 체계 상에서 종 레벨에 속하는 복수개의 특정 바이러스들(예: Rotavirus A, Influenza virus A, Rabies lyssavirus)의 핵산 서열들이 학습용 입력 데이터로서 이용될 수 있다.

[167] 일 실시예에서, fine-tuning의 과정에서의 핵산 서열을 포함하는 유기체가 생물학적 분류 체계 상에서 위치하는 hierarchical level은, 사전 학습 과정에서의 핵산 서열(220)을 포함하는 유기체가 생물학적 분류 체계 상에서 위치하는 hierarchical level보다 하위 level일 수 있다. 예컨대, 사전 학습의 과정에서 목 레벨에 속하는 바이러스 서열들이 대용량으로 수집되어 학습용 데이터로서 이용되었다면, fine-tuning의 과정에서는 목 레벨보다 하위 레벨인 종 레벨에 속하는 바이러스 서열들 중에서 사전설정된 생물학적 카테고리에 속하는 생명체들(예: Rotavirus A, Influenza virus A, Rabies lyssavirus)의 핵산 서열들이 이용될 수 있다.

[168] 일 실시예에서, fine-tuning의 과정에서 이용되는 학습용 데이터 세트의 개수는 host의 종류별로 또는 유기체의 종류별로 사전설정된 범위 내에 있도록 제한될 수 있다. 예컨대, host 종류별로 학습용 데이터 세트의 개수 분포 차이가 기 설정된 수준 이상 크지 않도록, host의 생물학적 카테고리별로 정해진 개수나 범위 내의 데이터셋이 학습용 데이터 세트로서 선택될 수 있다.

[169] 일 실시예에서, fine-tuning의 학습용 정답 데이터로서 이용되는 host(또는 유기체)에 대한 라벨 데이터는, 해당 핵산 서열을 포함하는 유기체의 host의 종류(또는 유기체의 종류)에 대한 라벨 데이터를 의미할 수 있다. 여기서, host의 종류는 해당 host(또는 유기체)의 생물학적 카테고리로서, 생물학적 분류 체계를 구성하는 복수개의 hierarchical level 중 어느 하나의 hierarchical level에 위치하는 카테고리(예: 종명)를 포함할 수 있다. 여기서, 생물학적 카테고리가 의미하는 바는 생물학적 분류 체계를 통해 정의될 수 있으며, 이에 관해서는 도 6을 더 참조하여 설명하도록 한다.

[170] 도 6은 일 실시예에 따른 생물학적 분류 체계와 생물학적 카테고리에 대한 개념도를 예시적으로 도시한다.

[171] 도 6에 도시된 것처럼, 생물학적 분류 체계는 유기체가 속하는 범위를 구분하기 위한 분류 체계로서, 복수개의 레벨(610)로 구성될 수 있다. 복수개의 레벨(610)은 상위 레벨이 하위 레벨을 포괄하는 형태의 생물학적 계통 구조를 가질 수 있으며, 예컨대, 종, 속, 과, 목, 강, 문, 계 및 역을 포함하는 레벨들 중 적어도 하나로 표현될 수 있다.

[172] 일 실시예에서, 복수개의 생물학적 카테고리(620)는 복수개의 레벨(610) 중 어느 하나의 레벨에 위치할 수 있다. 예컨대, 복수개의 생물학적 카테고리(620)는 생물학적 분류 체계에서의 종 레벨(611)에 위치하고, 일 예로, Homo sapiens, Sus scrofa, Bos taurus, Equus caballus 및 Gallus gallus 등 유기체의 host가 될 수 있는 개체들의 종명(species name)으로 표현될 수 있다.

[173] 도 6에는 host의 생물학적 카테고리로서 추정될 수 있는 후보군으로서 일부 종들이 예시적으로 도시되어 있으나, 이에 제한되지 않으며, 임의의 유기체의 host로서 공지된 다양한 개체들이 적용될 수 있다. 또한, 이러한 생물학적 카테고리는 해당 핵산 서열을 포함하는 유기체의 종류를 나타낼 때에도 같은 방식으로 해석될 수 있다.

[174] 일 실시예에서, 복수개의 생물학적 카테고리(620)는 host 또는 유기체에 대해 추정될 수 있는 생물학적 카테고리의 후보군으로서, 해당 핵산 서열을 포함하는 유기체의 종류에 기초하여 결정되거나 분류될 수 있다. 일 예로, 유기체의 종류가 Rotavirus 계열인 경우, host에 대한 복수개의 생물학적 카테고리(620)는 Homo sapiens, Sus scrofa, Bos taurus, Equus caballus, Gallus gallus 및 Vicugna pacos 등을 포함할 수 있다. 다른 일 예로, 생명체의 종류가 Influenza 계열인 경우, host에 대한 복수개의 생물학적 카테고리(620)는 Homo sapiens, Sus scrofa, Anas platyrhynchos, Gallus gallus, Anas discors, Arenaria interpres, Anas acuta, Meleagris gallopavo, Anas clypeata, Equus caballus, Anas carolinensis 등을 포함할 수 있다. 또 다른 일 예로, 생명체가 Rabies lyssavirus 계열인 경우, host에 대한 복수개의 생물학적 카테고리(620)는 Canis lupus, Bos taurus, Desmodus rotundus, Vulpes vulpes, Eptesicus fuscus, Mephitis mephitis, Homo sapiens, Procyon lotor, Felis catus, Tadarida brasiliensis 및 Vulpes lagopus 등을 포함할 수 있다.

[175] 한편, 도 6에서는 계층 구조를 갖는 생물학적 분류 체계가 도시되어 있으나, 이에 제한되지 않는다. 실시예에 따라, 생물학적 카테고리에는 분류군의 성격에 따른 계통군 분류 구조 등 다양한 분류 체계가 적용될 수 있다.

[176] 표 1은 일 실시예에 따른 fine-tuning에 이용되는 N개의 학습용 데이터 세트를 예시적으로 도시한다. 표 1에서, 각 학습용 데이터 세트는 각 샘플별로 핵산 서열의 입력(학습용 입력 데이터)과, 해당 핵산 서열을 포함하는 유기체 또는 해당 유기체의 host의 생물학적 카테고리에 대한 라벨 데이터(학습용 정답 데이터)을 포함할 수 있다. 하나의 생물학적 카테고리(예: 종명)는 하나의 카테고리 식별자

(예: M)과 매핑될 수 있고, 예컨대, 학습용 정답 데이터에는 표 1에 도시된 것처럼, 해당 host의 생물학적 카테고리와 매핑된 카테고리 식별자가 포함될 수 있다.

[177] [표1]

	핵산 서열	라벨 데이터
샘플1	AGCATTGTGGGTAGTAAGGTATAAA ... AGCTCAAAATCTACA	1
샘플2	AGCGTTATTGTTGAGAAATGGATTG ... AGCACAAAAAAATTT	2
...	...	...
샘플N	AGCTGTTTTTTTTTTTGTTGTGGGTAA ... AGCCTATAAATCC	M

[178] 일 실시예에서, fine-tuning에 이용되는 학습용 데이터 세트는 사전 검증된 데이터 세트일 수 있다. 일 예로, 소정의 매체나 수단을 통해 생명공학 학술 정보(예: 논문 등)에 게시되어 있거나, NCBI나 GISAID 등 공개 데이터베이스에 등재되어 있는 핵산 서열과 해당 host 및/또는 해당 유기체에 대한 정보가 수집될 수 있고, 수집된 정보들 중에서 사전설정된 검증 방식(예: 관리자 검증, 룰 기반 검증)에 의해 host 및/또는 유기체에 대한 정보가 정확한 것으로 검증된 핵산 서열들이 학습용 데이터 세트로서 선택될 수 있다.일 실시예에서, fine-tuning에 이용되는 학습용 데이터 세트의 수는 사전 학습에 이용되는 학습용 데이터의 수보다 작을 수 있다. 예컨대, 사전 학습에 수십만 개 내지 수백만 개 이상의 대용량 학습용 데이터가 사용되었다면, fine-tuning에는 약 천 개 내지 수천 개 수준의 저용량 학습용 데이터 세트가 사용될 수 있다.

[179] 전술한 학습용 데이터 세트를 이용해서, 추정 모델은 사전 학습된 모델(510)이 카테고리별 확률값(540)을 출력하도록 fine-tunning될 수 있다. 예컨대, host 추정을 위한 fine-tuning의 과정에서 추정 모델은 학습용 입력 데이터인 각 핵산 서열에 포함된 베이스들의 종류 및 순서에 기초하여, 해당 핵산 서열에서의 host의 카테고리별 확률값(540)을 출력할 수 있다. 또한, 상기 학습용 데이터 세트를 labeled data로서 이용하여 fine-tuning을 통한 supervised learning 방식으로 학습(550)이 이루어질 수 있다. 또는, 전술한 것과 유사한 방식으로, 유기체 추정을 위한 fine-tuning의 과정에서 추정 모델은 해당 핵산 서열에서의 유기체의 카테고리별 확률값(540)을 출력하고, 학습용 데이터 세트를 이용해서 supervised learning 방식으로 학습이 이루어질 수 있다.

[180] 예를 들어, fine-tuning의 과정에서, 추정 모델의 출력 데이터로서 유기체의 host의 생물학적 카테고리가 복수개의 생물학적 카테고리(620)(예: Homo sapiens, Sus scrofa, Bos taurus 등) 각각일 확률값(예: 93%, 2%, 0.3% 등)을 의미하는 카테고리별 확률값(540)을 출력하고, 각 핵산 서열에 정답으로서 라벨링되어 있는 생물학

적 카테고리(예: Homo sapiens)와 비교함으로써 에러를 계산할 수 있으며, 에러를 줄이기 위한 역전파에 따라 추정 모델의 파라미터들이 업데이트될 수 있다.

[181] 일 실시예에 따른 fine-tuning은 (i) 학습용 입력 데이터에 포함된 핵산 서열을 토큰화(tokenization)하여 복수의 토큰들을 획득하는 과정, (ii) 획득된 복수의 토큰들로부터 생성되는 컨텍스트 벡터(context vector)를 이용해서 학습용 입력 데이터에 포함된 핵산 서열의 host 또는 유기체에 대해 추정하는 과정 및 (iii) 추정된 결과와 학습용 정답 데이터 간의 차이가 줄어들도록 사전 학습된 모델(510)을 훈련시키는 과정을 포함할 수 있다. 여기서, 사전 학습된 모델(510)을 훈련시키는 과정은 전술한 사전 학습된 모델(510)이 포함된 구조를 갖는 추정 모델에 대한 훈련을 수행함으로써 학습된 추정 모델이 구현되는 과정에 대응될 수 있다. 상술한 추정 모델이 BERT를 포함하는 예시에서, BERT에 포함된 복수의 인코더들에 복수의 토큰들이 입력됨에 따라, 핵산 서열이 종합적으로 고려된 컨텍스트 벡터가 하나의 압축된 벡터 표현으로서 출력될 수 있다. 또한, 이러한 컨텍스트 벡터가 인코더와 연결된 분류기(classifier)에 입력되고, 분류기로부터 host 또는 유기체에 대해 예측된 결과로서 분류값 또는 확률값이 출력될 수 있다. 또한, 이러한 예측된 결과를 해당 학습용 정답 데이터에 포함된 host에 대한 라벨 데이터 또는 유기체에 대한 라벨 데이터와 비교함으로써 에러가 계산되고, 에러를 줄이기 위해 추정 모델의 파라미터들이 업데이트될 수 있다.

[182] 전술한 훈련 과정은 복수의 학습용 데이터 세트 각각을 이용해서 수행될 수 있으며, 이에 관한 보다 상세한 설명은 도 7을 참조하여 살펴보도록 한다.

[183] 도 7은 일 실시예에 따른 fine-tuning의 과정에서 BERT 기반의 추정 모델의 구조 및 동작을 예시적으로 도시한다.

[184] 도 7을 더 참조하면, 추정 모델은 BERT를 포함할 수 있다. 추정 모델이 BERT를 포함한다는 예시에서, BERT는 semi-supervised learning을 이용하여 사전 학습된 모델(510)에 supervised learning 기반의 fine-tunning이 적용되는 모델을 의미할 수 있다. 실시예에 따라서, 추정 모델은 ALBERT, RoBERTa 및 ELECTRA 등과 같은 BERT의 파생 모델을 포함할 수도 있다.

[185] 일 실시예에서, 추정 모델은 입력 임베딩 레이어(710), 사전 학습된 BERT(pre-trained BERT)(720) 및 제2 분류 레이어(730) 중 적어도 하나를 포함할 수 있다.

[186] 일 실시예에서, 사전 학습된 BERT(720)는 사전 학습된 모델(510)의 레이어들 중 적어도 일부를 포함할 수 있다. 예컨대, 추정 모델에 적용되는 사전 학습된 BERT(720)는 사전 학습에 의해 가중치가 기 연산된 인코더 레이어(320)가 그대로 포함될 수 있다.

[187] 일 실시예에서, 사전 학습된 BERT(720)는 사전 학습된 모델(510)을 적어도 부분적으로 변형하여 획득될 수 있다. 예컨대, 추정 모델에 적용되는 사전 학습된 BERT(720)에는 사전 학습에 의해 가중치가 기 연산된 인코더 레이어(320)가 포함되되, 해당 인코더 레이어(320) 내의 인코더들 중 일부는 가중치가 고정되도록 freeze 처리될 수 있다. 이로 인해, fine-tuning의 과정에서 해당 freezed layer(721)

의 가중치는 사전 학습에서의 가중치 값으로 고정되며, 해당 freezed layer(721)를 제외한 나머지의 가중치들만 추가 훈련을 통해 가중치 값이 갱신될 수 있다. 이에 따라, 학습 시간이 단축되고, overfitting 문제가 개선될 수 있다.

[188] 일 실시예에서, 입력 임베딩 레이어(710)는 입력되는 핵산 서열을 전처리하여 사전 학습된 BERT(720)에 제공할 수 있다. 입력 임베딩 레이어(710)은 일련의 입력 데이터인 핵산 서열을 인코더가 연산 가능한 형태로 변환할 수 있다. 실시예에 따라서, 입력 임베딩 레이어(710)은 사전 학습된 모델(510)의 입력 임베딩 레이어(310)와 대응될 수도 있다.

[189] 일 실시예에서, 입력 임베딩 레이어(710)은 서열 임베딩을 위한 라벨링된 서열 입력 레이어(711), 토큰화를 위한 토큰 임베딩 레이어(712), 세그먼트 정보 부가를 위한 세그먼트 임베딩 레이어(713), 및 위치 정보 부가를 위한 포지셔널(또는 포지션) 임베딩 레이어(714) 중 적어도 하나를 포함할 수 있다.

[190] 라벨링된 서열 입력 레이어(711)는 입력되는 핵산 서열을 임베딩하기 적절한 형태로 처리할 수 있다. 예컨대, 라벨링된 서열 입력 레이어(711)는 핵산 서열의 첫 번째 위치에 서열 시작을 나타내는 특수 토큰인 CLS(Special Classification token)을 추가하고, 핵산 서열의 마지막 위치에 서열 끝을 나타내는 특수 토큰인 SEP(Special Classification token) 토큰을 추가할 수 있다.

[191] 토큰 임베딩 레이어(712)는 핵산 서열을 토큰화하여 복수의 토큰들을 획득하는 토큰화 과정을 수행할 수 있다. 일 실시예에서, 토큰 임베딩 레이어(712)에서 수행되는 토큰화 과정은 사전 학습의 과정에서 기술한 토큰화 과정의 실시예들을 포함할 수 있다. 일 예로, 토큰 임베딩 레이어(712)는 라벨링된 서열 입력 레이어(711)에서 출력되는 입력 서열에 포함된 베이스들을 k-mer 기법으로 분할하여 토큰화하거나, 기능 단위의 유전자 추정 기법으로 slicing하여 토큰화할 수 있다. 또한, 토큰 임베딩 레이어(712)는 토큰화 과정에서 각 토큰을 벡터로 처리할 수 있다.

[192] 세그먼트 임베딩 레이어(713)는 복수의 토큰들에 서열 구분을 위한 구분 정보가 인가되도록 복수의 토큰들을 처리할 수 있다. 일 실시예에서, 세그먼트 임베딩 레이어(713)는 하나의 핵산 서열로부터 획득된 두 개 이상의 부분 서열이 함께 입력되는 경우, 각 부분 서열을 구분하기 위해 상이한 벡터 표현을 사용하여 세그먼트 임베딩을 수행할 수 있다. 예컨대, 입력되는 부분 서열이 2개인 경우, 2개의 벡터 표현이 사용될 수 있으며, 첫 번째 벡터(인덱스 0)는 제1 부분 서열에 속하는 모든 토큰에 할당되고, 마지막 벡터(인덱스 1)는 제2 부분 서열에 속하는 모든 토큰에 할당될 수 있다. 다른 실시예에서, 추정 모델에 핵산 서열 또는 부분 서열이 하나씩 입력되는 경우, 세그먼트 임베딩 레이어(713)는 생략되거나 동일한 벡터 표현이 사용될 수 있다.

[193] 포지셔널(또는 포지션) 임베딩 레이어(714)는 상술한 토큰 임베딩 및 세그먼트 임베딩을 통해 복수의 토큰들로부터 생성되는 복수의 임베딩 벡터들이 인코더의 입력으로 사용되기 전에, 임베딩 벡터들에 위치 정보를 인가할 수 있다.

[194] 상술한 입력 임베딩 레이어(710)를 거치면서 인코더가 연산가능한 형태로 가공된 임베딩된 데이터들은 사전 학습된 BERT(720)에 입력될 수 있다. 사전 학습된 BERT(720)는 입력된 임베딩된 데이터들을 종합적으로 고려한 context vector를 출력할 수 있다.

[195] 일 실시예에서, 제2 분류 레이어(730)는 사전 학습된 BERT(720)에서 출력된 결과를 이용해서 host 또는 유기체의 생물학적 카테고리를 추정할 수 있다. 예컨대, 제2 분류 레이어(730)는 사전 학습된 BERT(720)의 마지막 인코더 블록에서 출력되는 임베딩 벡터(예: context vector)를 입력으로 사용하여 host의 카테고리별 확률값(740)을 출력하기 위한 분류 기능을 수행하는 classifier를 포함할 수 있다. 일 예로, 제2 분류 레이어(720)는 classifier를 통해 복수개의 생물학적 카테고리(620)에 대응되는 복수개의 클래스 각각에 대한 host의 카테고리별 확률값(540)을 출력할 수도 있고, 또는, 복수개의 클래스 중에서 확률값이 가장 크거나 기설정 기준값보다 큰 하나 이상의 클래스에 대한 카테고리별 확률값(540)을 출력할 수도 있다.

[196] 일 실시예에서, 제2 분류 레이어(730)는 host 또는 유기체의 생물학적 카테고리를 추정하기 위한 FC(fully connected) 신경망과 소프트맥스 함수를 포함할 수 있다. 제2 분류 레이어(730)는 host의 카테고리별 확률값(740) 또는 유기체의 카테고리별 확률값의 결과 출력을 위한 분류 기능을 수행하도록 구성될 수 있다. 예컨대, 사전 학습된 BERT(720)로부터 출력되는 모든 임베딩 벡터들은 FC 구조로 feed forward 신경망에 입력될 수 있고, 피드 포워드 신경망의 출력층에 활성 함수로서 소프트맥스 함수가 사용될 수 있다. 신경망에서 출력되는 특정 차원의 벡터는 소프트맥스 함수를 지나면서 0과 1 사이의 실수값을 갖고 총 합은 1인 벡터로 변환될 수 있고, 해당 벡터가 각 카테고리별 확률값(740)으로서 출력될 수 있다.

[197] 일 실시예에서, 제2 분류 레이어(730)에서 출력된 host의 카테고리별 확률값(740) 또는 유기체의 카테고리별 확률값은 학습용 정답 데이터인 host 또는 유기체의 생물학적 카테고리에 대한 라벨 데이터와 비교될 수 있고, 비교 결과에 따라 에러가 최소화되도록 추정 모델은 추가 훈련될 수 있다.

[198] 한편, 추정 모델에 제공되는 핵산 서열은, 베이스들의 개수가 사전설정된 범위 내에 있을 수 있다. 구체적으로, 추정 모델에 입력되는 핵산 서열에 포함된 베이스들의 개수는 사전설정된 제1 cutoff 이상 또는 제2 cutoff 이하일 수 있다. 예컨대, 추정 모델의 입력 임베딩 레이어(710)에 입력되는 서열 내 베이스들의 개수는 사전설정된 제1 cutoff 이상이거나 제2 cutoff 이하일 수 있고, 제1 cutoff 이상이면서 제2 cutoff 이하일 수 있다.

[199] 일 실시예에서, 제1 cutoff는 추정 모델에 입력되는 서열의 베이스들의 개수에 대한 최소 기준 값이다. 예컨대, 제1 cutoff는 1,500 bp(base pair) 내지 2,000 bp 사이의 값으로 설정될 수 있다.

[200] 일 실시예에서, 제2 cutoff는 추정 모델에 입력되는 서열의 베이스들의 개수에 대한 최대 기준 값이다. 예컨대, 제2 cutoff는 10,000 bp 내지 20,000 bp 사이의 값으로 설정될 수 있다.

[201] 일 실시예에서, 제1 cutoff 및/또는 제2 cutoff는 fine-tuning의 과정에서 추정 모델의 성능이 최소 목표 수준을 달성하였는지 여부에 기초하여 변경될 수 있다. 예컨대, 학습용 데이터 세트들 중 사전설정된 비율의 데이터 세트들을 이용해서 추가 훈련된 추정 모델의 추정 정확도가 기준값 미만일 경우, 제1 cutoff는 추정 정확도를 향상시키기 위해 소정의 비율만큼 상향될 수 있다.

[202] 일 실시예에서, 제1 cutoff 및/또는 제2 cutoff는, 타겟으로 하는 유기체의 종류가 사전설정된 경우, 해당 종류에 속하는 유기체의 유전체(genome) 길이, 기준 서열의 길이, 기준 서열의 개수 및 다양성 정보 중 적어도 하나에 기초하여 결정될 수 있다. 여기서, 기준 서열은 해당 생명체를 대표하기 위한 핵산 서열을 의미하며, 예컨대, 해당 유기체의 종류에 대해서 최초로 발견된 염기 서열 또는 가장 많은 비중을 차지하거나 사전설정된 비중 이상을 차지하는 염기 서열을 포함할 수 있다. 또한, 다양성 정보는 예컨대, 핵산 서열의 기준 서열과 다른 서열들 간의 차이에 관한 서열 차이 수준, 핵산 서열 내에서 서열 차이 수준이 상대적으로 큰 비보존성 영역(non-conserved region), 및 상대적으로 낮은 보존성 영역(conserved region)의 범위나 개수 등을 포함할 수 있다. 일 예로, 타겟으로 하는 바이러스 과에 속하는 바이러스들의 유전체의 평균적인 베이스들 개수가 많을수록, 기준 서열의 베이스들 개수가 많을수록, 기준 서열이 소수 개일수록, 비보존성 영역의 범위가 넓거나 영역 개수가 많을수록, 보존성 영역의 범위가 좁거나 영역 개수가 작을수록, 제1 cutoff 및/또는 제2 cutoff는 상대적으로 큰 값으로 설정될 수 있으며, 이에 따라 추정 모델의 추정 정확도가 증가할 수 있다.

[203] 일 실시예에서, 제2 cutoff는 추정 모델에 입력 가능한 최대 값에 기초하여 설정될 수도 있다. 예컨대, 제2 cutoff는 추정 모델에서 연산 가능한 최대 토큰 사이즈를 고려하여 설정될 수 있다.

[204] 상술한 것처럼, 베이스들의 개수가 상기 사전설정된 범위 내에 있는 경우, 컴퓨터 장치(100)는 핵산 서열을 추정 모델에 입력하고, 추정 모델은 입력된 핵산 서열을 이용해 추정한 host(또는 유기체)와 학습용 정답 데이터의 host에 대한 라벨 데이터(또는 유기체에 대한 라벨 데이터)를 비교하는 방식으로 추가 훈련될 수 있다.

[205] 그러나, 베이스들의 개수가 상술한 사전설정된 범위를 벗어나는 경우, 실시예에 따라서, 컴퓨터 장치(100)는 핵산 서열을 부분 서열로 나누는 전처리를 수행하고, 부분 서열을 추정 모델에 입력할 수 있다. 이러한 경우, 추정 모델은 입력된 부분 서열을 이용해 추정한 host(또는 유기체)와 학습용 정답 데이터의 host에 대한 라벨 데이터(또는 유기체에 대한 라벨 데이터)를 비교하는 방식으로 추가 훈련될 수 있다. 이에 대해서는 도 8을 더 참조하여 설명하도록 한다.

[206] 도 8은 일 실시예에 따라 핵산 서열이 부분 서열로 전처리되는 과정을 예시적으로 도시한다.

[207] 도 8에 도시된 것처럼, 컴퓨터 장치(100)는 fine-tuning의 과정에서 학습용 입력 데이터인 핵산 서열(810)에 포함된 베이스들의 개수가 제2 cutoff를 초과하는 경우, 제2 cutoff 개수 이하의 베이스들이 포함되도록 핵산 서열(810)로부터 하나 이상의 부분 서열(820)을 획득할 수 있다.

[208] 일 실시예에서, 컴퓨터 장치(100)는 핵산 서열(810)로부터 제2 cutoff 개수 이하의 베이스들을 포함하는 하나의 제1 부분 서열(821)을 획득할 수 있다. 예컨대, 컴퓨터 장치(100)는 핵산 서열(810)(예: 30,000bp) 중에서 사전설정된 시작 지점(예: 300bp 내지 400 bp 지점)을 기준으로 제2 cutoff 개수(예: 20,000bp)의 베이스들을 포함하는 단 하나의 제1 부분 서열(821)을 추출할 수 있다. 또한, 컴퓨터 장치(100)는 제1 부분 서열(821)만 추정 모델에 제공하고, 제1 부분 서열(821)을 제외한 나머지 서열은 추정 모델에 제공하지 않을 수 있다.

[209] 전술한 것처럼, 부분 서열(820)이 단수개인 경우, 추정 모델은 해당 부분 서열(820)에 대한 host(또는 유기체)를 추정하고, 추정된 host(또는 유기체)와 학습용 정답 데이터를 비교한 결과에 따라 훈련될 수 있다. 예컨대, 추정 모델은 제1 부분 서열(821)을 이용해서 host를 추정하고, 추정된 host와 핵산 서열(810)의 host에 대한 라벨 데이터를 비교한 결과에 따라 모델의 가중치를 업데이트할 수 있다.

[210] 다른 실시예에서, 컴퓨터 장치(100)는 핵산 서열(810)로부터 각각 제2 cutoff 개수 이하의 베이스들을 포함하는 복수개의 부분 서열(820)을 획득할 수 있다. 예컨대, 컴퓨터 장치(100)는 핵산 서열(810)(예: 30,000bp) 중에서 시작 지점을 기준으로 제2 cutoff 개수(예: 20,000bp)의 베이스들을 포함하는 제1 부분 서열(821)과, 끝 지점을 기준으로 제2 cutoff 개수(예: 20,000bp)의 베이스들을 포함하는 제2 부분 서열(822)을 분할할 수 있다. 또한, 컴퓨터 장치(100)는 분할된 제1 부분 서열(821) 및 제2 부분 서열(822)을 각각 추정 모델에 제공할 수 있다.

[211] 또 다른 실시예에서, 컴퓨터 장치(100)는 핵산 서열(810)로부터 각각 제1 cutoff 개수 이상 제2 cutoff 개수 이하의 베이스들을 포함하는 복수개의 부분 서열(820)을 획득할 수 있다. 일 예로, 컴퓨터 장치(100)는 핵산 서열(810)(예: 30,000bp) 중에서 시작 지점을 기준으로 제2 cutoff 개수(예: 20,000bp)의 베이스들을 포함하는 제1 부분 서열(821)과, 분할된 제1 부분 서열(821)의 끝 지점을 기준으로 제1 cutoff 개수(예: 2,000bp) 이상 제2 cutoff 개수(예: 20,000bp) 이하인 나머지 베이스들(예: 10,000bp)을 포함하는 제2 부분 서열(822)을 분할할 수 있다. 다른 일 예로, 컴퓨터 장치(100)는 핵산 서열(810)(예: 30,000bp)를 절반으로 나누어 제1 부분 서열(821) 및 제2 부분 서열(822)로 분할할 수도 있다. 또한, 마찬가지로, 컴퓨터 장치(100)는 분할된 제1 부분 서열(821) 및 제2 부분 서열(822)을 각각 추정 모델에 제공할 수 있다.

[212] 전술한 것처럼, 부분 서열(820)이 복수개인 경우, 일 실시예에 따른 추정 모델은 해당 복수개의 부분 서열(820) 각각에 대한 host(또는 유기체)를 추정하고, 복수

개의 부분 서열(820) 각각에 대해 추정된 host(또는 유기체)를 통계 처리할 수 있다. 또한, 추정 모델은 통계 처리한 결과를 이용해서 핵산 서열(810)에 대해 추정된 host(또는 유기체)를 결정하고, 핵산 서열(810)에 대해 추정된 host(또는 유기체)와 학습용 정답 데이터를 비교한 결과에 따라 훈련될 수 있다. 예컨대, 추정 모델은 제1 부분 서열(821)을 이용해서 추정된 제1 host, 제2 부분 서열(822)을 이용해서 추정된 제2 host, 내지 제P 부분 서열(P는 3 이상의 자연수)을 이용해서 추정된 제P host에 대해 통계 처리한 결과에 따라 전체 서열인 핵산 서열(810)에 대한 host를 결정하고, 결정된 host와 핵산 서열(810)의 host에 대한 라벨 데이터를 비교한 결과에 따라 모델의 가중치를 업데이트하는 방식으로 훈련될 수 있다.

[213] 상술한 통계 처리에는, majority vote 방식, 평균 방식 및 표준 편차 방식 중 적어도 하나가 포함될 수 있다. 일 실시예에서, majority vote 방식은 복수개의 부분 서열(820) 각각에 대해 추정된 host(또는 유기체) 중에서, 가장 많은 수를 차지하는 host(또는 유기체)를 선택하는 simple majority vote 방식, 및 과반수를 차지하는 host(또는 유기체)를 선택하는 absolute majority vote 방식 중 적어도 하나를 포함할 수 있다. 일 실시예에서, 평균 방식(또는 표준 편차 방식)은 복수개의 부분 서열(820) 각각에 대해 추정된 host(또는 유기체)의 카테고리별 확률값을 평균하여 카테고리별 평균값(또는 표준 편차값)을 구하고, 카테고리별 평균값(또는 표준 편차값)이 가장 크거나 기준값 이상인 host(또는 유기체)를 선택하는 방식일 수 있다. 위에서는 majority vote 방식, 평균 방식 및 표준 편차 방식이 예시되었으나, 이에 제한되지 않으며, 다양한 통계 처리 방식이 이용될 수 있다.

[214] 부분 서열(820)이 복수개인 경우, 다른 일 실시예에 따른 추정 모델은 복수개의 부분 서열(820) 각각에 대해 추정된 host(또는 유기체)와 학습용 정답 데이터를 각각 비교한 결과에 따라 각각 훈련될 수 있다. 예컨대, 추정 모델은 전술한 제1 host와 핵산 서열(810)의 host에 대한 라벨 데이터를 비교한 결과에 따라 모델의 가중치를 업데이트하고, 제2 host와 핵산 서열(810)의 host에 대한 라벨 데이터를 비교한 결과에 따라 모델의 가중치를 업데이트하는 방식으로 각각에 대해 추가 훈련할 수 있다.

[215] 실시예에 따라서, 핵산 서열(810)로부터 하나 이상의 부분 서열(820)를 나누기 위한 기준 위치로서 활용되는 상술한 지점들(예: 시작 지점, 끝 지점 등)은 fine-tuning의 과정에서 변경될 수 있다. 일 예로, 상술한 지점들은 매번 정해진 범위 내에서 랜덤한 값으로 설정될 수 있다. 다른 일 예로, 상술한 시작 지점이나 끝 지점은 일종의 하이퍼 파라미터로서 관리될 수 있으며, 예컨대, 서로 다른 시작 지점이나 끝 지점의 값으로 사전 학습된 모델(510)을 fine-tuning하여 구현된 복수개의 추정 모델 후보들 중에서 더 높은 추정 정확도를 보이는 시작 지점이나 끝 지점의 값이 추정 모델에 적용될 수 있다.

[216] 전술한 것처럼, 컴퓨터 장치(100)는 핵산 서열의 길이가 소정의 길이 이상인 경우, 서열들을 정해진 규칙에 따라 또는 무작위로 여러 부분 서열로 나누어 전처

리한 후 추정 모델에 제공할 수 있다. 이에 따라, 추정 모델에 입력되는 서열의 길이가 유사 범위 내에 있도록 조정될 수 있으며, 추정 정확도가 향상될 수 있다.

[217] 한편, 실시예에 따라서, 상술한 fine-tuning의 과정에서 host 정보와 유기체 정보 중 적어도 어느 하나가 이용될 수 있으며, fine-tuning의 결과로서 host 추정 및 유기체 추정 중 적어도 어느 하나를 위한 추정 모델이 구현될 수 있다.

[218] 제1 실시예에서, 추정 모델은 host 정보를 이용해서 핵산 서열의 host를 추정하도록 fine-tuning될 수 있다. 예컨대, fine-tuning의 과정에 이용되는 학습용 정답 데이터는 해당 핵산 서열의 host의 생물학적 카테고리에 대한 라벨 데이터를 포함할 수 있다. 또한, 이러한 학습용 데이터 세트들을 이용해서, 도 5에 도시된 것처럼, 사전 학습된 모델(510)에 host 추정용 레이어(520)를 적용한 추정 모델이 훈련될 수 있으며, 학습 결과로서 핵산 서열이 입력되면 해당 host의 카테고리별 확률값을 추정하도록 학습된 추정 모델이 구현될 수 있다.

[219] 제2 실시예에서, 추정 모델은 유기체 정보를 이용해서 핵산 서열을 포함하는 유기체를 추정하도록 fine-tuning될 수 있다. 예컨대, fine-tuning의 과정에 이용되는 학습용 정답 데이터는 해당 핵산 서열을 포함하는 유기체의 생물학적 카테고리에 대한 라벨 데이터를 포함할 수 있다. 또한, 이러한 학습용 데이터 세트들을 이용해서, 도 5에 도시된 것과 유사한 방식으로, 사전 학습된 모델(510)에 유기체 추정용 레이어(식별번호 520 참조)를 적용한 추정 모델이 훈련될 수 있으며, 학습 결과로서 핵산 서열이 입력되면 해당 유기체의 카테고리별 확률값을 추정하도록 학습된 추정 모델이 구현될 수 있다.

[220] 제3 실시예에서, 추정 모델은 host 정보와 유기체 정보를 이용해서 핵산 서열의 host와 유기체를 추정하도록 fine-tuning될 수 있다. 일 예로, fine-tuning의 과정에 이용되는 학습용 정답 데이터는 해당 핵산 서열을 포함하는 유기체의 생물학적 카테고리에 대한 라벨 데이터와 해당 유기체의 host의 생물학적 카테고리에 대한 라벨 데이터를 포함할 수 있다. 또한, 이러한 학습용 데이터 세트들을 이용해서 사전 학습된 모델(510)에 추정용 레이어(식별번호 520 참조)를 적용한 추정 모델이 훈련될 수 있으며, 학습 결과로서 핵산 서열이 입력되면 해당 host의 카테고리별 확률값과 해당 유기체의 카테고리별 확률값을 함께 추정하도록 학습된 추정 모델이 구현될 수 있다. 이러한 경우, 입력으로서 핵산 서열만을 가지고 host 종과 유기체 종을 함께 추정할 수 있으므로, 핵산 서열의 수집 과정에서 함께 획득된 host와 유기체에 대한 정보들이 정확한지 보다 효율적으로 검증할 수 있다.

[221] 제4 실시예에서, 추정 모델은 host 정보와 유기체 정보를 이용해서 핵산 서열의 host 또는 유기체를 추정하도록 fine-tuning될 수 있다.

[222] 예를 들면, fine-tuning 과정에서 학습용 입력 데이터로서 핵산 서열과 해당 유기체의 생물학적 카테고리가 이용되고, 학습용 정답 데이터로서 해당 핵산 서열에 대한 host의 생물학적 카테고리가 이용될 수 있다. 이에 따라, 추정 모델은 핵산 서열과 해당 유기체의 생물학적 카테고리가 함께 입력되면 두 입력을 모두 고려하여 해당 host의 카테고리별 확률값을 출력하고, 출력을 학습용 정답 데이터와

비교하여 해당 host 종을 맞추도록 학습이 이루어질 수 있다. 이러한 경우, 입력으로서 핵산 서열 이외에도 해당 유기체 정보를 함께 활용할 수 있으므로, host를 보다 정확하게 추정할 수 있다.

[223] 다른 예를 들면, 추정 모델은 host 정보를 이용해서 host를 추정하도록 fine-tuning되되, 사전 설정된 복수개의 유기체의 종류에 따라 마련될 수 있다. 예컨대, 바이러스 계열에 속하는 유기체들의 학습용 데이터로 학습된 제1 사전 학습된 모델(510)을 fine-tuning한 제1 추정 모델, 박테리아 계열에 속하는 유기체들의 학습용 데이터로 학습된 제2 사전 학습된 모델(510)을 fine-tuning한 제2 추정 모델, 및 진균 계열에 속하는 유기체들의 학습용 데이터로 학습된 제3 사전 학습된 모델(510)을 fine-tuning한 제3 추정 모델 등이 구현될 수 있다. 또는, 대용량의 유기체들의 학습용 데이터로 학습된 사전 학습된 모델(510)을 바이러스 계열에 속하는 유기체들의 host를 추정하도록 fine-tuning한 제4 추정 모델, 해당 사전 학습된 모델(510)을 박테리아 계열에 속하는 유기체들의 host를 추정하도록 fine-tuning한 제5 추정 모델, 진균 계열에 속하는 유기체들의 host를 추정하도록 fine-tuning한 제6 추정 모델 등이 구현될 수 있다. 이와 같이, 바이러스 계열이나 박테리아 계열 등과 같이 사전에 정해진 여러 범주에서 병원체 계열을 분류하여 각각에 최적화되도록 fine-tuning된 추정 모델을 구현함으로써, 각 병원체 계열별로 핵산 서열의 특성을 고려하여 host를 보다 정확하게 추정할 수 있다.

[224] 한편, 전술한 예시와는 다른 일 실시예로서, 추정 모델은 유기체의 종류에 무관하게 마련될 수 있다. 예컨대, 사전 학습과 fine-tuning 과정에서 유기체의 종류가 제한되지 않은 핵산 서열들을 학습용 데이터 세트로 이용함으로써, 하나의 추정 모델이 구현될 수 있다. 이러한 경우, 하나의 추정 모델로 모든 종류의 유기체의 핵산 서열로부터 host 또는 유기체를 추정할 수 있다.

[225] 한편, 상술한 사전 학습 및 fine-tuning 각각의 과정에서 복수개의 하이퍼 파라미터들이 이용될 수 있다. 하이퍼 파라미터는 사용자에 의해 가변되는 변수일 수 있다. 하이퍼 파라미터는 예를 들어, 학습률(learning rate), 비용 함수(cost function), 학습 사이클 반복 횟수, 가중치 초기화(Weight initialization)(예를 들어, 가중치 초기화 대상이 되는 가중치 값의 범위 설정), Hidden Unit 개수(예를 들어, 히든 레이어의 개수, 히든 레이어의 노드 수)를 포함할 수 있다. 또한, 하이퍼 파라미터는 상술한 토큰화 기법(예: k-mer, gene prediction), k-mer 기법인 경우 k의 설정값, 경사 하강법 학습시 스텝 크기(gradient accumulation step), 배치 사이즈(batch size), 드롭아웃(drop out) 등을 더 포함할 수 있다.

[226] 한편, 명세서 전반에서 추정 모델이 BERT 기반의 전이학습 방식으로 구현되는 실시예들을 주로 기술하였으나, 위에서 이미 기술한 바와 같이, 추정 모델은 이에 제한되지 않으며, 신경망 기반의 모델로서 그밖의 다양한 뉴럴 네트워크(예: CNN, RNN 등)들을 통해 구현될 수 있다.

[227] 전술한 것처럼, 추정 모델은 이하에서 후술될 바와 같이, 핵산 서열이 입력되면, 해당 핵산 서열을 포함하는 유기체의 host 또는 해당 유기체에 대한 추정 결과를

출력하도록 전이학습될 수 있다. 컴퓨터 장치(100)는 추정 모델을 저장 및 관리할 수 있고, 추정 모델을 제공할 수 있다. 예컨대, 컴퓨터 장치(100)는 서버에 의해 전이학습 방식으로 학습된 추정 모델이 저장 및 관리되며, 사용자 단말이 추정 모델을 요청함에 따라 서버가 추정 모델을 사용자 단말에 제공하도록 구현될 수 있다.

[228]

[229] 도 9는 일 실시예에 따라 추정 모델을 획득하기 위한 예시적인 흐름도를 도시한다. 도 9는 도 1 내지 도 8에 도시된 실시예들을 참조하여 이해될 수 있다.

[230] 일 실시예에서, 도 9에서 도시되는 단계들은 컴퓨터 장치(100)에 의해 수행될 수 있다. 추가적인 일 실시예에서, 도 9에서 도시되는 단계들은 서버에서 수행되는 방식과 같이 하나의 엔티티에 의해 구현될 수 있다. 추가적인 다른 실시예에서, 도 9에서 도시되는 단계들 중 일부는 사용자 단말에서 수행되고 다른 일부는 서버에서 수행되는 방식과 같이 도 9에서의 단계들은 복수개의 엔티티들에 의해 구현될 수도 있다.

[231] 단계 S910에서 컴퓨터 장치(100)는 사전 학습된 모델(510)을 획득할 수 있다. 일 실시예에서, 컴퓨터 장치(100)는 복수개의 핵산 서열(220)을 획득하고, 복수개의 핵산 서열(220)에 대한 사전 학습을 수행하여 사전 학습된 모델(510)을 획득할 수 있다. 다른 일 실시예에서, 컴퓨터 장치(100)는 다른 장치에 의해 기 사전 학습된 모델(510)을 네트워크를 통해 다른 장치(또는 저장부 등)로부터 수신할 수 있다.

[232] 일 실시예에서, 사전 학습된 모델(510)은 학습용 데이터로서 복수개의 핵산 서열(220)을 이용할 수 있다. 일 실시예에서, 사전 학습된 모델(510)은 복수개의 핵산 서열(220) 각각에 포함된 베이스들 중 일부의 베이스에 마스크를 적용한 뒤, 마스킹된 베이스를 맞추는 semi-supervised learning 방식에 의해 학습된 것일 수 있다. 일 실시예에서, 사전 학습된 모델(510)은 각각 두 개 이상의 베이스들을 갖는 토큰들로 토큰화되는 핵산 서열을 이용해서 학습된 것일 수 있다. 일 실시예에서, 상기 토큰들은 (i) 핵산 서열을 k 개씩 분할하거나 (ii) 핵산 서열을 기능 단위로 분할하여서 토큰화되는 베이스들을 각각 포함할 수 있다.

[233] 단계 S920에서 컴퓨터 장치(100)는 사전 학습된 모델(510)을 fine-tuning하여서 추정 모델을 획득할 수 있다. 일 실시예에서, 컴퓨터 장치(100)는 복수의 학습용 데이터 세트를 획득하고, 복수의 학습용 데이터 세트를 이용해서 사전 학습된 모델(510)에 대한 fine-tuning을 수행할 수 있다.

[234] 일 실시예에서, fine-tuning은 복수의 학습용 데이터 세트를 이용해서 수행되되, 각 학습용 데이터 세트는 (i) 핵산 서열을 포함하는 학습용 입력 데이터 및 (ii) 해당 핵산 서열을 포함하는 유기체 또는 해당 유기체의 host에 대한 라벨 데이터를 포함하는 학습용 정답 데이터를 포함할 수 있다. 일 실시예에서, fine-tuning은 (i) 학습용 입력 데이터에 포함된 핵산 서열을 토큰화하여 복수의 토큰들을 획득하는 과정, (ii) 복수의 토큰들로부터 생성되는 컨텍스트 벡터를 이용해서 학습용 입력 데이터에 포함된 핵산 서열의 유기체 또는 host에 대해 추정하는 과정 및

(iii) 추정된 결과와 학습용 정답 데이터 간의 차이가 줄어들도록 사전 학습된 모델(510)을 훈련시키는 과정을 포함할 수 있다.

[235] 제1 실시예에서, 컴퓨터 장치(100)는 핵산 서열 및 host에 대한 라벨 데이터를 포함하는 각 학습용 데이터 세트를 이용해서 사전 학습된 모델(510)을 fine-tuning하고, 학습 결과로서 핵산 서열이 입력되면 해당 핵산 서열의 host를 추정하도록 학습된 추정 모델을 획득할 수 있다.

[236] 제2 실시예에서, 컴퓨터 장치(100)는 핵산 서열 및 유기체에 대한 라벨 데이터를 포함하는 각 학습용 데이터 세트를 이용해서 사전 학습된 모델(510)을 fine-tuning하고, 학습 결과로서 핵산 서열이 입력되면 해당 핵산 서열의 유기체를 추정하도록 학습된 추정 모델을 획득할 수 있다.

[237] 제3 실시예에서, 컴퓨터 장치(100)는 핵산 서열, host에 대한 라벨 데이터 및 유기체에 대한 라벨 데이터를 포함하는 각 학습용 데이터 세트를 이용해서 사전 학습된 모델(510)을 fine-tuning하고, 학습 결과로서 핵산 서열이 입력되면 해당 핵산 서열의 host와 유기체를 추정하도록 학습된 추정 모델을 획득할 수 있다.

[238] 제4 실시예에서, 컴퓨터 장치(100)는 기 설정된 복수개의 유기체 종류별로 분류된 학습용 데이터를 이용해서 학습된 복수개의 사전 학습된 모델(510)을 획득하고, 핵산 서열 및 host에 대한 라벨 데이터를 포함하는 학습용 데이터 세트들을 이용해서 복수개의 사전 학습된 모델(510) 각각을 fine-tuning한 결과로서 복수개의 추정 모델을 획득할 수 있다. 또는, 컴퓨터 장치(100)는 하나의 사전 학습된 모델(510)을 가져오되, 기 설정된 복수개의 유기체 종류별로 분류된 학습용 데이터 세트들을 각각 이용해서 fine-tuning하고, 각각의 학습 결과로서 복수개의 유기체 종류별로 fine-tuning된 복수개의 추정 모델을 획득할 수 있다.

[239] 제5 실시예에서, 컴퓨터 장치(100)는 제1 실시예에 따른 host 추정 모델 및 제2 실시예에 따른 유기체 추정 모델을 각각 획득하고, host 추정 모델 및 유기체 추정 모델을 포함하는 추정 모델을 획득할 수 있다. 실시예에 따라서, host 추정 모델과 유기체 추정 모델은 상이한 모델의 구조를 가지거나, 각 모델을 위한 뉴럴 네트워크의 종류가 상이할 수도 있다.

[240] 이상에서는, 추정 모델이 전이학습 방식으로 학습되는 과정의 다양한 실시예들에 대해 살펴보았다. 위에서 살펴본 바와 같이, 일 실시예에 따라 분자 진단 시약에 사용되는 올리고뉴클레오타이드(예: 프라이머, 프로브)의 서열들이 타겟 분석물(예: 바이러스)의 서열들의 일부인 점을 고려하여, 유기체의 핵산 서열들에 대한 사전 학습 및 host 추정(또는 유기체 추정)을 위한 fine-tuning을 통해 추정 모델을 획득하는 기술적 특징이 제안된다. 이에 따라, 상대적으로 적은 양의 labeled data를 이용하더라도 높은 추정 성능을 갖는 추정 모델을 구현할 수 있다. 또한, 생명체의 종류에 따라 여러 개의 추정 모델이 마련될 필요가 없으며, 하나의 추정 모델만으로 여러 다양한 종류의 유기체의 핵산 서열로부터 해당 host 또는 해당 유기체를 추정 가능한 효과가 있다.

[241]

[242] 이하에서는, 위에서 제시한 추정 모델을 이용하여 host 또는 유기체를 추정하는 과정의 다양한 실시예들에 대해 서술하도록 한다.

[243] 컴퓨터 장치(100)는 추정 모델을 제공받을 수 있다. 일 실시예에서, 컴퓨터 장치(100)는 서버에서 전이 학습 방식으로 추정 모델을 학습시키면 사용자 단말이 서버로부터 추정 모델을 수신하도록 구현될 수 있다.

[244] 컴퓨터 장치(100)는 추정을 위한 핵산 서열을 획득할 수 있다. 여기서, 핵산 서열은 host 또는 유기체를 알고자 하는 핵산 서열 중 적어도 일부를 포함한다. 일 실시예에서, 컴퓨터 장치(100)는 핵산 서열을 공개 데이터베이스나 메모리(110)로부터 또는 사용자 입력을 통해 획득할 수 있다. 예컨대, 컴퓨터 장치(100)는 NCBI나 GISAID와 같은 공개 데이터베이스에 등재되어 있는 생명체(예: Rotavirus A)의 핵산 서열을 수신하여 메모리(110)에 저장할 수 있다.

[245] 일 실시예에서, 컴퓨터 장치(100)는 핵산 서열과 해당 핵산 서열의 유기체 및/또는 host에 대한 정보를 포함하는 서열 관련 정보를 획득할 수 있으며, 서열 관련 정보로부터 핵산 서열을 획득할 수 있다. 예컨대, 핵산 서열과 해당 유기체의 종류 및/또는 해당 host의 종류를 포함하는 서열 관련 정보가 공개 데이터베이스로부터 수집될 수 있으며, 이하의 과정에서 추정 모델을 통해 해당 핵산 서열의 유기체 및/또는 host의 종류에 대한 검증이 이루어질 수 있다.

[246] 컴퓨터 장치(100)는 추정 모델에 핵산 서열을 제공하여서, 추정 모델로부터 해당 핵산 서열을 포함하는 유기체 또는 해당 유기체의 host를 추정할 수 있다. 상술한 것처럼, 해당 핵산 서열을 포함하는 유기체는 해당 핵산 서열에 대응되는 유기체로서, 예컨대, 해당 핵산 서열을 유전 물질로서 적어도 부분적으로 포함하는 유기체를 지칭할 수 있다. 컴퓨터 장치(100)는 추정 모델에 유기체 또는 host를 추정하고자 하는 핵산 서열을 입력하고, 추정 모델은복수개의 생물학적 카테고리(620) 중에서 해당 핵산 서열을 포함하는 유기체 또는 host의 카테고리를 출력할 수 있다.

[247] 도 10은 일 실시예에 따라 추정 모델(1010)의 추론(inference) 동작에 대한 개념도를 예시적으로 도시한다. 일 실시예에서, 도 10에서 도시되는 추정 모델(1010)은 사전 학습된 모델(510)에 대한 fine-tuning 또는 전이학습이 완료된 모델을 의미할 수 있다. 마찬가지로, 도 10에는 일 실시예에 따른 host 추정을 위한 추론 과정이 도시되어 있으나, 이와 유사한 방식으로 다른 일 실시예에 따른 유기체 추정을 위한 추론 과정이 수행될 수 있다.

[248] 도 10을 참조하면, 추정 모델(1010)에는 핵산 서열(1020)이 입력될 수 있다. 상술한 것처럼, 핵산 서열(1020)은 host 또는 유기체를 알고자 하는 핵산 서열로서, 예컨대, 유기체의 유전체 중 적어도 일부의 영역에 위치한 베이스들을 나열한 서열 정보를 의미할 수 있다.

[249] 일 실시예에서, 핵산 서열(1020)은 해당 핵산 서열을 포함하는 유기체의 종류가 제한되어 있지 않을 수 있다. 예컨대, fine-tuning의 과정에서 특정 유기체 종(예:

Rotavirus A, Influenza virus A)의 서열들이 이용되었다면, 추정 모델(1010)에는 여러 다양한 유기체 종의 서열들이 입력으로서 제공될 수 있다.

[250] 다른 실시예에서, 핵산 서열(1020)은 해당 핵산 서열을 포함하는 유기체의 종류가 일부 제한될 수도 있다. 예컨대, fine-tuning의 과정에서 특정 유기체 종(예: Rotavirus A, Influenza virus A)의 서열들이 이용된 경우, 해당 종을 포괄하는 상위 hierarchical level의 기 설정된 생물학적 카테고리(예: 바이러스 목)에 속하는 유기체의 핵산 서열들이 추정 모델(1010)에 입력으로서 제공되거나 이를 권장하는 메시지가 출력될 수 있다.

[251] 일 실시예에서, 추정 모델(1010)은 핵산 서열(1020)이 입력됨에 따라, host의 카테고리별 확률값(1030) 또는 유기체의 카테고리별 확률값을 출력할 수 있다. 이상에서 기술한 실시예들에 따라, 추정 모델(1010)은 host를 추정하도록 학습되는 경우에는 host의 카테고리별 확률값(1030)을 출력하고, 유기체를 추정하도록 학습되는 경우에는 유기체의 카테고리별 확률값을 출력할 수 있다. 여기서, host의 카테고리별 확률값(1030)은 핵산 서열(1020)의 host가 기 정의된 복수개의 생물학적 카테고리(620) 각각에 해당할 것으로 추정되는 확률값을 나타내고, 유기체의 카테고리별 확률값은 핵산 서열(1020)의 유기체가 복수개의 생물학적 카테고리(620) 각각에 해당할 것으로 추정되는 확률값을 나타낸다.

[252] 예컨대, 추정 모델(1010)은 이상에서 fine-tuning에 의해 학습된 바에 따라, 입력되는 핵산 서열(1020)에 포함된 베이스들의 종류 및 순서에 기초하여, 해당 핵산 서열(1020)의 host가 추정 모델(1010)의 복수개의 클래스에 대응되는 복수개의 생물학적 카테고리(620)(예: Homo sapiens, Sus scrofa, Bos taurus 등) 각각에 대한 host의 카테고리별 확률값(1030)(예: 93%, 2%, 0.3% 등)을 출력할 수 있다.

[253] 도 11은 일 실시예에 따라 추정 모델(1010)에서 카테고리별 확률값을 추정하는 예시적인 방식을 도시한다.

[254] 도 11을 더 참조하면, 추정 모델(1010)에는 입력 데이터로서 핵산 서열(1110)이 제공될 수 있다. 도 11에 도시된 핵산 서열(1110)은 핵산 서열(1020)에 대응될 수 있다.

[255] 일 실시예에서, 추정 모델(1010)은 핵산 서열(1110)에 특수 토큰이 삽입된 서열(1120)을 획득할 수 있다. 예컨대, 추정 모델(1010)은 핵산 서열(1110)의 첫 번째 위치에 서열 시작 표시용 특수 토큰으로서 [CLS] 토큰을 삽입하고, 마지막 위치에 서열 구분용 특수 토큰으로서 [SEP] 토큰을 삽입할 수 있다.

[256] 일 실시예에서, 추정 모델(1010)은 특수 토큰이 삽입된 서열(1120)에 대응되는 복수개의 토큰들(1130)을 획득할 수 있다. 예컨대, [CLS] 토큰과 [SEP] 토큰이 삽입된 서열(1020)에 포함된 베이스들 각각에 대해 적어도 하나의 토큰이 생성될 수 있다. 일 예시로, 추정 모델(1010)은 핵산 서열(1110)로서 복수개의 토큰들(1130)을 입력 받을 수 있다. 다른 예시로, 추정 모델(1010)은 핵산 서열(1110)을 입력 받고, 입력된 핵산 서열(1110)에 대한 전처리를 수행하여 복수개의 토큰들(1130)을 생성할 수 있다.

[257] 일 실시예에서, 복수개의 토큰들(1130) 각각은 복수개의 베이스들을 포함하고 그리고 복수개의 토큰들 중 서로 인접한 토큰들에서 핵산 서열(1110)의 적어도 일부가 중첩될 수 있다. 실시예에 따라서, 복수개의 토큰들(1130)은 상술한 k-mer 기법 또는 유전자 추정 기법을 이용하여 획득될 수 있다.

[258] 일 실시예에서, 추정 모델(1010)은 복수개의 토큰들(1130)을 이용하여, 추정 모델(1010)의 하나 이상의 클래스(1140)에 대응되는 하나 이상의 추정값(1150)을 출력할 수 있다. 실시예에 따라서, 각 클래스(1140)는 기 정의된 복수개의 생물학적 카테고리(620) 중 어느 하나이고, 각 클래스(1140)에 대응되는 추정값(1150)은 host 또는 유기체가 해당 생물학적 카테고리에 해당될 확률값일 수 있다. 예를 들어, 컴퓨터 장치(100)는 추정 모델(1010)을 사용하여 복수개의 토큰들(1130)에 대하여 사전결정된 개수의 클래스들(예: 복수개의 생물학적 카테고리)에 대응하는 추정값들(1050)(예: host의 카테고리별 확률값(1030))을 생성할 수 있다. 일 실시예에서, 추정 모델(1010)의 클래스의 종류나 개수는 구현 양태에 따라 가변적으로 결정될 수 있다.

[259] 일 예로, 도 11에 도시된 것처럼, 추정값의 합계가 1인 범위 내에서, Homo sapiens의 제1 클래스에 대응되는 제1 추정값은 95%이고, Bos taurus의 제2 클래스에 대응되는 제2 추정값은 2%이고, Sus scrofa의 제3 클래스에 대응되는 제3 추정값은 0.3%일 수 있다.

[260] 한편, 추정 모델(1010)에 제공되는 핵산 서열(1020)은, 베이스들의 개수가 사전설정된 범위 내에 있을 수 있다. 예컨대, 추정 모델(1010)에 입력되는 핵산 서열(1020)은 베이스들의 개수가 제1 cutoff 이상 또는 제2 cutoff 이하일 수 있다.

[261] 그러나, 핵산 서열(1020)에 포함된 베이스들의 개수가 사전설정된 범위에서 벗어난 경우, 일 실시예에 따른 컴퓨터 장치(100)는 베이스들의 개수가 사전설정된 범위 내에 있도록 핵산 서열(1020)로부터 하나 이상의 부분 서열을 획득하고, 추정 모델(1010)에 하나 이상의 부분 서열을 제공할 수 있다.

[262] 이에 관한 실시예들에 대해 도 8을 다시 참조하여 설명하도록 한다. 일부 실시예들은 fine-tuning의 과정에서 상술한 실시예들을 포함할 수 있으며, 중복되는 설명은 생략하도록 한다. 이하의 실시예들에서 핵산 서열(810)은 host 또는 유기체의 추정 대상인 핵산 서열(1020)에 대응될 수 있다.

[263] 일 실시예에서, 컴퓨터 장치(100)는 베이스들의 개수가 제2 cutoff를 초과하는 경우, 제2 cutoff 개수의 베이스들이 포함되도록 또는 사전설정된 시작 지점을 기준으로 제2 cutoff 개수의 베이스들이 포함되도록 핵산 서열(810)로부터 하나의 제1 부분 서열(821)을 획득할 수 있다. 또한, 컴퓨터 장치(100)는 제1 부분 서열(821)만 추정 모델(1010)에 제공하고, 제1 부분 서열(821)을 제외한 나머지 서열은 추정 모델(1010)에 제공하지 않을 수 있다.

[264] 이처럼 부분 서열(820)이 단수개인 경우, 추정 모델(1010)은 입력되는 제1 부분 서열(821)을 이용해서 제1 부분 서열(821)을 포함하는 유기체의 host(또는 유기체)를 추정할 수 있다. 컴퓨터 장치(100)는 추정 모델(1010)로부터 제1 부분 서열

(821)에 대해 추정된 host(또는 유기체)를 얻을 수 있고, 이렇게 추정된 host를 핵산 서열(820)의 host(또는 유기체)로서 출력할 수 있다.

[265] 다른 실시예에서, 컴퓨터 장치(100)는 베이스들의 개수가 제2 cutoff를 초과하는 경우, 각각 제2 cutoff 개수 이하의 베이스들이 포함되도록 핵산 서열(810)로부터 복수개의 부분 서열(820)을 획득할 수 있다. 또한, 컴퓨터 장치(100)는 복수개의 부분 서열(820) 각각을 추정 모델(1010)에 제공할 수 있다.

[266] 이처럼 부분 서열(820)이 복수개인 경우, 추정 모델(1010)은 복수개의 부분 서열(820) 각각에 대한 host(또는 유기체)를 추정하고, 각각에 대해 추정된 host(또는 유기체)를 출력할 수 있다. 예컨대, 제1 부분 서열(821)을 이용해서 제1 부분 서열(821)의 host의 카테고리별 확률값을 출력하고, 제2 부분 서열(822)을 이용해서 제2 부분 서열(822)의 host의 카테고리별 확률값을 출력할 수 있다.

[267] 또는, 컴퓨터 장치(100)는 복수개의 부분 서열(820) 각각에 대한 host(또는 유기체)를 통계 처리할 수 있고, 통계 처리한 결과를 이용해서 핵산 서열(810)에 대한 host(또는 유기체)를 추정할 수 있다. 예컨대, 컴퓨터 장치(100)는 복수개의 부분 서열(820) 각각에 대해 추정된 host의 카테고리들에 대해서, majority vote 방식에 따라 과반수를 차지하는 host의 카테고리를 선택하거나, 평균 방식에 따라 카테고리별 확률값의 평균이 가장 큰 host의 카테고리를 선택할 수 있으며, 선택된 host의 카테고리를 전체 서열인 핵산 서열(820)에 대해 추정된 host로서 출력할 수 있다.

[268] 일 실시예에서, 컴퓨터 장치(100)는 핵산 서열(810)에 대해 추정된 host(또는 유기체)를 출력할 때, 복수개의 부분 서열(820) 각각에 대한 host(또는 유기체)를 통계 처리한 결과를 함께 출력할 수 있다. 일 예로, majority vote 방식이 이용된 경우, 과반수를 차지하는 host 종의 하이라이트 표시와 함께, 기준값(예: 1) 이상의 vote 수를 갖는 다른 host 종을 함께 출력할 수 있다. 다른 일 예로, 평균 방식이 이용된 경우, 카테고리별 평균값이 가장 큰 host 종의 하이라이트 표시와 함께, 카테고리별 평균값이 기준값 이상인 다른 host 종을 함께 출력할 수 있다.

[269] 일 실시예에서, 컴퓨터 장치(100)는 베이스들의 개수가 제1 cutoff 미만인 경우, 제1 cutoff 이상 및 제2 cutoff 개수 이하의 베이스들을 포함하는 핵산 서열의 입력을 권장하는 메시지를 출력할 수 있다.

[270] 이에 따라, 추정 모델(1010)에 입력되는 핵산 서열의 길이가 적정 범위 내에 있도록 조정될 수 있으며, 이로 인해 여러 추정 결과들이 얻어진 경우 이들에 대한 통계 처리를 통해 추정 정확도가 향상될 수 있다.

[271] 일 실시예에서, 컴퓨터 장치(100)는 추정 모델(1010)에 의해 추정된 결과에 기초해서 추정 모델(1010)에 제공되는 핵산 서열에 포함되는 베이스들의 개수를 갱신할 수 있다. 구체적으로, 컴퓨터 장치(100)는 추정된 host(또는 유기체)의 개수가 사전설정된 제1 개수 이상 또는 제2 개수 이하인 경우, 상기한 제1 cutoff 또는 제2 cutoff를 상이한 값으로 갱신할 수 있다. 예컨대, 컴퓨터 장치(100)는 추정 모델(1010)로부터 출력되는 복수개의 생물학적 카테고리(620) 중에서 각각의 카

테고리별 확률값(1030)이 기 설정된 기준값 이상인 카테고리의 개수가 제1 개수(예: 3개)보다 크거나, 또는 제2 개수(예: 1개)보다 작은 경우, 상기한 제1 cutoff 또는 제2 cutoff를 기 설정된 다른 값으로 갱신하고, 베이스들의 개수가 갱신된 제1 cutoff 이상 또는 제2 cutoff 이하가 되도록 핵산 서열로부터 처리되는 핵산 서열을 추정 모델(1010)에 입력하여서 host를 재추정할 수 있다. 또한, 컴퓨터 장치(100)는 재추정 결과가 소정 조건(예: 전술한 카테고리 개수 조건의 충족, 재추정 횟수 등)을 충족할 때까지 상기한 cutoff 값의 갱신 및 재추정을 반복할 수 있으며, 추정 결과 및/또는 재추정 결과를 제공할 수 있다.

[272] 일 실시예에서, 추정 모델(1010)에 의해 추정된 유기체 또는 host의 개수가 제1 개수 이상인 경우에는 제1 cutoff를 더 큰 값으로 갱신하고/하거나, 제2 개수 이하인 경우에는 제2 cutoff를 더 작은 값으로 갱신할 수 있다. 예컨대, 상술한 것처럼 복수개의 생물학적 카테고리(620) 중에서 확률값이 기준값 이상인 카테고리의 개수가 제1 개수보다 큰 경우, 여러 종이 추정됨에 따라 부정확한 결과가 도출되었으므로, 베이스들의 개수가 더 많아지도록 처리된 핵산 서열을 추정 모델(1010)에 입력하여 보다 타이트한 추정 결과가 얻어지도록 할 수 있다. 또는, 확률값이 기준값 이상인 카테고리의 개수가 제2 개수보다 작은 경우, 적합한 종이 추정되지 않음에 따라 부정확한 결과가 도출되었으므로, 베이스들의 개수가 더 줄어들도록 처리된 핵산 서열을 추정 모델(1010)에 입력하여 보다 느슨한 추정 결과가 얻어지도록 할 수 있다.

[273] 통상적으로, 유기체의 종류별로 유전체의 길이나 해당 유기체를 결정짓는 주요 서열의 길이 등에 상당한 차이가 존재한다. 이로 인해 모델에서 이용되는 적정한 입력 서열의 길이(베이스들의 개수)를 결정하는데 어려움이 있으며, 이로 인해 추정 결과의 정확성이 감소하는 문제점이 있다. 그러나, 전술한 실시예에 따르면, 적정한 추정 결과가 얻어질 때까지 추정 모델(1010)에 입력되는 서열 길이를 유동적으로 조절 가능하며, 상기한 문제점을 해결 가능한 이점이 있다.

[274] 한편, 컴퓨터 장치(100)는 추정 모델(1010)에서 추정한 host 또는 유기체를 다양한 방식으로 출력할 수 있다.

[275] 일 실시예에서, 컴퓨터 장치(100)는 핵산 서열(1020)와 host의 카테고리별 확률값(1030)(또는 유기체의 카테고리별 확률값)이 포함된 결과 화면을 디스플레이할 수 있다. 예컨대, 결과 화면에는 복수개의 생물학적 카테고리(620) 각각의 종명과 각각의 확률값이 표시될 수 있다.

[276] 다른 일 실시예에서, 컴퓨터 장치(100)는 host의 카테고리별 확률값(1030) (또는 유기체의 카테고리별 확률값) 중 소정의 조건을 충족하는 하나 이상의 host(또는 유기체)가 포함된 결과 화면을 디스플레이할 수 있다. 일 예로, 결과 화면에는 host의 카테고리별 확률값(1030)(또는 유기체의 카테고리별 확률값) 중 가장 큰 확률값 또는 기준값 이상의 확률값을 갖거나 값이 큰 순으로 상위 특정 개수의 host 종명(또는 유기체 종명)이 해당 확률값과 함께 표시될 수 있다.

[277] 이러한 결과 화면은 표의 형태로 디스플레이될 수도 있고, 구현 양태에 따라 그래프나 이미지 등의 형태로 디스플레이될 수도 있으며, 이러한 표나 그래프의 종류, 스케일(scale) 등은 상이할 수 있다.

[278] 일 실시예에서, 컴퓨터 장치(100)는 host(또는 유기체)의 추정에서 이용된 추정 근거 데이터를 획득하여 출력할 수 있다. 예컨대, 컴퓨터 장치(100)는 추정된 host를 포함하는 결과 화면과 근거 요청 버튼을 함께 디스플레이하고, 근거 요청 버튼에 대한 사용자의 선택 입력이 수신되는 경우, host의 카테고리별 확률값(1030)에 대한 추정 근거 데이터를 획득하여 디스플레이할 수 있다. 일 예로, 일련의 핵산 서열 중에서 host의 추정 과정에 주요한 특징인 것으로 분석된 서열 조각이 추정 근거 데이터로서 표시될 수 있다. 실시예에 따라, 추정 근거 데이터는 이미지의 형태로 표시될 수도 있으며, 예컨대, 추정된 host가 Homo sapiens인 경우, 해당 핵산 서열이 사람에게서만 유독 높은 확률값이 추정된 주요 근거가 되는 서열 조각이 핵산 서열 이미지 상에서 시각적으로 강조되어 표시될 수 있다.

[279] 일 실시예에서, 추정 근거 데이터는 XAI(explainable artificial intelligence) 방식으로 산출될 수 있다. 예를 들면, 컴퓨터 장치(100)는 설명 가능한 딥러닝 모델로 구현된 추정 모델(1010)로부터 설명 가능한 특징을 추출하는 방식(예: SmoothGrad 등)이 이용하여, 추출된 특징 등에 대한 설명 정보를 기 설정된 설명 인터페이스로 제공할 수 있다. 실시예에 따라, 그 밖에도 모델에 종속적으로 입력 데이터의 특징, 가중치, 주요 객체 위치 등을 확인하는 방식(예: LRP 등)이 이용될 수 있다. 다른 예를 들면, 컴퓨터 장치(100)는 추정 모델(1010)에서 이용되는 어텐션 알고리즘을 이용하여 추정 근거 데이터를 획득할 수 있다. 예컨대, 어텐션 알고리즘에 따라 계산된 유사도, 유사도가 반영된 키와 값, 어텐션값 등을 이용해 추정 근거 데이터를 생성할 수 있다. 또한, 추정 근거 데이터의 획득 과정에서 BERT 내부 구조분석 알고리즘(ACL 2019 참조) 등이 이용될 수도 있다.

[280] 한편, 추정 근거 데이터의 획득 과정에서 실시예에 따라, 모델에 종속되지 않고 입력을 조정하여 나온 출력을 보면서 원인 분석을 하는 모델 비종속적 설명 방식(예: LIME) 등이 이용될 수도 있다.

[281] 도 12는 일 실시예에 따라 유기체 또는 host를 추정하기 위한 예시적인 흐름도를 도시한다. 도 12는 도 1 내지 도 11에 도시된 실시예들을 참조하여 이해될 수 있다.

[282] 일 실시예에서, 도 12에서 도시되는 단계들은 컴퓨터 장치(100)에 의해 수행될 수 있다. 추가적인 일 실시예에서, 도 12에서 도시되는 단계들은 단말에서 수행되는 방식과 같이 하나의 엔티티에 의해 구현될 수 있다. 추가적인 다른 실시예에서, 도 12에서 도시되는 단계들 중 일부는 사용자 단말에서 수행되고 다른 일부는 서버에서 수행되는 방식과 같이 도 12에서의 단계들은 복수개의 엔티티들에 의해 구현될 수도 있다.

[283] 단계 S1210에서 컴퓨터 장치(100)는 사전 학습된 모델(510)을 fine-tuning하여서 획득된 추정 모델(1010)에 접근할 수 있다

[284] 일 실시예에서, 컴퓨터 장치(100)는 서버에 의해 전이학습 방식으로 학습된 추정 모델(1010)을 사용자 단말이 서버로부터 수신하고 실행시키는 방식으로 추정 모델(1010)에 접근하도록 구현될 수 있다. 다른 일 실시예에서, 컴퓨터 장치(100)는 서버에 의해 학습된 추정 모델(1010)이 데이터베이스에 저장되고 사용자 단말이 데이터베이스에 접근하여 추정 모델(1010)을 수신하고 실행시키는 방식으로 추정 모델(1010)에 접근하도록 구현될 수 있다. 다른 일 실시예에서, 컴퓨터 장치(100)는 사용자 단말이 메모리(110) 또는 다른 저장매체에 기 저장된 추정 모델(1010)을 로딩하고 실행시키는 방식으로 추정 모델(1010)에 접근하도록 구현될 수 있다. 다른 일 실시예에서, 컴퓨터 장치(100)는 사용자 단말이 서버에 의해 학습된 추정 모델(1010)의 실행을 위한 요청을 추정 모델(1010)을 실행시키기 위해 요구되는 데이터(예: 입력 데이터 등)와 함께 서버에 전송하여서 서버로부터 추정 모델(1010)의 실행 결과를 수신하는 방식으로 추정 모델(1010)에 접근하도록 구현될 수 있다. 그러나, 본 개시에서 추정 모델(1010)에 접근하는 것은 이에 제한되지 않으며, 다양하게 변형되어 실시될 수 있다.

[285] 단계 S1220에서 컴퓨터 장치(100)는 추정 모델(1010)에 핵산 서열(1020)을 제공할 수 있다. 일 실시예에서, 컴퓨터 장치(100)는 핵산 서열에 대한 사용자 입력에 기초하여 핵산 서열(1020)을 획득하거나, 메모리(110), 다른 장치 또는 저장 매체로부터 핵산 서열(1020)을 수신할 수 있다. 일 실시예에서, 추정 대상인 핵산 서열(1020)은 특정 타겟 분석물의 검출을 위한 진단 시약에 이용되는 서열 후보들 중 하나일 수 있다.

[286] 일 실시예에서, 추정 모델(1010)에 제공되는 핵산 서열(1020)은 베이스들의 개수가 사전설정된 제1 cutoff 이상 또는 제2 cutoff 이하일 수 있다.

[287] 일 실시예에서, 핵산 서열(1020)에 포함된 베이스들의 개수가 제2 cutoff를 초과하는 경우, 컴퓨터 장치(100)는 제2 cutoff 개수 이하의 베이스들이 포함되도록 핵산 서열(1020)로부터 하나 이상의 부분 서열을 획득할 수 있다. 이러한 경우, 단계 S1220에서 추정 모델(1010)에는 상기 하나 이상의 부분 서열이 제공될 수 있다.

[288] 일 실시예에서, 단계 S1220에서 추정 모델(1010)에는 사전설정된 시작 지점을 기준으로 제2 cutoff 개수 이하의 베이스들을 포함하는 하나의 부분 서열이 제공되고 해당 하나의 부분 서열을 제외한 나머지 서열은 제공되지 않을 수 있다. 또는, 일 실시예에서, 단계 S1220에서 추정 모델(1010)에는 제2 cutoff 개수 이하의 베이스들을 포함하는 복수개의 부분 서열 각각이 제공될 수 있다.

[289] 단계 S1230에서 컴퓨터 장치(100)는 추정 모델(1010)로부터, 핵산 서열(1020)을 포함하는 유기체 또는 해당 유기체의 host를 추정할 수 있다.

[290] 일 실시예에서, 상기 추정에서는 유기체 또는 host의 생물학적 카테고리로서, 생물학적 분류 체계를 구성하는 복수개의 hierarchical level 중 어느 하나의 hierarchical level에 위치하는 카테고리가 추정될 수 있다. 일 실시예에서, 상기 어느 하나의 hierarchical level은 생물학적 분류 체계에서의 종 레벨일 수 있다.

[291] 일 실시예에서, 단계 S1220에서 추정 모델(1010)에 하나의 부분 서열이 제공되고 나머지 서열은 제공되지 않은 경우, 단계 S1230에서는 하나의 부분 서열에 대한 host(또는 유기체)가 추정될 수 있고, 핵산 서열(1020)에 대해 추정된 host(또는 유기체)는 상기 하나의 부분 서열 정보에 대한 host일 수 있다.

[292] 일 실시예에서, 단계 S1220에서 추정 모델(1010)에 복수개의 부분 서열 각각이 제공된 경우, 단계 S1230에서는 복수개의 부분 서열 각각에 대한 host(또는 유기체)가 추정될 수 있다. 이러한 경우, 컴퓨터 장치(100)는 복수개의 부분 서열 각각에 대한 host(또는 유기체)를 통계 처리할 수 있으며, 핵산 서열(1020)에 대해 추정된 host(또는 유기체)는 복수개의 부분 서열 각각에 대한 host(또는 유기체)를 통계 처리한 결과를 이용해서 획득된 것일 수 있다.

[293] 상술한 fine-tuning의 실시예들에 따라 컴퓨터 장치(100)는 상이한 추정 결과를 제공할 수 있다.

[294] 제1 실시예에서, 컴퓨터 장치(100)는 핵산 서열(1020)을 추정 모델(1010)에 제공하고, 추정 모델(1010)로부터 출력되는 host의 카테고리별 확률값(1030)를 포함하는 추정 결과를 제공할 수 있다. 예컨대, 추정 결과는 host와 관련된 복수개의 생물학적 카테고리(예: Homo sapiens, Bos taurus, Sus scrofa 등) 각각에 대한 host의 카테고리별 확률값(예: 95%, 3%, 0.1% 등)을 포함할 수 있다.

[295] 제2 실시예에서, 컴퓨터 장치(100)는 핵산 서열(1020)을 추정 모델(1010)에 제공하고, 추정 모델(1010)로부터 출력되는 유기체의 카테고리별 확률값을 포함하는 추정 결과를 제공할 수 있다. 예컨대, 추정 결과는 유기체와 관련된 복수개의 생물학적 카테고리(예: Rotavirus A, Influenza virus A, Rabies lyssavirus 등) 각각에 대한 유기체의 카테고리별 확률값(예: 95%, 3%, 0.1% 등)을 포함할 수 있다.

[296] 제3 실시예에서, 컴퓨터 장치(100)는 핵산 서열(1020)을 추정 모델(1010)에 제공하고, 추정 모델(1010)로부터 출력되는 host의 카테고리별 확률값(1030)과 유기체의 카테고리별 확률값을 포함하는 추정 결과를 제공할 수 있다.

[297] 제4 실시예에서, 컴퓨터 장치(100)는 복수개의 유기체 종류별로 마련된 복수개의 추정 모델(1010)을 획득하고, 기 획득된 서열 관련 정보로부터 핵산 서열(1020)과 유기체 정보를 각각 획득하고, 복수개의 추정 모델(1010) 중 해당 유기체 정보에 대응되는 어느 하나의 추정 모델(1010)에 핵산 서열(1020)을 제공할 수 있다. 또한, 컴퓨터 장치(100)는 추정 모델(1010)로부터 출력되는 host의 카테고리별 확률값(1030)를 포함하는 추정 결과를 제공할 수 있다.

[298] 제5 실시예에서, 컴퓨터 장치(100)는 핵산 서열(1020)을 추정 모델(1010)에 포함된 host 추정 모델과 유기체 추정 모델 각각에 제공하고, host 추정 모델로부터 출력되는 host의 카테고리별 확률값(1030)과 유기체 추정 모델로부터 출력되는 유기체의 카테고리별 확률값을 포함하는 추정 결과를 제공할 수 있다.

[299] 단계 S1230 이후, 일 실시예에 따른 컴퓨터 장치(100)는 핵산 서열(1020)의 유기체에 대한 정보와 추정된 유기체 간의 비교 결과 또는 host에 대한 정보와 추정된 host 간의 비교 결과를 획득할 수 있다.

[300] 구체적으로, 컴퓨터 장치(100)는 단계 S1220 또는 그 이전에 핵산 서열(1020), 해당 유기체 및/또는 해당 host에 대한 정보를 포함하는 서열 관련 정보를 획득하고, 서열 관련 정보로부터 핵산 서열(1020)을 획득하여서 추정 모델(1020)에 제공할 수 있다. 예컨대, 컴퓨터 장치(100)는 서열 관련 정보를 공개 데이터베이스로부터 수집할 수 있다. 또한, 단계 S1230 이후, 컴퓨터 장치(100)는 추정 모델(1020)에 의해 추정된 유기체의 생물학적 카테고리와 서열 관련 정보에 포함된 유기체의 생물학적 카테고리를 비교한 결과를 생성하거나, 추정 모델(1020)에 의해 추정된 host의 생물학적 카테고리와 서열 관련 정보에 포함된 host의 생물학적 카테고리를 비교한 결과를 생성할 수 있다. 이러한 비교 결과에는 예컨대, 생물학적 카테고리의 일치 여부 및 각 생물학적 카테고리에 대해 추정된 확률값 등이 포함될 수 있다.

[301] 일 실시예에서, 컴퓨터 장치(100)는 상기한 비교 결과를 출력하거나, 사용자 요청에 대응하여 분자 진단용 시약의 개발에 이용되는 사용자 단말에 전송할 수 있다.

[302] 전술한 바와 같이, 본 명세서 내에서 기재된 유기체 또는 host를 추정하는 방법을 사용하여, 핵산 서열로부터 해당 핵산 서열을 포함하는 유기체 또는 해당 유기체의 host를 추정할 수 있다. 이러한 실시예는 전이학습 방식으로 추정 모델을 획득하는 기술적 특징과 결합되지 않고, 추정 모델을 사용하여 유기체 또는 host를 추정하는 기술적 특징을 독립적으로 사용할 수 있다.

[303]

[304] 이하에서는, 위에서 제시한 유기체 또는 host의 추정 결과가 분자 진단용 시약 개발에 이용되는 다양한 실시예들에 대해 서술하도록 한다. 실시예에 따라서, 분자 진단용 시약의 개발에는 타겟 분석물의 검출에 이용되는 프라이머 및 프로브 중 적어도 하나가 개발될 수 있으나, 이에 제한되는 것은 아니다.

[305] 일 실시예에서, 핵산 서열(1020)과 추정된 host 및/또는 추정된 유기체는 해당 host를 대상으로 하거나 해당 유기체의 검출을 위한 분자 진단용 시약 개발에 이용될 수 있다. 예컨대, 대상으로 하는 특정 종의 host(예: Homo sapiens)로부터 채취되는 샘플 내에 특정 병원체의 핵산 서열이 존재하는지 여부를 검출하기 위해 이용되는 프라이머 또는 프로브의 서열을 디자인하는 개발 과정이 진행될 수 있으며, 이를 위해 해당 병원체의 서열인 것으로 등재된 다양한 핵산 서열들에 대한 서열 관련 정보들이 수집될 수 있다. 일 예로, 수집된 핵산 서열들 중에서 (i) 추정 모델(1010)이 추정한 host 종과 해당 서열 관련 정보에 포함된 host 종이 일치하는 것으로 검증된 핵산 서열, 및/또는 (ii) 추정 모델(1010)이 추정한 병원체 종과 해당 서열 관련 정보에 포함된 병원체 종이 일치하는 것으로 검증된 핵산 서열은 해당 개발 과정에서 해당 병원체의 핵산 서열에 특이적인 프라이머 또는 프로브의 디자인을 위한 후보 서열들 중 하나로서 이용될 수 있다.

[306] 다른 일 실시예에서, 서열 관련 정보에 포함된 유기체에 대한 정보와 추정된 유기체 간의 비교 결과 및/또는 서열 관련 정보에 포함된 host에 대한 정보와 추정

된 host 간의 비교 결과는 해당 host를 대상으로 하고/하거나 해당 유기체의 검출을 위한 분자 진단용 시약 개발에 이용될 수 있다. 예컨대, 상술한 예시와 같은 개발 과정이 진행될 때, 해당 개발 과정에서 검출하고자 하는 유기체에 대한 정보, 해당 개발 과정에서 대상으로 하는 host에 대한 정보 및 이에 대응되는 핵산 서열이 포함되어 있는 서열 관련 정보가 수집될 수 있다. 수집된 서열 관련 정보에 포함된 핵산 서열들 중에서 추정 모델(1010)이 추정한 host 종과 해당 서열 관련 정보에 포함된 host 종이 동일하거나, 추정 모델(1010)이 추정한 병원체 종과 해당 서열 관련 정보에 포함된 병원체 종이 동일한 것으로 검증된 핵산 서열들만 해당 개발 과정에 이용되고, 상이한 것으로 확인된 핵산 서열들은 해당 개발 과정에서 배제될 수 있다.

[307] 이에 관한 실시예들에 대해서는 도 13 내지 도 14를 참조하여 보다 상세하게 서술하도록 한다.

[308] 도 13은 일 실시예에 따른 컴퓨터 장치(100)가 추정 모델(1010)에 의한 추정 결과가 분자 진단용 시약 개발에 이용되도록 하는 과정을 예시하는 순서도이다. 또한, 도 14는 일 실시예에 따른 컴퓨터 장치(100)가 추정 모델(1010)에 의한 추정 결과가 분자 진단용 시약 개발에 이용되도록 제어하는 예시적인 방식을 도시한다.

[309] 도 13 내지 도 14를 참조하면, 데이터베이스(200)는 컴퓨터 장치(100) 및 하나 이상의 정보 제공자 단말(300)과 네트워크를 통해 서로 연결될 수 있다.

[310] 데이터베이스(200)에는 복수개의 핵산 서열이 저장될 수 있다. 실시예에 따라서, 데이터베이스(200)는 복수개의 단말에 의해 액세스될 수 있는 데이터베이스로 구현될 수 있고, 예컨대, NCBI 등과 같은 public 데이터베이스이거나, 또는 컴퓨터 장치(100)의 운영자에 의해 제공되는 private 데이터베이스일 수 있다.

[311] 정보 제공자 단말(300)은 하나 이상의 핵산 서열을 제공하는 정보 제공자의 단말로서, 서버 또는 다른 컴퓨팅 장치와 상호작용 가능한 임의의 형태의 단말을 포함할 수 있다. 여기서, 정보 제공자는 핵산 서열을 제공할 수 있는 사용자로서, 예컨대, 기관이나 기업, 연구소 등이거나 이에 소속된 개인일 수 있으며, 이에 제한되지 않는다. 일 예로, 정보 제공자는 host로부터 채취된 샘플에 대한 분석을 통해 해당 샘플에 포함된 특정 병원체의 핵산 서열에 대한 정보를 얻거나 또는 기타의 방식으로 특정 병원체의 핵산 서열과 해당 host에 대한 정보를 얻을 수 있다.

[312] 단계 S1310에서 정보 제공자 단말(300)은 핵산 서열과 해당 핵산 서열을 포함하는 유기체 및/또는 해당 유기체의 host에 대한 정보를 데이터베이스(200)에 제공할 수 있다. 일 예로, 정보 제공자 단말(300)은 host로부터 채취된 샘플에 대한 분석을 통해 해당 샘플에 포함된 특정 병원체의 핵산 서열, 해당 병원체 종 및 해당 host 종에 대한 정보를 획득하고, 획득한 정보를 데이터베이스(200)에 등재할 수 있다. 다른 일 예로, 정보 제공자 단말(300)은 특정 병원체의 핵산 서열, 해당 병원체 종 및 해당 병원체의 host로서 알려져 있는 종들에 대한 정보를 데이터베이스(200)에 등재할 수 있다.

[313] 단계 S1320에서 데이터베이스(200)는 제공된 핵산 서열과 해당 핵산 서열을 포함하는 유기체 및/또는 해당 유기체의 host에 대한 정보를 저장할 수 있다. 일 실시예에서, 데이터베이스(200)에는 정보 제공자 단말(300)로부터 수신되는 핵산 서열이 해당 유기체의 종류별로 또는 해당 host의 종류별로 분류되어 저장될 수 있다.

[314] 실시예에 따라, 다수의 정보 제공자 단말(300)에 의해 단계 S1310 내지 S1320의 과정이 반복적으로 수행될 수 있으며, 이에 따라 데이터베이스(200)에는 다수의 핵산 서열과 해당 유기체 및/또는 host에 대한 정보들이 수집될 수 있다.

[315] 단계 S1330에서 컴퓨터 장치(100)는 데이터베이스(200)로부터 핵산 서열과 해당 유기체 및/또는 해당 host에 대한 정보를 포함하는 서열 관련 정보를 획득할 수 있다. 일 실시예에서, 컴퓨터 장치(100)는 데이터베이스(200)에 액세스하고, 데이터베이스(200)에서 제공하는 핵산 서열에 관한 다양한 정보들 중에서 시약 개발에서 대상으로 하는 host 및/또는 유기체에 대응되는 핵산 서열에 대한 정보들을 검색하여서, 이에 대응되는 검색 결과로서 해당 host와 해당 유기체에 대한 정보 및 해당 핵산 서열을 각각 포함하는 복수개의 서열 관련 정보를 획득할 수 있다. 또한, 컴퓨터 장치(100)는 획득된 복수개의 서열 관련 정보에 기초해서, 대상으로 하는 특정 병원체 및 특정 host에 대응되는 복수개의 핵산 서열들이 포함된 서열 관련 정보들을 포함하는 후보 핵산 서열 리스트(1410)를 획득할 수 있다. 예컨대, 분자 진단용 시약 개발이 Rotavirus에 속하는 유기체의 검출을 타겟으로 하고 Homo sapiens에 속하는 host를 대상으로 개발되는 경우, 후보 핵산 서열 리스트(1410)는 복수개의 서열 관련 정보 중에서 유기체에 대한 정보는 Rotavirus에 대응되고 host에 대한 정보는 Homo sapiens에 대응되는 서열 관련 정보들에 있는 핵산 서열들(예: Sequence 1 ~ N)을 포함할 수 있다.

[316] 단계 S1340에서 컴퓨터 장치(100)는 획득된 서열 관련 정보로부터 핵산 서열을 획득하고, 추정 모델(1010)을 이용해서 해당 핵산 서열을 포함하는 유기체 또는 해당 유기체의 host를 추정할 수 있다. 일 실시예에서, 컴퓨터 장치(100)는 후보 핵산 서열 리스트(1410)에 포함된 복수개의 핵산 서열 각각을 추정 모델(1010)에 입력하고, 추정 모델(1010)로부터 복수개의 핵산 서열 각각에 대한 host의 추정 결과(1420)(또는 유기체의 추정 결과)를 획득할 수 있다.

[317] 단계 S1350에서 컴퓨터 장치(100)는 서열 관련 정보에 포함된 유기체에 대한 정보와 추정된 유기체 간의 비교 결과, 또는, 서열 관련 정보에 포함된 host에 대한 정보와 추정된 host 간의 비교 결과를 획득할 수 있다. 상술한 것처럼, 비교 결과에는 해당 서열 관련 정보에 포함된 host(또는 유기체)와 추정된 host(또는 유기체)가 상호 대응되는지 여부를 포함할 수 있다. 일 실시예에서, 비교 결과는 추정된 host(또는 유기체)의 카테고리별 확률값 중에서 서열 관련 정보에 포함된 host(또는 유기체)에 대해 추정된 확률값이 가장 큰지, 확률값이 큰 순서로 상위 사전설정된 개수 이내에 있는지, 및 기준값 이상인지 여부 중 적어도 하나가 포함될 수 있다. 일 실시예에서, 컴퓨터 장치(100)는 상기한 비교 결과를 출력할 수

있다. 다른 일 실시예에서, 이러한 비교 결과는 해당 host를 대상으로 하는(또는 해당 유기체의 검출을 위한) 분자 진단용 시약의 개발 과정에 제공될 수 있다.

[318] 일 실시예에서, 컴퓨터 장치(100)는 서열 관련 정보에 포함된 host(또는 유기체)와 추정된 host(또는 유기체)가 상이한 핵산 서열은 해당 host를 대상으로 하는(또는 해당 유기체의 검출을 위한) 분자 진단용 시약 개발에 이용되지 않도록 제어할 수 있다. 예컨대, 컴퓨터 장치(100)는 서열 관련 정보에 포함된 host의 생물학적 카테고리(예: Homo sapiens)와 추정된 host의 생물학적 카테고리(예: Sus scrofa)가 상이한 추정 결과(1421)를 검출하고, 후보 핵산 서열 리스트(1410) 중에서 해당 상이한 추정 결과(1421)에 대응되는 핵산 서열(예: sequence 2)가 제외된 필터링된 후보 핵산 서열 리스트(1430)를 획득할 수 있다. 실시예에 따라서, 필터링된 후보 핵산 서열 리스트(1430)는 sequence alignment 기반으로, 타겟 분석물의 종에서는 높은 빈도로 발견되고 그 외의 종에서는 거의 발견되지 않아 타겟 분석물의 종에 특이적인 서열(예: gene region, intergenic region)을 탐색하는 과정에 이용되거나, 이를 기반으로 프라이머 및/또는 프로브를 위한 디자인 가능 영역(designable region)을 결정하는 과정 등에 이용될 수 있다.

[319] 일 실시예에서, 컴퓨터 장치(100)는 서열 관련 정보에 포함된 host(또는 유기체)와 추정된 host(또는 유기체)가 동일한 핵산 서열이 해당 host를 대상으로 하는(또는 해당 유기체의 검출을 위한) 분자 진단용 시약 개발에 이용되도록 제어할 수 있다. 예컨대, 컴퓨터 장치(100)는 필터링된 후보 핵산 서열 리스트(1430)를 분자 진단용 시약 개발과 연관된 사용자 단말에 제공함으로써, 검증된 핵산 서열들이 개발 과정에서 이용되도록 지원할 수 있다.

[320] 전술한 실시예들에 따라, 올리고뉴클레오타이드의 디자인 과정에서 대상으로 하는 host와 추정된 host가 상이한 핵산 서열 또는 검출하고자 하는 유기체와 추정된 유기체가 상이한 핵산 서열은 진단 시약 개발에서 배제되므로, 이를 기반으로 디자인된 올리고뉴클레오타이드는 보다 강건한 성능을 보유할 수 있게 된다. 일례로, 상기 디자인된 올리고뉴클레오타이드를 이용하여 타겟 분석물을 검출하는 경우, host로부터 채취된 샘플 내에서 타겟 분석물에 보다 특이적으로 결합하거나 또는 병원체를 보다 특이적으로 검출 가능한 올리고뉴클레오타이드가 설계될 수 있고, 위양성 가능성이 줄어들 수 있다.

[321] 단계 S1360에서 컴퓨터 장치(100)는 서열 관련 정보에 포함된 유기체 또는 host에 대한 정보가 추정된 유기체 또는 host와 상이하면, 해당 유기체 또는 host에 대한 정보가 추정된 유기체 또는 host로 수정되도록 제어할 수 있다. 일 실시예에서, 컴퓨터 장치(100)는 서열 관련 정보에 포함된 host(또는 유기체)에 대한 정보가 추정된 host(또는 유기체)와 상이한 경우, 서열 관련 정보에 포함된 host(또는 유기체)에 대한 정보를 추정된 host(또는 유기체)로 갱신할 수 있다. 다른 일 실시예에서, 컴퓨터 장치(100)는 상기 비교 결과를 데이터베이스(200)에 송신함으로써, 해당 핵산 서열의 host인 것으로 등재되어 있는 host(또는 유기체)가 추정된

host(또는 유기체)로 수정되도록 데이터베이스(200)에 요청하거나 권장할 수 있다.

[322] 실시예에 따라, 데이터베이스(200)에서의 정보 오류가 정정되는 경우, 공개 데이터베이스에 등재된 핵산 서열을 다양한 목적으로 활용하는 사용자들의 잠재적인 이익 향상을 도모할 수 있다.

[323] 이상에서 기술한 바와 같이, 유기체 또는 host의 추정 결과는 해당 host를 대상으로 하거나 해당 유기체의 검출을 위한 분자 진단용 시약 개발에 이용될 수 있으나, 상술한 실시예들에 제한되지 않는다. 실시예에 따라서, 알려진 host 종들 이외에도 변이(variation) 등으로 인해 추후 host로서 나타날 수 있는 잠재적 host 후보군들을 예측하는데 이용될 수 있다. 또는, 실시예에 따라서, 상이한 유기체 종의 핵산 서열들에 대해서 host들 간의 유사성을 예측하고 이를 기반으로 동물원성(原性) 감염증(zoonosis) 등을 연구하거나 분석하는데 이용될 수도 있다.

[324] 본 개시의 기술 분야에서 통상의 지식을 가진 자는 여기에 개시된 실시예들과 관련하여 설명된 다양한 예시적인 논리 블록들, 모듈들, 프로세서들, 수단들, 회로들 및 알고리즘 단계들이 전자 하드웨어, (편의를 위해, 여기서 소프트웨어로 지칭되는) 다양한 형태들의 프로그램 또는 설계 코드 또는 이들 모두의 결합에 의해 구현될 수 있다는 것을 이해할 것이다. 하드웨어 및 소프트웨어의 이러한 상호 호환성을 명확하게 설명하기 위해, 다양한 예시적인 컴포넌트들, 블록들, 모듈들, 회로들 및 단계들이 이들의 기능과 관련하여 위에서 일반적으로 설명되었다. 이러한 기능이 하드웨어 또는 소프트웨어로서 구현되는지 여부는 특정한 애플리케이션 및 전체 시스템에 대하여 부과되는 설계 제약들에 따라 좌우된다. 본 개시의 기술 분야에서 통상의 지식을 가진 자는 각각의 특정한 애플리케이션에 대하여 다양한 방식들로 설명된 기능을 구현할 수 있으나, 이러한 구현 결정들은 본 개시의 범위를 벗어나는 것으로 해석되어서는 안 될 것이다.

[325] 여기서 제시된 다양한 실시예들은 방법, 장치, 또는 표준 프로그래밍 및/또는 엔지니어링 기술을 사용한 제조 물품(article)으로 구현될 수 있다. 용어 제조 물품은 임의의 컴퓨터-판독가능 저장장치로부터 액세스 가능한 컴퓨터 프로그램, 캐리어, 또는 매체(media)를 포함한다. 예를 들어, 컴퓨터-판독가능 저장매체는 자기 저장 장치(예를 들면, 하드 디스크, 플로피 디스크, 자기 스트립, 등), 광학 디스크(예를 들면, CD, DVD, 등), 스마트 카드, 및 플래쉬 메모리 장치(예를 들면, EEPROM, 카드, 스틱, 키 드라이브, 등)를 포함하지만, 이들로 제한되는 것은 아니다. 또한, 여기서 제시되는 다양한 저장 매체는 정보를 저장하기 위한 하나 이상의 장치 및/또는 다른 기계-판독가능한 매체를 포함한다.

[326] 제시된 프로세스들에 있는 단계들의 특정한 순서 또는 계층 구조는 예시적인 접근들의 일례임을 이해하도록 한다. 설계 우선순위들에 기반하여, 본 개시의 범위 내에서 프로세스들에 있는 단계들의 특정한 순서 또는 계층 구조가 재배열될 수 있다는 것을 이해하도록 한다. 첨부된 방법 청구항들은 샘플 순서로 다양한

단계들의 엘리먼트들을 제공하지만 제시된 특정한 순서 또는 계층 구조에 한정되는 것을 의미하지는 않는다.

청구범위

[청구항 1] 메모리, 프로세서 및 상기 메모리에 저장되고 상기 프로세서에 의해 실행되도록 구성된 하나 이상의 프로그램을 사용하는 컴퓨터 장치에 의해 수행되는 컴퓨터 구현 방법에 있어서,
사전 학습된 모델을 fine-tuning하여서 획득된 추정 모델에 접근하는 단계;
상기 추정 모델에 핵산 서열을 제공하는 단계; 및
상기 추정 모델로부터, 상기 핵산 서열을 포함하는 유기체 또는 상기 유기체의 host를 추정하는 단계를 포함하는,
컴퓨터 구현 방법.

[청구항 2] 제 1 항에 있어서,
상기 추정에서는,
상기 유기체 또는 상기 host의 생물학적 카테고리로서, 생물학적 분류 체계를 구성하는 복수개의 hierarchical level 중 어느 하나의 hierarchical level에 위치하는 카테고리가 추정되는,
컴퓨터 구현 방법.

[청구항 3] 제 2 항에 있어서,
상기 어느 하나의 hierarchical level은,
상기 생물학적 분류 체계에서의 종(species) 레벨인,
컴퓨터 구현 방법.

[청구항 4] 제 1 항에 있어서,
상기 사전 학습된 모델은,
학습용 데이터로서 복수개의 핵산 서열이 이용된 것을 특징으로 하는,
컴퓨터 구현 방법.

[청구항 5] 제 4 항에 있어서,
상기 사전 학습된 모델은
상기 복수개의 핵산 서열 각각에 포함된 베이스들 중 일부의 베이스에 마스크(mask)를 적용한 뒤, 마스킹된(masked) 베이스를 맞추는 semi-supervised learning 방식에 의해 학습된 것인,
컴퓨터 구현 방법.

[청구항 6] 제 4 항에 있어서,
상기 사전 학습된 모델은,
각각 두 개 이상의 베이스들을 갖는 토큰들로 토큰화되는 핵산 서열을 이용해서 학습된 것인,
컴퓨터 구현 방법.

[청구항 7] 제 6 항에 있어서,

상기 토큰들은 (i) 상기 핵산 서열을 k 개(상기 k는 자연수)씩 분할하거나 (ii) 상기 핵산 서열을 기능 단위로 분할하여서 토큰화되는 베이스들을 각각 포함하는,
컴퓨터 구현 방법.

[청구항 8] 제 1 항에 있어서,
상기 fine-tuning은 복수의 학습용 데이터 세트를 이용해서 수행되되, 각 학습용 데이터 세트는 (i) 핵산 서열을 포함하는 학습용 입력 데이터 및 (ii) 해당 핵산 서열을 포함하는 유기체 또는 해당 유기체의 host에 대한 라벨 데이터를 포함하는 학습용 정답 데이터를 포함하는,
컴퓨터 구현 방법.

[청구항 9] 제 8 항에 있어서,
상기 fine-tuning은,
(i) 상기 학습용 입력 데이터에 포함된 핵산 서열을 토큰화(tokenization)하여 복수의 토큰들을 획득하는 과정, (ii) 상기 복수의 토큰들로부터 생성되는 컨텍스트 벡터를 이용해서 상기 학습용 입력 데이터에 포함된 핵산 서열의 유기체 또는 host에 대해 추정하는 과정 및 (iii) 상기 추정된 결과와 상기 학습용 정답 데이터 간의 차이가 줄어들도록 상기 사전 학습된 모델을 훈련시키는 과정을 포함하는,
컴퓨터 구현 방법.

[청구항 10] 제 1 항에 있어서,
상기 추정 모델에 제공되는 상기 핵산 서열은,
베이스들의 개수가 사전설정된 제1 cutoff 이상 또는 제2 cutoff 이하인,
컴퓨터 구현 방법.

[청구항 11] 제 1 항에 있어서,
상기 방법은,
상기 핵산 서열에 포함된 베이스들의 개수가 소정의 제2 cutoff를 초과하는 경우, 상기 제2 cutoff 개수 이하의 베이스들이 포함되도록 상기 핵산 서열로부터 하나 이상의 부분 서열을 획득하는 단계를 더 포함하며,
상기 핵산 서열을 제공하는 단계에서 상기 추정 모델에는, 상기 하나 이상의 부분 서열이 제공되는,
컴퓨터 구현 방법.

[청구항 12] 제 11 항에 있어서,
상기 핵산 서열이 제공되는 단계에서 상기 추정 모델에는,
사전설정된 시작 지점을 기준으로 상기 제2 cutoff 개수 이하의 베이스들을 포함하는 하나의 부분 서열이 제공되고 상기 하나의 부분 서열을 제외한 나머지 서열은 제공되지 않거나, 또는
상기 제2 cutoff 개수 이하의 베이스들을 포함하는 복수개의 부분 서열 각각이 제공되는,

컴퓨터 구현 방법.

[청구항 13] 제 11 항에 있어서,
상기 추정하는 단계에서는,
상기 부분 서열이 복수개인 경우, 상기 복수개의 부분 서열 각각에 대한 유기체 또는 host가 추정되고,
상기 방법은,
상기 복수개의 부분 서열 각각에 대한 유기체 또는 host를 통계 처리하는 단계를 더 포함하며,
상기 추정된 유기체 또는 상기 추정된 host는,
상기 통계 처리한 결과를 이용해서 획득된 것인,
컴퓨터 구현 방법.

[청구항 14] 제 13 항에 있어서,
상기 통계 처리에는,
Majority vote 방식, 평균 방식 및 표준 편차 방식 중 적어도 하나가 포함되는,
컴퓨터 구현 방법.

[청구항 15] 제 1 항에 있어서,
상기 host를 대상으로 하는 분자 진단용 시약 개발에는,
(i) 상기 추정된 유기체 또는 상기 추정된 host 및 (ii) 상기 핵산 서열이 이용되는,
컴퓨터 구현 방법.

[청구항 16] 제 15 항에 있어서,
상기 분자 진단용 시약의 개발에서는,
상기 유기체의 검출에 이용되는 프라이머 및 프로브 중 적어도 하나가 개발되는,
컴퓨터 구현 방법.

[청구항 17] 제 1 항에 있어서,
상기 핵산 서열을 제공하는 단계에서는,
상기 핵산 서열과 상기 유기체 또는 상기 host에 대한 정보를 포함하는 서열 관련 정보로부터 상기 핵산 서열을 획득하고,
상기 유기체에 대한 정보와 상기 추정된 유기체 간의 비교 결과 또는 상기 host에 대한 정보와 상기 추정된 host 간의 비교 결과를 획득하는 단계를 더 포함하는,
컴퓨터 구현 방법.

[청구항 18] 제 1 항에 있어서,
상기 핵산 서열을 제공하는 단계에서는,
상기 핵산 서열과 상기 유기체 또는 상기 host에 대한 정보를 포함하는 서열 관련 정보로부터 상기 핵산 서열을 획득하고,

상기 유기체 또는 상기 host에 대한 정보가 상기 추정된 유기체 또는 host와 상이하면, 상기 유기체 또는 상기 host에 대한 정보가 상기 추정된 유기체 또는 host로 수정되도록 제어하는 단계를 더 포함하는,
컴퓨터 구현 방법.

[청구항 19] 제 1 항에 있어서,
상기 추정 모델에 제공되는 상기 핵산 서열은,
베이스들의 개수가 사전설정된 제1 cutoff 이상 또는 제2 cutoff 이하이고,
상기 방법은,
상기 추정된 유기체 또는 host의 개수가 사전설정된 제1 개수 이상 또는 제2 개수 이하인 경우, 상기 제1 cutoff 또는 상기 제2 cutoff를 상이한 값으로 갱신하는 단계를 더 포함하는,
컴퓨터 구현 방법.

[청구항 20] 제 19 항에 있어서,
상기 갱신하는 단계에서는,
상기 추정된 유기체 또는 host의 개수가 상기 제1 개수 이상인 경우에는 상기 제1 cutoff를 더 큰 값으로 갱신하고, 상기 제2 개수 이하인 경우에는 상기 제2 cutoff를 더 작은 값으로 갱신하는 것을 특징으로 하는,
컴퓨터 구현 방법.

[청구항 21] 제 1 항에 있어서,
상기 추정된 유기체 또는 host의 개수가 기 설정된 기준 개수 이상인 경우, 상기 핵산 서열이 상기 유기체 또는 상기 host를 대상으로 하는 분자 진단용 시약 개발에 이용되지 않도록 제어하는 단계를 더 포함하는 것을 특징으로 하는,
컴퓨터 구현 방법.

[청구항 22] 제 1 항의 방법에 포함된 각 단계를 수행하도록 프로그램된,
컴퓨터 판독 가능한 기록 매체에 저장되어 있는 컴퓨터 프로그램.

[청구항 23] 제 1 항의 방법에 포함된 각 단계를 수행하도록 프로그램된 컴퓨터 프로그램이 저장되어 있는,
컴퓨터 판독 가능한 기록 매체.

[청구항 24] 적어도 하나의 명령어를 저장하는 메모리; 및
프로세서를 포함하며,
상기 프로세서는 상기 적어도 하나의 명령어를 실행시킴으로써,
사전 학습된 모델을 fine-tuning하여서 획득된 추정 모델에 접근하고,
상기 추정 모델에 핵산 서열을 제공하고,
상기 추정 모델로부터, 상기 핵산 서열을 포함하는 유기체 또는 상기 유기체의 host를 추정하는,
컴퓨터 장치.

[청구항 25] 메모리, 프로세서 및 상기 메모리에 저장되고 상기 프로세서에 의해 실행되도록 구성된 하나 이상의 프로그램을 사용하는 컴퓨터 장치에 의해 수행되는 컴퓨터 구현 방법에 있어서,
사전 학습된 모델을 fine-tuning하여서 획득된 추정 모델에 접근하는 단계;
상기 추정 모델에 핵산 서열을 제공하는 단계; 및
상기 추정 모델로부터, 상기 핵산 서열을 포함하는 유기체 또는 상기 유기체의 host를 추정하는 단계를 포함하고,
상기 사전 학습된 모델은 학습용 데이터로서 복수개의 핵산 서열을 이용하고,
상기 fine-tuning은 복수의 학습용 데이터 세트를 이용해서 수행되되, 각 학습용 데이터 세트는 (i) 핵산 서열을 포함하는 학습용 입력 데이터 및 (ii) 해당 핵산 서열의 유기체 또는 host에 대한 라벨 데이터를 포함하는 학습용 정답 데이터를 포함하는,
컴퓨터 구현 방법.

[청구항 26] 제 25 항에 있어서,
상기 fine-tuning은,
(i) 상기 학습용 입력 데이터에 포함된 핵산 서열을 토큰화(tokenization)하여 복수의 토큰들을 획득하는 과정, (ii) 상기 복수의 토큰들로부터 생성되는 컨텍스트 벡터를 이용해서 상기 학습용 입력 데이터에 포함된 핵산 서열의 유기체 또는 host에 대해 추정하는 과정 및 (iii) 상기 추정된 결과와 상기 학습용 정답 데이터 간의 차이가 줄어들도록 상기 사전 학습된 모델을 훈련시키는 과정을 포함하는,
컴퓨터 구현 방법.

[청구항 27] 메모리, 프로세서 및 상기 메모리에 저장되고 상기 프로세서에 의해 실행되도록 구성된 하나 이상의 프로그램을 사용하는 컴퓨터 장치에 의해 수행되는 컴퓨터 구현 방법에 있어서,
사전 학습된 모델을 획득하는 단계; 및
상기 사전 학습된 모델을 fine-tuning하여서, 핵산 서열이 제공되면 상기 핵산 서열을 포함하는 유기체 또는 상기 유기체의 host를 추정하도록 학습된 추정 모델을 획득하는 단계를 포함하고,
상기 fine-tuning은 복수의 학습용 데이터 세트를 이용해서 수행되되, 각 학습용 데이터 세트는 (i) 핵산 서열을 포함하는 학습용 입력 데이터 및 (ii) 해당 핵산 서열의 유기체 또는 host에 대한 라벨 데이터를 포함하는 학습용 정답 데이터를 포함하는,
컴퓨터 구현 방법.

[청구항 28] 제 27 항에 있어서,
상기 라벨 데이터는,

상기 유기체 또는 상기 host의 생물학적 카테고리로서, 생물학적 분류 체계를 구성하는 복수개의 hierarchical level 중 어느 하나의 hierarchical level에 위치하는 카테고리에 대한 라벨 데이터인,
컴퓨터 구현 방법.

[청구항 29] 제 28 항에 있어서,
상기 사전 학습된 모델은,
학습용 데이터로서 복수개의 핵산 서열이 이용된 것을 특징으로 하는,
컴퓨터 구현 방법.

[청구항 30] 제 29 항에 있어서,
상기 사전 학습된 모델은
상기 복수개의 핵산 서열 각각에 포함된 베이스들 중 일부의 베이스에 마스크(mask)를 적용한 뒤, 마스킹된(masked) 베이스를 맞추는 semi-supervised learning 방식에 의해 학습된 것인,
컴퓨터 구현 방법.

[청구항 31] 제 29 항에 있어서,
상기 사전 학습된 모델은,
각각 두 개 이상의 베이스들을 갖는 토큰들로 토큰화되는 핵산 서열을 이용해서 학습된 것인,
컴퓨터 구현 방법.

[청구항 32] 제 31 항에 있어서,
상기 토큰들은 (i) 상기 핵산 서열을 k 개(상기 k는 자연수)씩 분할하거나 (ii) 상기 핵산 서열을 기능 단위로 분할하여서 토큰화되는 베이스들을 각각 포함하는,
컴퓨터 구현 방법.

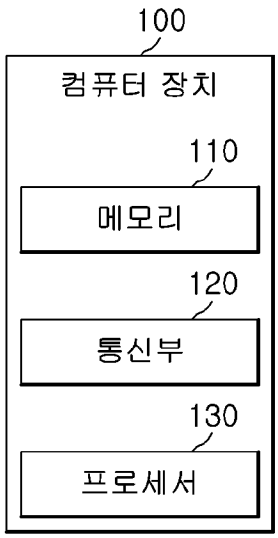
[청구항 33] 제 27 항에 있어서,
상기 fine-tuning은,
(i) 상기 학습용 입력 데이터에 포함된 핵산 서열을 토큰화(tokenization)하여 복수의 토큰들을 획득하는 과정, (ii) 상기 복수의 토큰들로부터 생성되는 컨텍스트 벡터를 이용해서 상기 학습용 입력 데이터에 포함된 핵산 서열의 유기체 또는 host에 대해 추정하는 과정 및 (iii) 상기 추정된 결과와 상기 학습용 정답 데이터 간의 차이가 줄어들도록 상기 사전 학습된 모델을 훈련시키는 과정을 포함하는,
컴퓨터 구현 방법.

[청구항 34] 제 27 항에 있어서,
상기 추정 모델을 획득하는 단계에서, 상기 추정 모델은,
상기 핵산 서열에 포함된 베이스들의 개수가 소정의 제2 cutoff를 초과하는 경우, 상기 제2 cutoff 개수 이하의 베이스들이 포함되도록 상기 핵산 서

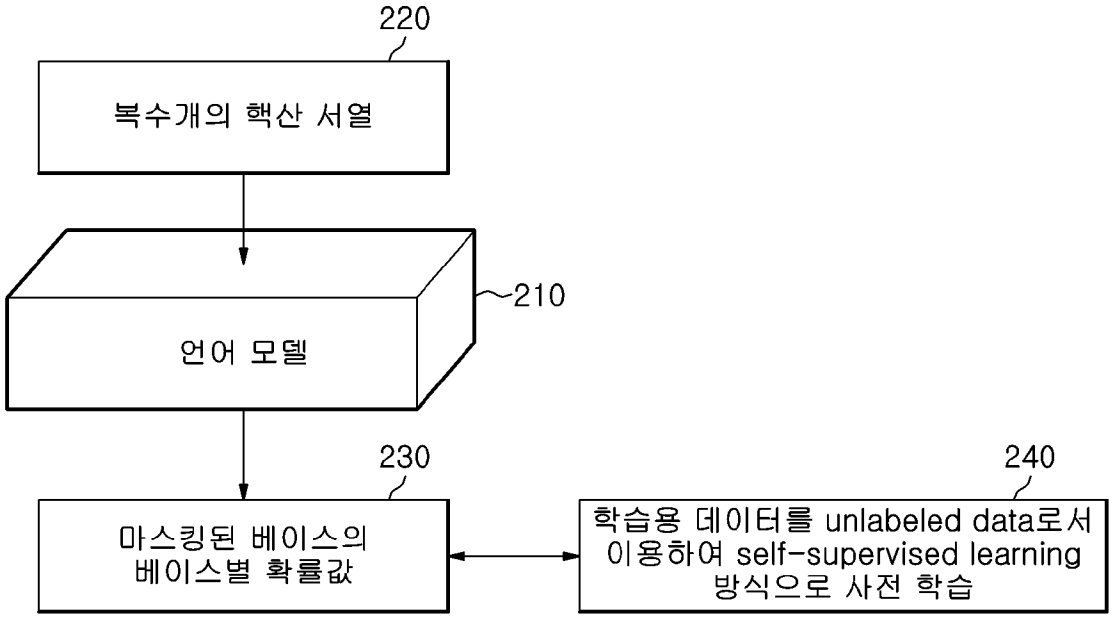
열로부터 획득되는 하나 이상의 부분 서열이 제공되면 상기 하나 이상의 부분 서열의 유기체 또는 host를 추정하도록 학습되는,
컴퓨터 구현 방법.

[청구항 35] 제 34 항에 있어서,
상기 추정 모델을 획득하는 단계에서, 상기 추정 모델은,
상기 부분 서열이 복수개인 경우, 상기 복수개의 부분 서열 각각의 유기체 또는 host를 추정하고, 상기 복수개의 부분 서열 각각의 유기체 또는 host를 통계 처리하고, 상기 통계 처리한 결과를 이용해서 상기 핵산 서열의 유기체 또는 host를 추정하도록 학습되는,
컴퓨터 구현 방법.

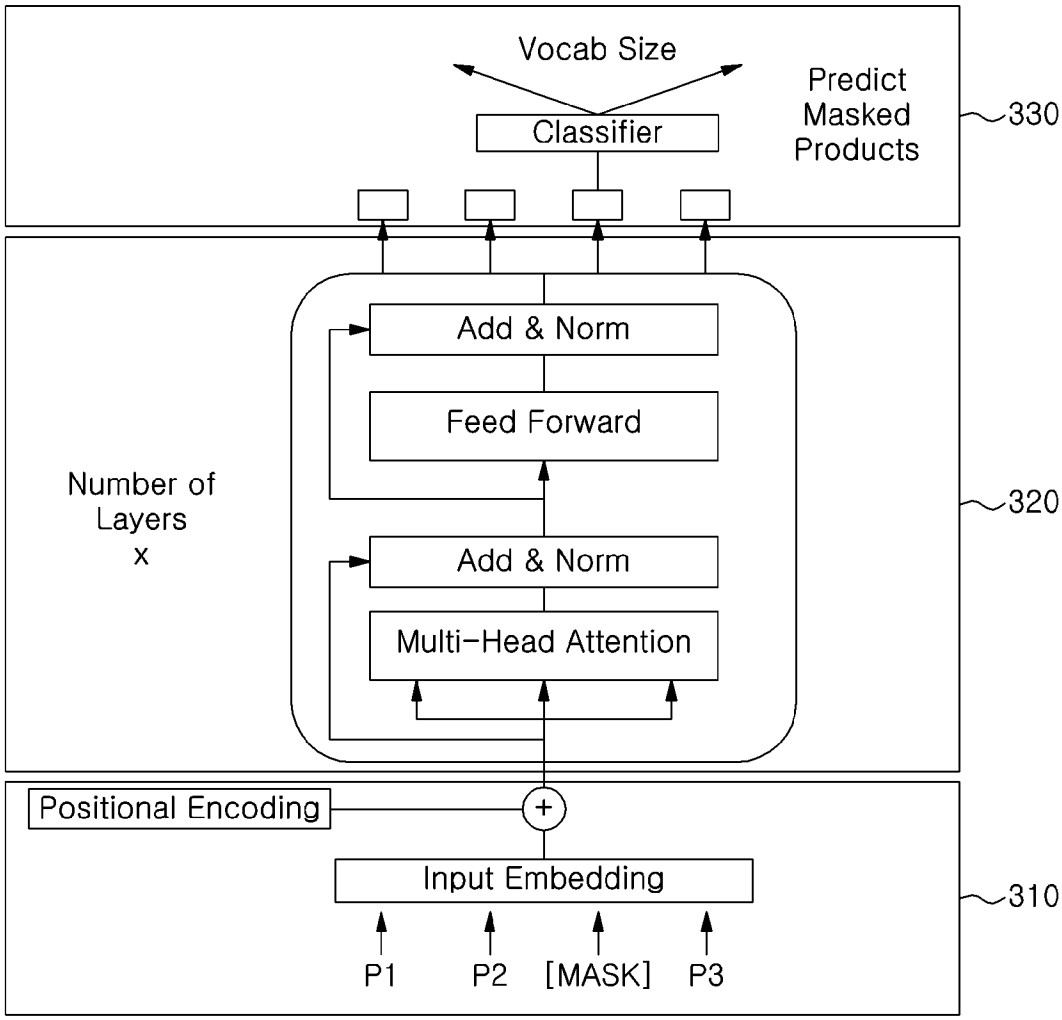
[도1]


100
컴퓨터 장치
110
메모리
120
통신부
130
프로세서

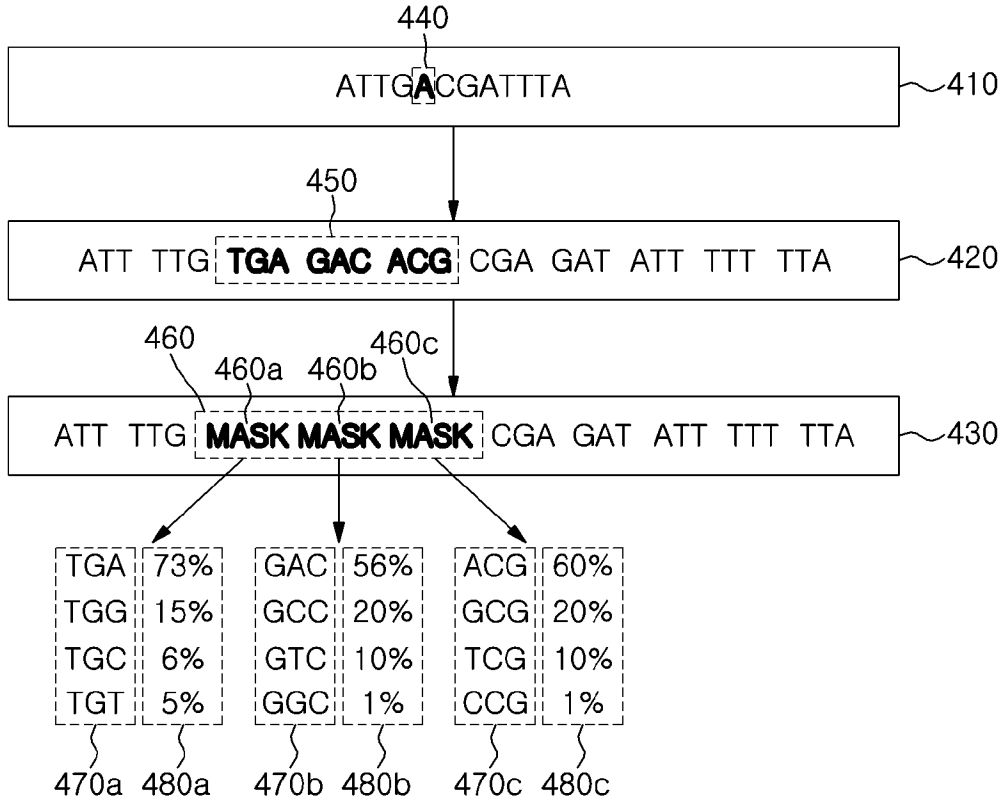
[도2]


220
복수개의 핵산 서열
210
언어 모델
230
마스킹된 베이스의
베이스별 확률값
240
학습용 데이터를 unlabeled data로서
이용하여 self-supervised learning
방식으로 사전 학습

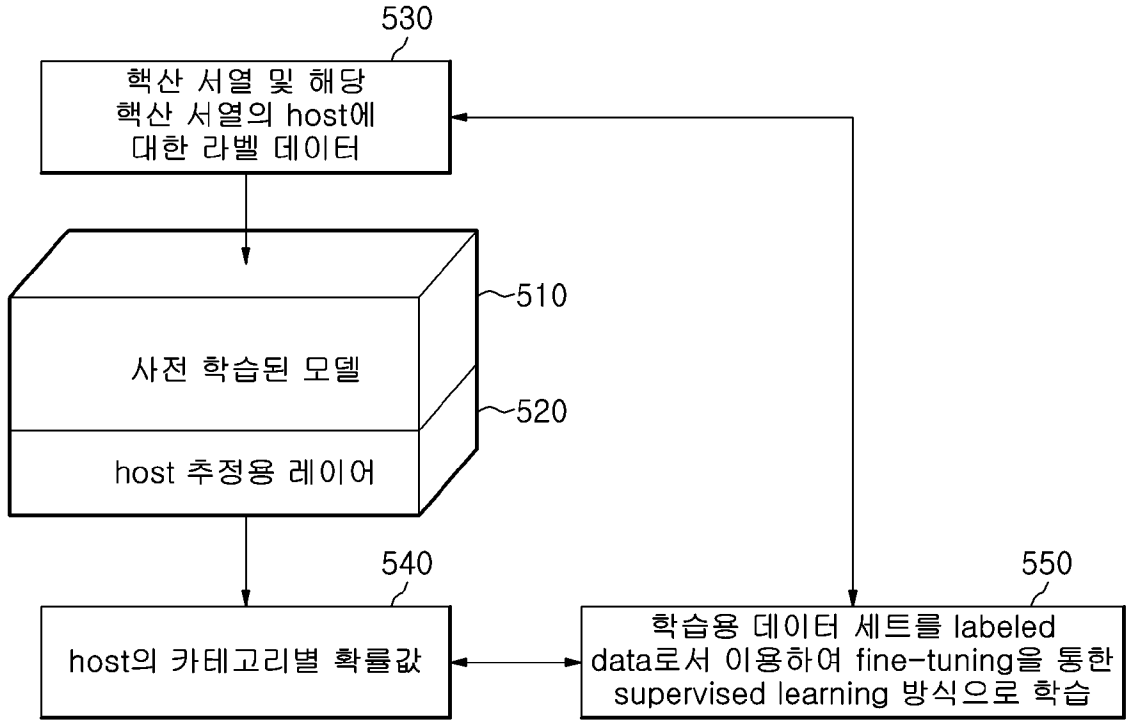
[도3]


Vocab Size
Classifier
Predict
Masked
Products
330
Add & Norm
Feed Forward
Number of
Layers
x
Add & Norm
Multi-Head Attention
320
Positional Encoding
+
Input Embedding
310
P1
P2
[MASK]
P3

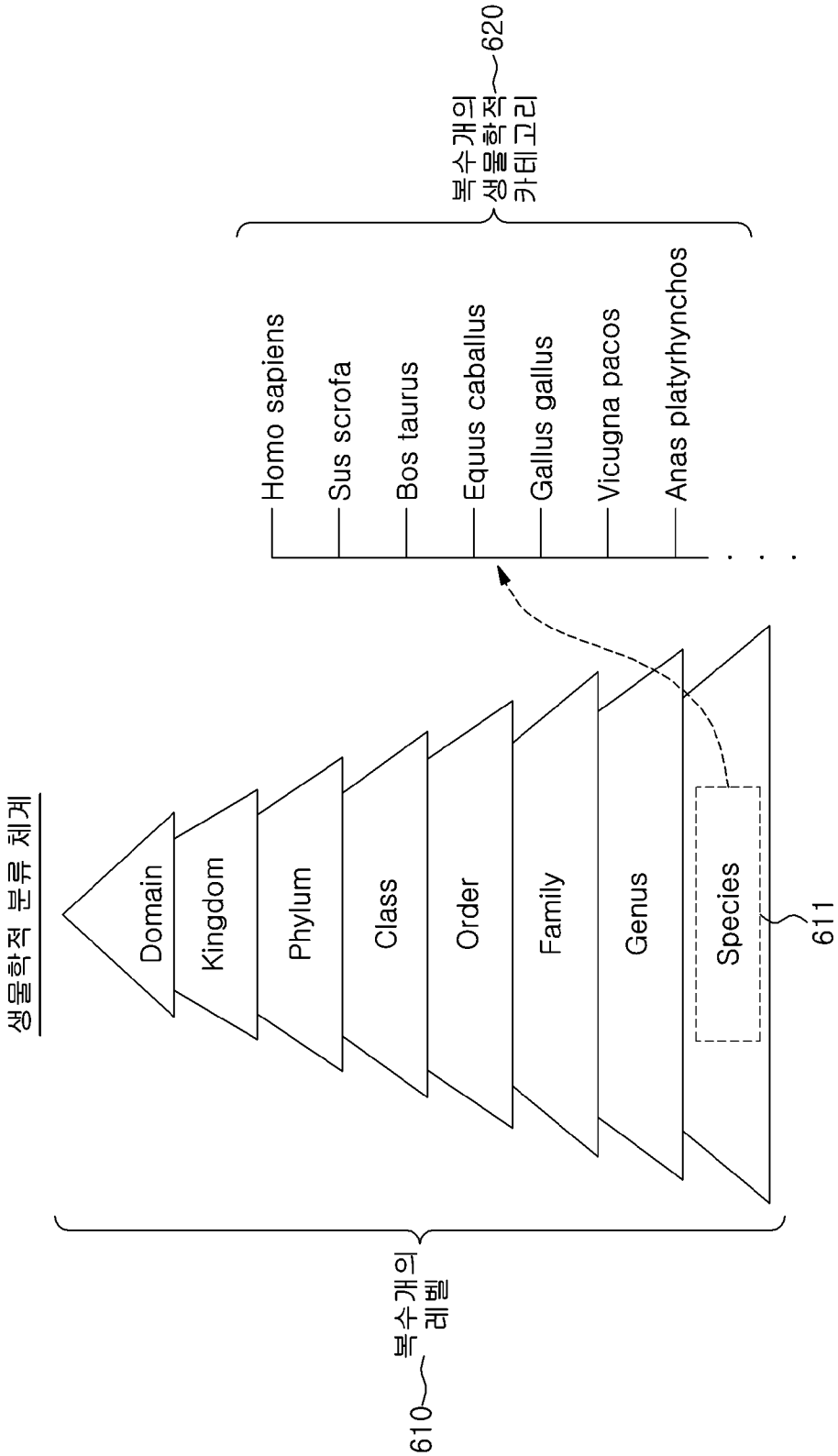
[도4]


440
ATTGACGATTTA
410
450
ATT TTG TGA GAC ACG CGA GAT ATT TTT TTA
420
460
460a
460b
460c
ATT TTG MASK MASK MASK CGA GAT ATT TTT TTA
430
TGA 73%
TGG 15%
TGC 6%
TGT 5%
GAC 56%
GCC 20%
GTC 10%
GGC 1%
ACG 60%
GCG 20%
TCG 10%
CCG 1%
470a 480a 470b 480b 470c 480c

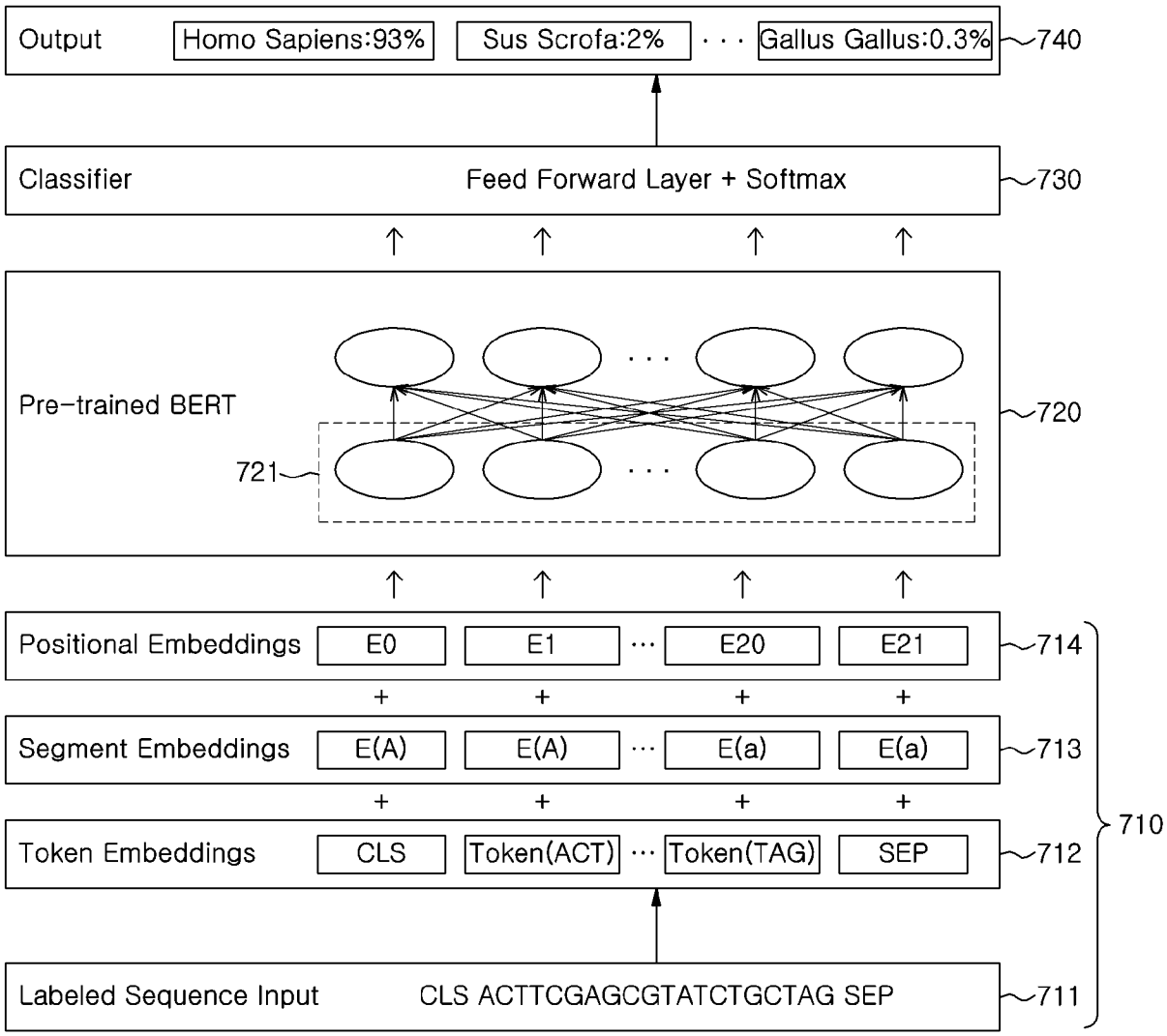
[도5]


530
핵산 서열 및 해당 핵산 서열의 host에 대한 라벨 데이터
510
사전 학습된 모델
520
host 추정용 레이어
540
host의 카테고리별 확률값
550
학습용 데이터 세트를 labeled data로서 이용하여 fine-tuning을 통한 supervised learning 방식으로 학습

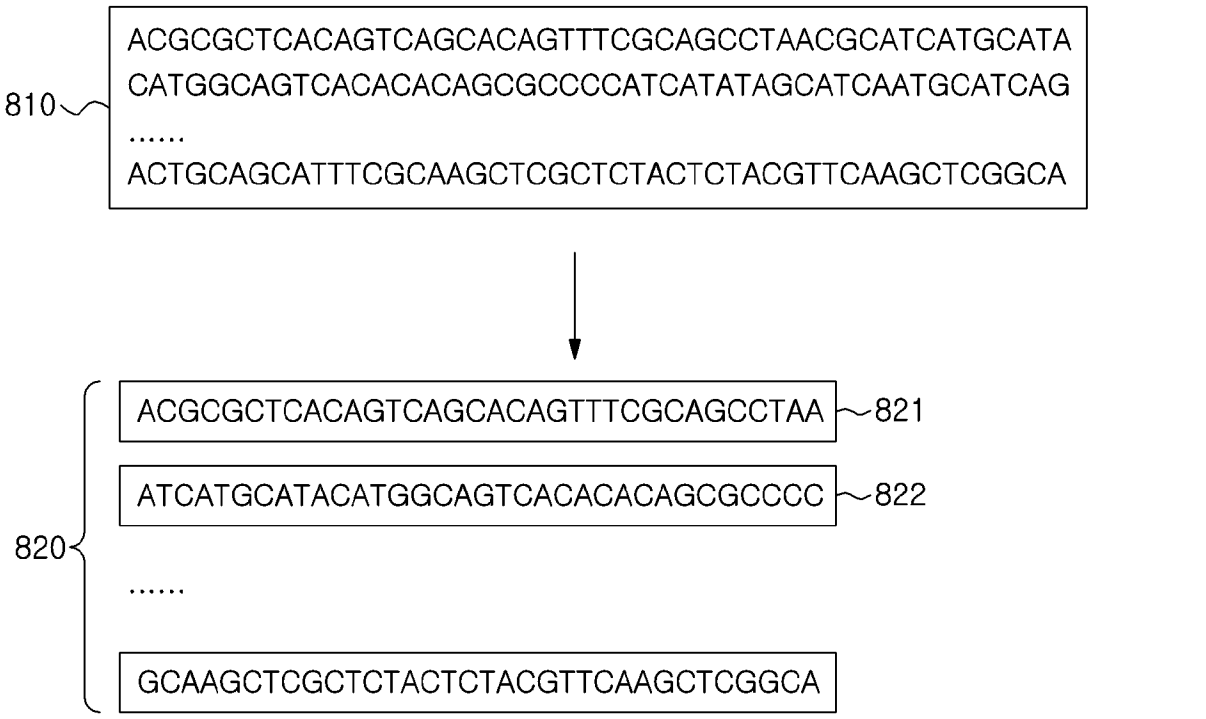
[도6]


생물학적 분류 체계
Domain
Kingdom
Phylum
Class
Order
Family
Genus
Species
611
복수개의
레벨
610
복수개의
생물학적
카테고리
620
Homo sapiens
Sus scrofa
Bos taurus
Equus caballus
Gallus gallus
Vicugna pacos
Anas platyrhynchos
. . .

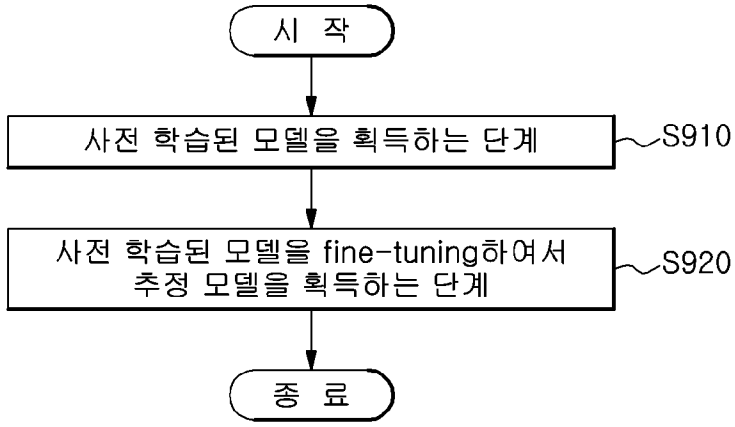
[도7]


Output
Homo Sapiens:93%
Sus Scrofa:2%
Gallus Gallus:0.3%
740
Classifier
Feed Forward Layer + Softmax
730
Pre-trained BERT
721
720
Positional Embeddings
E0
E1
E20
E21
714
Segment Embeddings
E(A)
E(A)
E(a)
E(a)
713
Token Embeddings
CLS
Token(ACT)
Token(TAG)
SEP
712
710
Labeled Sequence Input
CLS ACTTCGAGCGTATCTGCTAG SEP
711

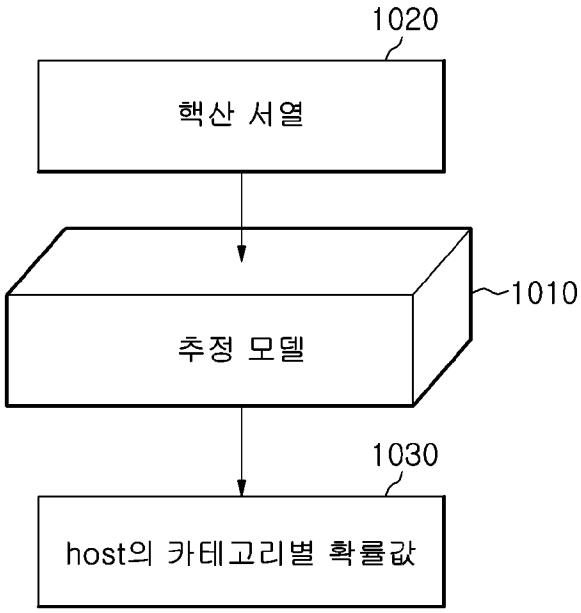
[도8]


810
ACGCGCTCACAGTCAGCACAGTTTCGCAGCCTAACGCATCATGCATA
CATGGCAGTCACACACAGCGCCCCATCATATAGCATCAATGCATCAG
.....
ACTGCAGCATTTCGCAAGCTCGCTCTACTCTACGTTCAAGCTCGGCA
820
ACGCGCTCACAGTCAGCACAGTTTCGCAGCCTAA
821
ATCATGCATACATGGCAGTCACACACAGCGCCCC
822
.....
GCAAGCTCGCTCTACTCTACGTTCAAGCTCGGCA

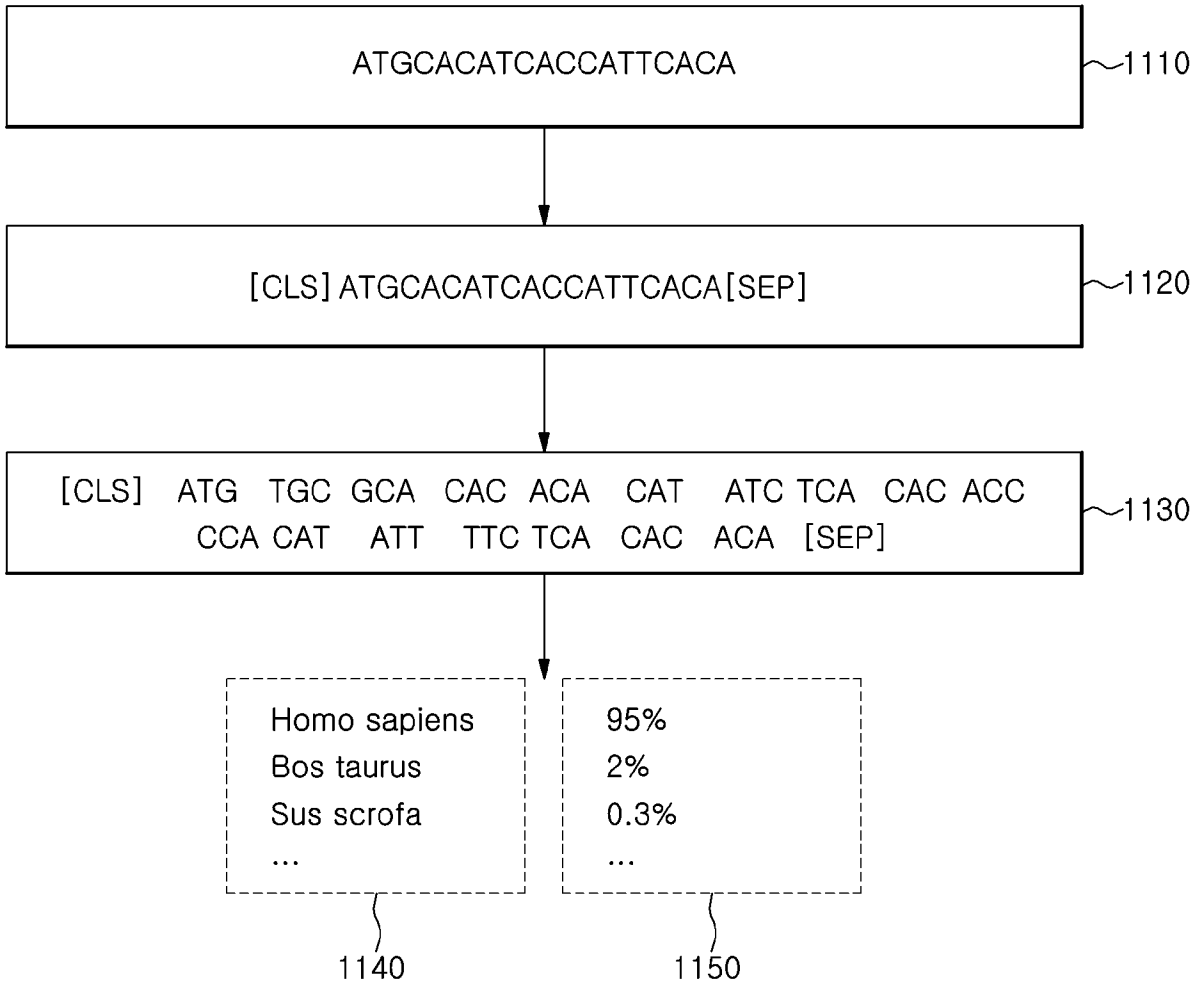
[도9]


시 작
사전 학습된 모델을 획득하는 단계
S910
사전 학습된 모델을 fine-tuning하여서
추정 모델을 획득하는 단계
S920
종 료

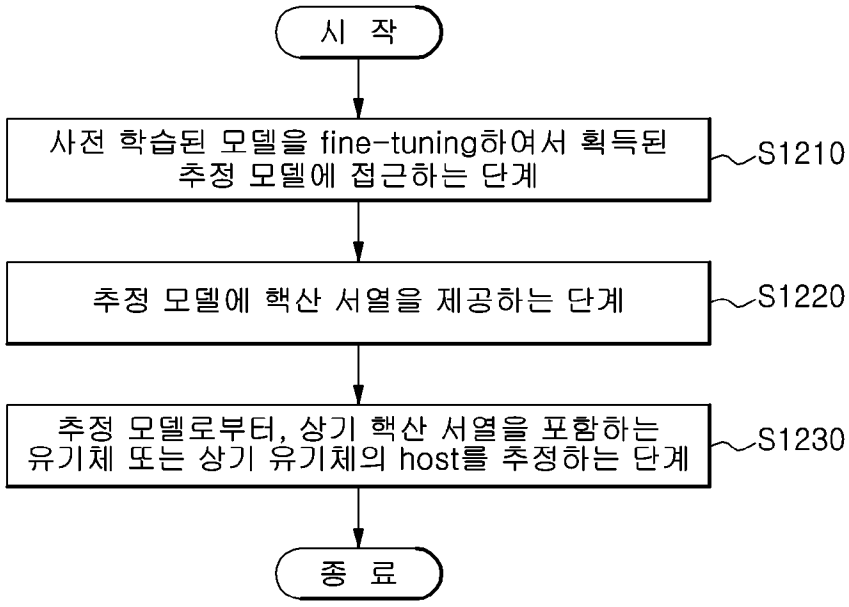
[도10]


1020
핵산 서열
1010
추정 모델
1030
host의 카테고리별 확률값

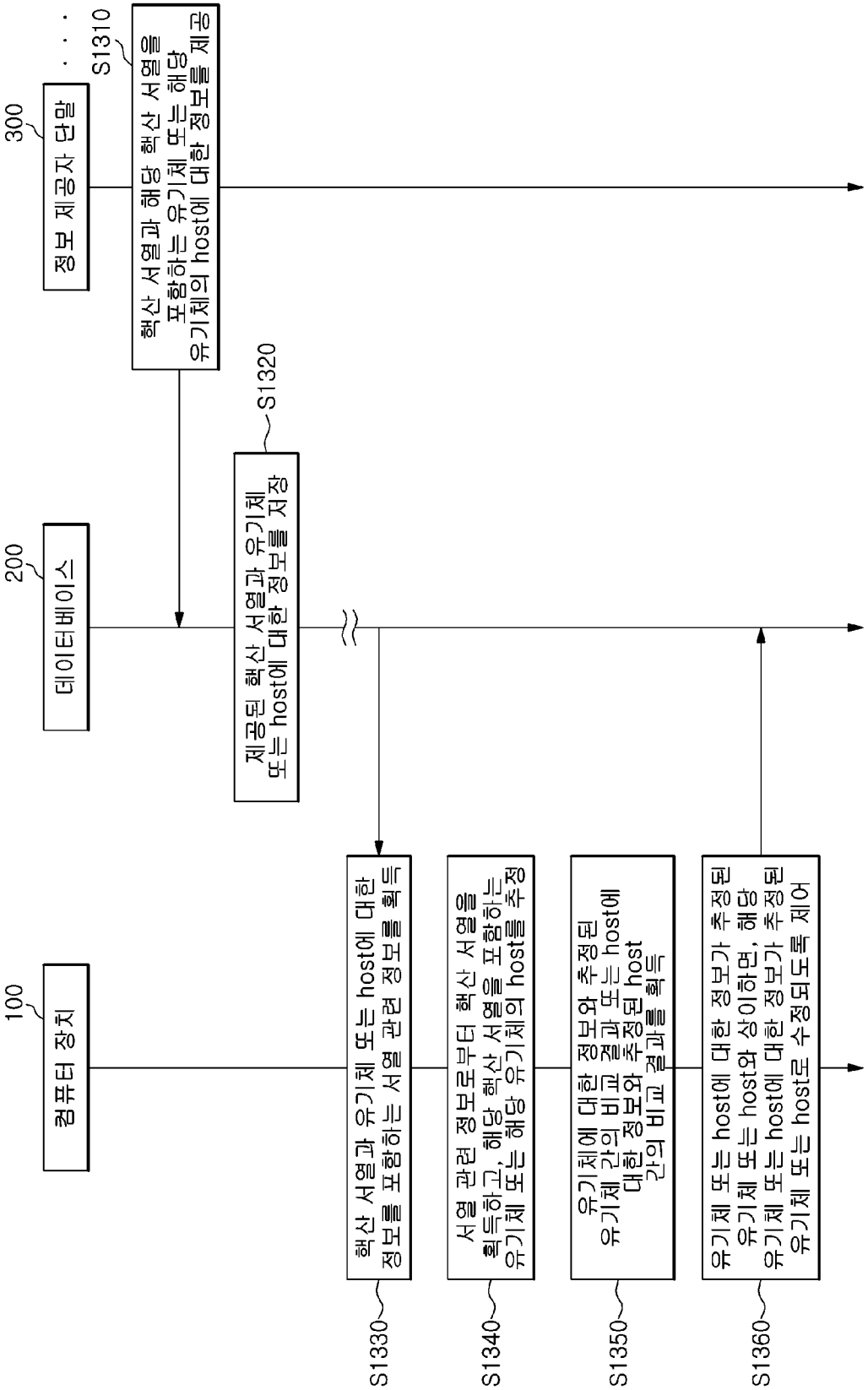
[도11]


ATGCACATCACCATTCACA
1110
[CLS] ATGCACATCACCATTCACA[SEP]
1120
[CLS] ATG TGC GCA CAC ACA CAT ATC TCA CAC ACC CCA CAT ATT TTC TCA CAC ACA [SEP]
1130
Homo sapiens
Bos taurus
Sus scrofa
...
95%
2%
0.3%
...
1140
1150

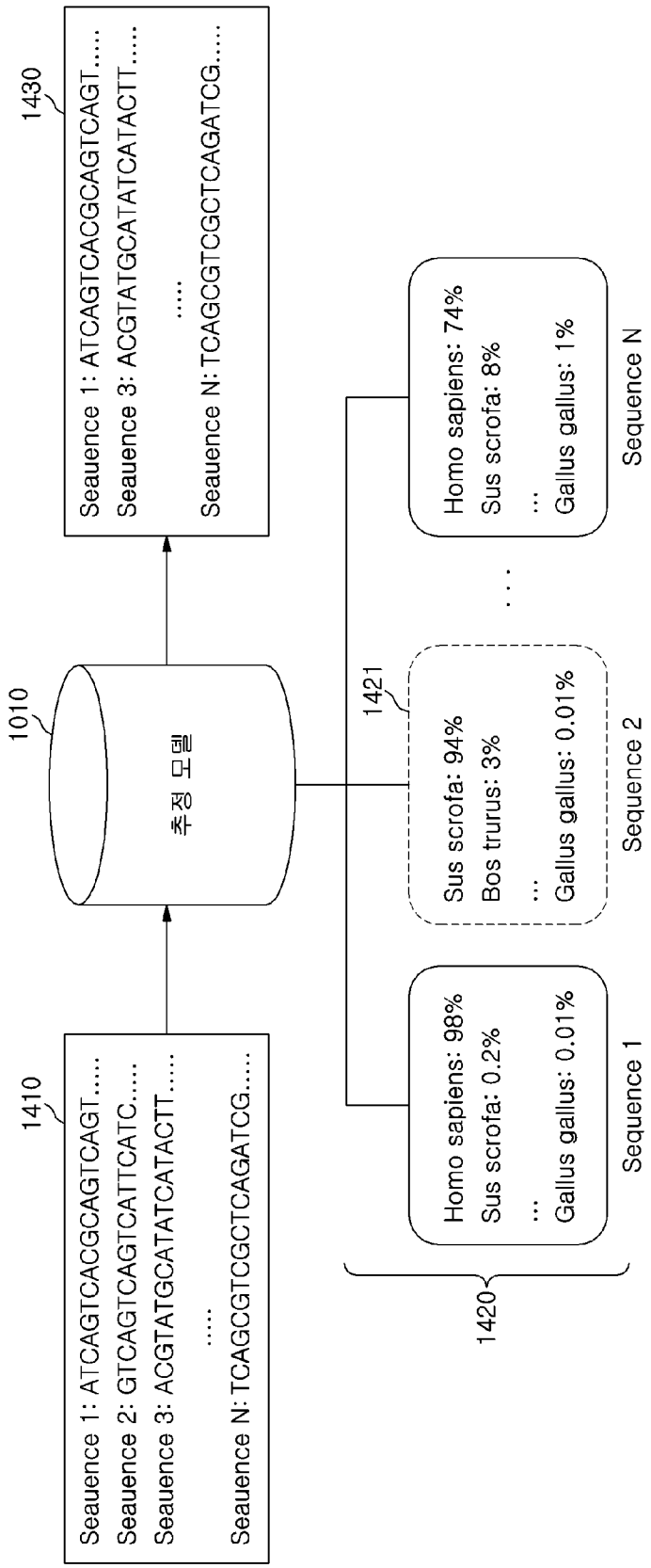
[도12]


시 작
사전 학습된 모델을 fine-tuning하여서 획득된 추정 모델에 접근하는 단계
S1210
추정 모델에 핵산 서열을 제공하는 단계
S1220
추정 모델로부터, 상기 핵산 서열을 포함하는 유기체 또는 상기 유기체의 host를 추정하는 단계
S1230
종 료

[도13]


100
컴퓨터 장치
200
데이터베이스
300
정보 제공자 단말
S1310
핵산 서열과 해당 핵산 서열을 포함하는 유기체 또는 해당 유기체의 host에 대한 정보를 제공
S1320
제공된 핵산 서열과 유기체 또는 host에 대한 정보를 저장
S1330
핵산 서열과 유기체 또는 host에 대한 정보를 포함하는 서열 관련 정보를 획득
S1340
서열 관련 정보로부터 핵산 서열을 획득하고, 해당 핵산 서열을 포함하는 유기체 또는 해당 유기체의 host를 추정
S1350
유기체에 대한 정보와 추정된 유기체 간의 비교결과 또는 host에 대한 정보와 추정된 host 간의 비교 결과를 획득
S1360
유기체 또는 host에 대한 정보가 추정된 유기체 또는 host와 상이하면, 해당 유기체 또는 host에 대한 정보가 추정된 유기체 또는 host로 수정되도록 제어

[도14]


1410
Seauence 1: ATCAGTCACGCAGTCAGT.....
Seauence 2: GTCAGTCAGTCATTCATC.....
Seauence 3: ACGTATGCATATCATACTT.....
.....
Seauence N: TCAGCGTCGCTCAGATCG.....
1010
추정 모델
1430
Seauence 1: ATCAGTCACGCAGTCAGT.....
Seauence 3: ACGTATGCATATCATACTT.....
.....
Seauence N: TCAGCGTCGCTCAGATCG.....
1420
1421
Homo sapiens: 98%
Sus scrofa: 0.2%
...
Gallus gallus: 0.01%
Sequence 1
Sus scrofa: 94%
Bos trurus: 3%
...
Gallus gallus: 0.01%
Sequence 2
Homo sapiens: 74%
Sus scrofa: 8%
...
Gallus gallus: 1%
Sequence N

INTERNATIONAL SEARCH REPORT

International application No.
PCT/KR2023/019095

A. CLASSIFICATION OF SUBJECT MATTER

G16B 40/00(2019.01)i; **G16B 30/10**(2019.01)i; **G16B 45/00**(2019.01)i; **G16B 5/00**(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G16B 40/00(2019.01); C12N 15/11(2006.01); C12Q 1/04(2006.01); G06F 19/22(2011.01); G06N 20/20(2019.01); G06N 3/04(2006.01); G16B 20/30(2019.01); G16B 25/10(2019.01); G16B 35/10(2019.01)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models: IPC as above
Japanese utility models and applications for utility models: IPC as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS (KIPO internal) & keywords: fine-tuning, host, 컴퓨터(computer), 핵산 서열(nucleic acid sequence), 유기체(organism), 토큰화(tokenization)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	KR 10-2014-0087044 A (PATHOGENICA, INC.) 08 July 2014 (2014-07-08) See abstract; and claims 1-40.	1-35
A	CN 114023380 A (INSTITUTE OF QUALITY STANDARD & TESTING TECHNOLOGY FOR AGRO-PRODUCTS, CHINA ACADEMY OF AGRICULTURAL SCIENCES et al.) 08 February 2022 (2022-02-08) See entire document.	1-35
A	KR 10-2396981 B1 (RIIID INC.) 13 May 2022 (2022-05-13) See entire document.	1-35
A	KR 10-2017-0023979 A (10X GENOMICS, INC.) 06 March 2017 (2017-03-06) See entire document.	1-35
A	KR 10-2020-0005607 A (ILLUMINA, INC.) 15 January 2020 (2020-01-15) See entire document.	1-35

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
16 February 2024	**16 February 2024**

Name and mailing address of the ISA/KR	Authorized officer
Korean Intellectual Property Office **Government Complex-Daejeon Building 4, 189 Cheongsa-ro, Seo-gu, Daejeon 35208**	
Facsimile No. **+82-42-481-8578**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/KR2023/019095

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
KR 10-2014-0087044 A	08 July 2014	EP 2788506 A2	15 October 2014
		US 2015-0344977 A1	03 December 2015
		WO 2013-067167 A2	10 May 2013
		WO 2013-067167 A3	11 July 2013
CN 114023380 A	08 February 2022	None	
KR 10-2396981 B1	13 May 2022	KR 10-2023-0014049 A	27 January 2023
		WO 2022-196955 A1	22 September 2022
KR 10-2017-0023979 A	06 March 2017	AU 2015-279546 A1	02 February 2017
		AU 2015-279546 B2	08 April 2021
		CA 2952503 A1	30 December 2015
		CN 106575322 A	19 April 2017
		CN 106575322 B	18 June 2019
		CN 110211637 A	06 September 2019
		CN 110211637 B	27 October 2023
		EP 3161700 A1	03 May 2017
		EP 3161700 B1	29 March 2023
		EP 4235677 A2	30 August 2023
		EP 4235677 A3	22 November 2023
		IL 249461 A	28 February 2017
		IL 249461 B	31 July 2019
		JP 2017-526046 A	07 September 2017
		MX 2016016713 A	23 May 2017
		SG 11201610691 A	27 January 2017
		US 10839939 B2	17 November 2020
		US 11133084 B2	28 September 2021
		US 2015-0379196 A1	31 December 2015
		US 2018-0165411 A1	14 June 2018
		US 2022-0115090 A1	14 April 2022
		WO 2015-200891 A1	30 December 2015
KR 10-2020-0005607 A	15 January 2020	AU 2018-359670 A1	21 November 2019
		AU 2018-359670 B2	26 August 2021
		AU 2021-266189 A1	02 December 2021
		CA 3067421 A1	09 May 2019
		CA 3067421 C	15 August 2023
		CA 3202587 A1	09 May 2019
		CN 110800064 A	14 February 2020
		EP 3707723 A2	16 September 2020
		EP 3707723 B1	25 October 2023
		IL 271239 A	30 January 2020
		JP 2020-528741 A	01 October 2020
		JP 2022-126742 A	30 August 2022
		JP 7091372 B2	27 June 2022
		KR 10-2023-0028569 A	28 February 2023
		US 2019-0218545 A1	18 July 2019
		WO 2019-090251 A2	09 May 2019
		WO 2019-090251 A3	16 January 2020

국 제 조 사 보 고 서

국제출원번호
PCT/KR2023/019095

A. 발명이 속하는 기술분류(국제특허분류(IPC))

G16B 40/00(2019.01)i; **G16B 30/10**(2019.01)i; **G16B 45/00**(2019.01)i; **G16B 5/00**(2019.01)i

B. 조사된 분야

조사된 최소문헌(국제특허분류를 기재)

G16B 40/00(2019.01); C12N 15/11(2006.01); C12Q 1/04(2006.01); G06F 19/22(2011.01); G06N 20/20(2019.01); G06N 3/04(2006.01); G16B 20/30(2019.01); G16B 25/10(2019.01); G16B 35/10(2019.01)

조사된 기술분야에 속하는 최소문헌 이외의 문헌

한국등록실용신안공보 및 한국공개실용신안공보: 조사된 최소문헌란에 기재된 IPC
일본등록실용신안공보 및 일본공개실용신안공보: 조사된 최소문헌란에 기재된 IPC

국제조사에 이용된 전산 데이터베이스(데이터베이스의 명칭 및 검색어(해당하는 경우))

eKOMPASS(특허청 내부 검색시스템) & 키워드: fine-tuning, host, 컴퓨터(computer), 핵산 서열(nucleic acid sequence), 유기체(organism), 토큰화(tokenization)

C. 관련 문헌

카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
A	KR 10-2014-0087044 A (패소제니카 아이엔씨.) 2014.07.08 요약; 청구항 1-40	1-35
A	CN 114023380 A (INSTITUTE OF QUALITY STANDARD AND TESTING TECHNOLOGY FOR AGRO-PRODUCTS, CHINESE ACADAMY OF AGRICULTURAL SCIENCES 등) 2022.02.08 전체 문헌	1-35
A	KR 10-2396981 B1 ((주)뤼이드) 2022.05.13 전체 문헌	1-35
A	KR 10-2017-0023979 A (10엑스 제노믹스, 인크.) 2017.03.06 전체 문헌	1-35
A	KR 10-2020-0005607 A (일루미나, 인코포레이티드) 2020.01.15 전체 문헌	1-35

☐ 추가 문헌이 C(계속)에 기재되어 있습니다.

☑ 대응특허에 관한 별지를 참조하십시오.

* 인용된 문헌의 특별 카테고리:

"A" 특별히 관련이 없는 것으로 보이는 일반적인 기술수준을 정의한 문헌
"D" 본 국제출원에서 출원인이 인용한 문헌
"E" 국제출원일보다 빠른 출원일 또는 우선일을 가지나 국제출원일 이후에 공개된 선출원 또는 특허 문헌
"L" 우선권 주장에 의문을 제기하는 문헌 또는 다른 인용문헌의 공개일 또는 다른 특별한 이유(이유를 명시)를 밝히기 위하여 인용된 문헌
"O" 구두 개시, 사용, 전시 또는 기타 수단을 언급하고 있는 문헌
"P" 우선일 이후에 공개되었으나 국제출원일 이전에 공개된 문헌
"T" 국제출원일 또는 우선일 후에 공개된 문헌으로, 출원과 상충하지 않으며 발명의 기초가 되는 원리나 이론을 이해하기 위해 인용된 문헌
"X" 특별한 관련이 있는 문헌. 해당 문헌 하나만으로 청구된 발명의 신규성 또는 진보성이 없는 것으로 본다.
"Y" 특별한 관련이 있는 문헌. 해당 문헌이 하나 이상의 다른 문헌과조합하는 경우로 그 조합이 당업자에게 자명한 경우 청구된 발명은 진보성이 없는 것으로 본다.
"&" 동일한 대응특허문헌에 속하는 문헌

국제조사의 실제 완료일	국제조사보고서 발송일
2024년02월16일(16.02.2024)	**2024년02월16일(16.02.2024)**

ISA/KR의 명칭 및 우편주소

대한민국 특허청
(35208) 대전광역시 서구 청사로 189, 4동 (둔산동, 정부대전청사)

팩스 번호 **+82-42-481-8578**

심사관

장정아

전화번호 **+82-41-481-5955**

국 제 조 사 보 고 서
대응특허에 관한 정보

국제출원번호
PCT/KR2023/019095

국제조사보고서에서 인용된 특허문헌	공개일	대응특허문헌	공개일
KR 10-2014-0087044 A	2014/07/08	EP 2788506 A2	2014/10/15
		US 2015-0344977 A1	2015/12/03
		WO 2013-067167 A2	2013/05/10
		WO 2013-067167 A3	2013/07/11
CN 114023380 A	2022/02/08	없음	
KR 10-2396981 B1	2022/05/13	KR 10-2023-0014049 A	2023/01/27
		WO 2022-196955 A1	2022/09/22
KR 10-2017-0023979 A	2017/03/06	AU 2015-279546 A1	2017/02/02
		AU 2015-279546 B2	2021/04/08
		CA 2952503 A1	2015/12/30
		CN 106575322 A	2017/04/19
		CN 106575322 B	2019/06/18
		CN 110211637 A	2019/09/06
		CN 110211637 B	2023/10/27
		EP 3161700 A1	2017/05/03
		EP 3161700 B1	2023/03/29
		EP 4235677 A2	2023/08/30
		EP 4235677 A3	2023/11/22
		IL 249461 A	2017/02/28
		IL 249461 B	2019/07/31
		JP 2017-526046 A	2017/09/07
		MX 2016016713 A	2017/05/23
		SG 11201610691 A	2017/01/27
		US 10839939 B2	2020/11/17
		US 11133084 B2	2021/09/28
		US 2015-0379196 A1	2015/12/31
		US 2018-0165411 A1	2018/06/14
		US 2022-0115090 A1	2022/04/14
		WO 2015-200891 A1	2015/12/30
KR 10-2020-0005607 A	2020/01/15	AU 2018-359670 A1	2019/11/21
		AU 2018-359670 B2	2021/08/26
		AU 2021-266189 A1	2021/12/02
		CA 3067421 A1	2019/05/09
		CA 3067421 C	2023/08/15
		CA 3202587 A1	2019/05/09
		CN 110800064 A	2020/02/14
		EP 3707723 A2	2020/09/16
		EP 3707723 B1	2023/10/25
		IL 271239 A	2020/01/30
		JP 2020-528741 A	2020/10/01
		JP 2022-126742 A	2022/08/30
		JP 7091372 B2	2022/06/27
		KR 10-2023-0028569 A	2023/02/28
		US 2019-0218545 A1	2019/07/18
		WO 2019-090251 A2	2019/05/09
		WO 2019-090251 A3	2020/01/16