(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0013547
(43) 공개일자 2024년01월30일

(51) 국제특허분류(Int. Cl.)
G16B 30/10 (2019.01) *C12Q 1/6869* (2018.01)
G16B 40/20 (2019.01) *G16B 50/00* (2019.01)
(52) CPC특허분류
G16B 30/10 (2019.02)
C12Q 1/6869 (2018.05)
(21) 출원번호 10-2022-0091233
(22) 출원일자 2022년07월22일
심사청구일자 2022년07월22일

전체 청구항 수 : 총 11 항

(71) 출원인
인천대학교 산학협력단
인천광역시 연수구 갯벌로 27, 인천대학교 이노베이션센터(송도동)
(72) 발명자
전광길
인천광역시 연수구 해송로 70, 209동 805호(송도동, 송도웰카운티2단지)
(74) 대리인
특허법인충정

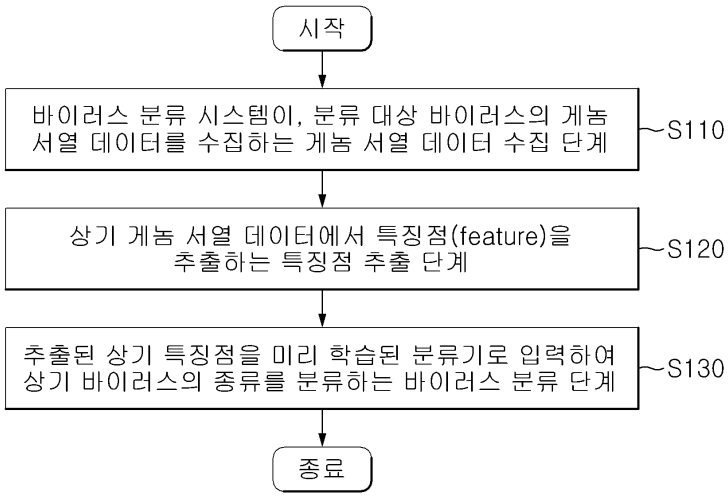
(54) 발명의 명칭 **인공 지능을 이용한 게놈 서열 분석 기반 ＣＯＶＩＤ－１９ 및 유사 바이러스의 분류 방법, 장치 및 시스템**

(57) 요 약

본 발명은 COVID-19 및 유사 바이러스의 분류 방법, 장치 및 시스템에 관한 것으로서, 보다 구체적으로는 인공지능을 이용한 게놈 서열 분석에 기반하여 COVID-19 및 유사 바이러스를 분류하는 방법, 장치 및 시스템에 관한 것이다.

본 발명에서는, 바이러스의 게놈 서열을 기반으로 바이러스의 종류를 분류하는 방법에 있어서, 바이러스 분류 시스템이, 분류 대상 바이러스의 게놈 서열 데이터를 수집하는 게놈 서열 데이터 수집 단계; 상기 게놈 서열 데이터에서 특징점(feature)을 추출하는 특징점 추출 단계; 및 추출된 상기 특징점을 미리 학습된 분류기로 입력하여 상기 바이러스의 종류를 분류하는 바이러스 분류 단계;를 포함하는 것을 특징으로 하는 바이러스 분류 방법을 개시한다.

대 표 도 - 도2



(52) CPC특허분류
G16B 40/20 (2019.02)
G16B 50/00 (2019.02)
C12Q 2525/149 (2013.01)

명 세 서

청구범위

청구항 1

바이러스의 게놈 서열을 기반으로 바이러스의 종류를 분류하는 방법에 있어서,

바이러스 분류 시스템이, 분류 대상 바이러스의 게놈 서열 데이터를 수집하는 게놈 서열 데이터 수집 단계;

상기 게놈 서열 데이터에서 특징점(feature)을 추출하는 특징점 추출 단계; 및

추출된 상기 특징점을 미리 학습된 분류기로 입력하여 상기 바이러스의 종류를 분류하는 바이러스 분류 단계;

를 포함하는 것을 특징으로 하는 바이러스 분류 방법.

청구항 2

제1항에 있어서,

상기 바이러스에는 코로나(COVID-19), 사스(SARS), 중동호흡기증후군(MERS), 에볼라(Ebola) 바이러스 중 하나 또는 둘 이상이 포함되는 것을 특징으로 하는 것을 특징으로 하는 바이러스 분류 방법.

청구항 3

제1항에 있어서,

상기 특징점 추출 단계에서는,

상기 게놈 서열 데이터에서 뉴클레오티드(nucleotide)의 빈도(frequency)를 특징점으로 추출하는 것을 특징으로 하는 것을 특징으로 하는 바이러스 분류 방법.

청구항 4

제1항에 있어서,

상기 특징점 추출 단계에서는,

상기 게놈 서열 데이터에서 트리-뉴클레오티드(tri-nucleotides)의 빈도(frequency)를 특징점으로 추출하는 것을 특징으로 하는 것을 특징으로 하는 바이러스 분류 방법.

청구항 5

제1항에 있어서,

상기 특징점 추출 단계에서는,

상기 게놈 서열 데이터에서 아미노산의 조성(composition of amino acids)을 특징점으로 추출하는 것을 특징으로 하는 것을 특징으로 하는 바이러스 분류 방법.

청구항 6

제1항에 있어서,

상기 특징점 추출 단계에서는,

상기 게놈 서열 데이터에서 뉴클레오티드 트리플렛의 서열(sequence of nucleotide triplets)을 특징점으로 추출하는 것을 특징으로 하는 것을 특징으로 하는 바이러스 분류 방법.

청구항 7

제1항에 있어서,

상기 특징점 추출 단계에서는,

상기 게놈 서열 데이터에서 코딩 서열(coding sequence, CDS)을 특징점으로 추출하는 것을 특징으로 하는 것을 특징으로 하는 바이러스 분류 방법.

청구항 8

제1항에 있어서,

상기 특징점 추출 단계에서는,

상기 게놈 서열 데이터에서 바이러스 종류간 정렬 유사도(alignment similarity)를 특징점으로 추출하는 것을 특징으로 하는 것을 특징으로 하는 바이러스 분류 방법.

청구항 9

제1항에 있어서,

상기 특징점 추출 단계에서는,

상기 게놈 서열 데이터에서 바이러스 종류간 게놈 서열의 차이를 보여주는 도트 플롯(Dot Plot) 이미지를 특징점으로 추출하는 것을 특징으로 하는 것을 특징으로 하는 바이러스 분류 방법.

청구항 10

제1항에 있어서,

상기 바이러스 분류 단계에서,

상기 분류기는, 훈련 데이터를 이용하여 뉴클레오티드(nucleotide)의 빈도(frequency), 트리-뉴클레오티드(tri-nucleotides)의 빈도(frequency), 아미노산의 조성(composition of amino acids), 뉴클레오티드 트리플렛의 서열(sequence of nucleotide triplets), 코딩 서열(coding sequence, CDS), 바이러스 종류간 정렬 유사도(alignment similarity), 바이러스 종류간 게놈 서열의 차이를 보여주는 도트 플롯(Dot Plot) 이미지 중 하나 이상을 이용하여 학습된 지원 벡터 머신(Support Vector Machine)을 기반으로 구성되는 것을 특징으로 하는 것을 특징으로 하는 바이러스 분류 방법.

청구항 11

바이러스의 게놈 서열을 기반으로 바이러스의 종류를 분류하는 바이러스 분류 시스템에 있어서,

분류 대상 바이러스의 게놈 서열 데이터를 수집하는 게놈 서열 데이터 수집부;

상기 게놈 서열 데이터에서 특징점(feature)을 추출하는 특징점 추출부; 및

추출된 상기 특징점을 미리 학습된 분류기로 입력하여 상기 바이러스의 종류를 분류하는 바이러스 분류부;

를 포함하는 것을 특징으로 하는 바이러스 분류 시스템.

발명의 설명

기 술 분 야

[0001] 본 발명은 COVID-19 및 유사 바이러스의 분류 방법, 장치 및 시스템에 관한 것으로서, 보다 구체적으로는 인공지능을 이용한 게놈 서열 분석에 기반하여 COVID-19 및 유사 바이러스를 분류하는 방법, 장치 및 시스템에 관한 것이다.

배 경 기 술

[0002] 최근 SARS-CoV-2로 인한 COVID-19(코로나 바이러스)가 빠른 속도로 확산하면서 수많은 사람들을 감염시키고 인류의 건강과 안위에 심각한 영향을 미치고 있다.

[0003] 중국에서 시작된 COVID-19 또는 SARS-CoV-2라고도 하는 신종 코로나 바이러스는 급속히 확산되면서 전 세계적인 보건 비상 사태를 야기하고 있으며, 이에 대하여 WHO(세계보건기구)에서는 2020년 3월 11일 전 세계적인 팬데믹

으로 공식 보고하였고, 2021년 7월 현재까지 COVID-19 전염병은 200개 이상의 국가/영토에 영향을 미쳤으며 4,067,517명의 사망을 포함하여 188,655,968명 이상의 감염 사례가 확인되었다.

[0004] 이에 따라, 각국 정부와 보건 기관들은 치명적인 바이러스의 전파 및 확산에 대응하기 위한 예방 및 검역 조치를 실행하고 있으며, 관련 연구자들도 효과적인 백신 또는 약물을 개발/생산하기 위해 COVID-19의 게놈, 그 기능 및 행동을 파악하고 대응 방안을 도출하기 위하여 노력하고 있다.

[0005] 이와 관련하여, 게놈 서열 분석과 첨단 인공 지능 기술은 연구자와 의료 전문가가 COVID-19 또는 SARS-CoV-2의 유전적 변이체를 이해하는 데 도움이 될 수 있으며, COVID-19의 게놈 서열 분석은 바이러스의 기원, 행동 및 구조를 이해하고, 백신, 항바이러스제의 생산/개발 및 효율적인 예방 전략의 수립에 도움이 될 수 있다.

[0006] 이에 따라, COVID-19 및 유사 바이러스에 대한 게놈 서열 분석을 기반으로 인공 지능을 이용하여 COVID-19 및 유사 바이러스를 분류할 수 있는 방안이 요구되고 있으나, 아직 이에 대한 적절한 해법이 제시되지 못하고 있다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 대한민국 공개특허공보 제10-2022-0053642호(2022년 4월 29일 공개)

발명의 내용

해결하려는 과제

[0008] 본 발명은 상기와 같은 종래 기술의 문제점을 해결하기 위해 창안된 것으로, COVID-19 및 유사 바이러스에 대한 게놈 서열 분석을 기반으로 인공 지능을 이용하여 COVID-19 및 유사 바이러스를 분류할 수 있는 인공 지능을 이용한 게놈 서열 분석 기반 COVID-19 및 유사 바이러스의 분류 방법, 장치 및 시스템을 제공하는 것을 목적으로 한다.

[0009] 그 외 본 발명의 세부적인 목적은 아래에 기재되는 구체적인 내용을 통하여 이 기술 분야의 전문가나 연구자에게 자명하게 파악되고 이해될 수 있을 것이다.

과제의 해결 수단

[0010] 상기 과제를 해결하기 위한 본 발명의 한 측면에 따른 바이러스 분류 방법은, 바이러스의 게놈 서열을 기반으로 바이러스의 종류를 분류하는 방법에 있어서, 바이러스 분류 시스템이, 분류 대상 바이러스의 게놈 서열 데이터를 수집하는 게놈 서열 데이터 수집 단계; 상기 게놈 서열 데이터에서 특징점(feature)을 추출하는 특징점 추출 단계; 및 추출된 상기 특징점을 미리 학습된 분류기로 입력하여 상기 바이러스의 종류를 분류하는 바이러스 분류 단계;를 포함하는 것을 특징으로 한다.

[0011] 여기서, 상기 바이러스에는 코로나(COVID-19), 사스(SARS), 중동호흡기증후군(MERS), 에볼라(Ebola) 바이러스 중 하나 또는 둘 이상이 포함될 수 있다.

[0012] 또한, 상기 특징점 추출 단계에서는, 상기 게놈 서열 데이터에서 뉴클레오티드(nucleotide)의 빈도(frequency)를 특징점으로 추출할 수 있다.

[0013] 또한, 상기 특징점 추출 단계에서는, 상기 게놈 서열 데이터에서 트리-뉴클레오티드(tri-nucleotides)의 빈도(frequency)를 특징점으로 추출할 수 있다.

[0014] 또한, 상기 특징점 추출 단계에서는, 상기 게놈 서열 데이터에서 아미노산의 조성(composition of amino acids)을 특징점으로 추출할 수 있다.

[0015] 또한, 상기 특징점 추출 단계에서는, 상기 게놈 서열 데이터에서 뉴클레오티드 트리플렛의 서열(sequence of nucleotide triplets)을 특징점으로 추출할 수 있다.

[0016] 또한, 상기 특징점 추출 단계에서는, 상기 게놈 서열 데이터에서 코딩 서열(coding sequence, CDS)을 특징점으로 추출할 수 있다.

[0017] 또한, 상기 특징점 추출 단계에서는, 상기 게놈 서열 데이터에서 바이러스 종류간 정렬 유사도(alignment similarity)를 특징점으로 추출할 수 있다.

[0018] 또한, 상기 특징점 추출 단계에서는, 상기 게놈 서열 데이터에서 바이러스 종류간 게놈 서열의 차이를 보여주는 도트 플롯(Dot Plot) 이미지를 특징점으로 추출할 수 있다.

[0019] 또한, 상기 바이러스 분류 단계에서, 상기 분류기는, 훈련 데이터를 이용하여 뉴클레오티드(nucleotide)의 빈도(frequency), 트리-뉴클레오티드(tri-nucleotides)의 빈도(frequency), 아미노산의 조성(composition of amino acids), 뉴클레오티드 트리플렛의 서열(sequence of nucleotide triplets), 코딩 서열(coding sequence, CDS), 바이러스 종류간 정렬 유사도(alignment similarity), 바이러스 종류간 게놈 서열의 차이를 보여주는 도트 플롯(Dot Plot) 이미지 중 하나 이상을 이용하여 학습된 지원 벡터 머신(Support Vector Machine)을 기반으로 구성될 수 있다.

[0020] 또한, 본 발명의 다른 측면에 따른 바이러스 분류 시스템은, 바이러스의 게놈 서열을 기반으로 바이러스의 종류를 분류하는 바이러스 분류 시스템에 있어서, 분류 대상 바이러스의 게놈 서열 데이터를 수집하는 게놈 서열 데이터 수집부; 상기 게놈 서열 데이터에서 특징점(feature)을 추출하는 특징점 추출부; 및 추출된 상기 특징점을 미리 학습된 분류기로 입력하여 상기 바이러스의 종류를 분류하는 바이러스 분류부;를 포함하는 것을 특징으로 한다.

발명의 효과

[0021] 이에 따라, 본 발명의 일 실시예에 따른 인공 지능을 이용한 게놈 서열 분석 기반 COVID-19 및 유사 바이러스의 분류 방법, 장치 및 시스템에서는, COVID-19 및 유사 바이러스에 대한 게놈 서열 분석을 기반으로 인공 지능을 이용하여 COVID-19 및 유사 바이러스를 분류할 수 있게 된다.

도면의 간단한 설명

[0022] 본 발명에 관한 이해를 돕기 위해 상세한 설명의 일부로 포함되는, 첨부도면은 본 발명에 대한 실시예를 제공하고, 상세한 설명과 함께 본 발명의 기술적 사상을 설명한다.

도 1 은 본 발명의 일 실시예에 따른 바이러스 분류 시스템(100)의 동작을 설명하는 도면이다.

도 2는 본 발명의 일 실시예에 따른 바이러스 분류 방법의 순서도이다.

도 3 내지 및 도 4는 본 발명의 일 실시예에 따른 바이러스 분류 방법의 동작을 설명하는 도면이다.

도 5 내지 도 11은 본 발명의 일 실시예에 따른 바이러스 분류 방법에서의 특징점 추출을 설명하는 도면이다.

도 12 내지 도 14는 본 발명의 일 실시예에 따른 바이러스 분류 방법의 결과와 성능을 예시하는 도면이다.

도 15는 본 발명의 일 실시예에 따른 바이러스 분류 시스템의 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0023] 본 발명은 다양한 변환을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 이하에서는 특정 실시예들을 첨부된 도면을 기초로 상세히 설명하고자 한다.

[0024] 이하의 실시예는 본 명세서에서 기술된 방법, 장치 및/또는 시스템에 대한 포괄적인 이해를 돕기 위해 제공된다. 그러나 이는 예시에 불과하며 본 발명은 이에 제한되지 않는다.

[0025] 본 발명의 실시예들을 설명함에 있어서, 본 발명과 관련된 공지기술에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략하기로 한다. 그리고, 후술되는 용어들은 본 발명에서의 기능을 고려하여 정의된 용어들로서 이는 사용자, 운용자의 의도 또는 관례 등에 따라 달라질 수 있다. 그러므로 그 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다. 상세한 설명에서 사용되는 용어는 단지 본 발명의 실시 예들을 기술하기 위한 것이며, 결코 제한적이어서는 안 된다. 명확하게 달리 사용되지 않는 한, 단수 형태의 표현은 복수 형태의 의미를 포함한다. 본 설명에서, "포함" 또는 "구비"와 같은 표현은 어떤 특성들, 숫자들, 단계들, 동작들, 요소들, 이들의 일부 또는 조합을 가리키기 위한 것이며, 기술된 것 이외에 하나 또는 그 이상의 다른 특성, 숫자, 단계, 동작, 요소, 이들의 일부 또는 조합의 존재 또는 가능성을 배제하도록 해석되어서는 안 된다.

[0026] 또한, 제1, 제2 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되는 것은 아니며, 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다.

[0027] 먼저, 도 1에서는 본 발명의 일 실시예에 따른 바이러스 분류 시스템(100)의 동작을 설명하기 위한 도면을 예시하고 있다. 도 1에서 볼 수 있는 바와 같이, 본 발명의 일 실시예에 따른 바이러스 분류 시스템(100)은 통신 네트워크(30)를 통해 하나 이상의 게놈 서열 입력 장치(20a, 20b)와 연결되어 분석 대상 바이러스에 대한 게놈 서열을 전송받아 바이러스에 대한 분류 작업을 수행할 수 있다.

[0028] 이때, 본 발명에서 상기 바이러스에는 코로나(COVID-19), 사스(SARS), 중동호흡기증후군(MERS), 에볼라(Ebola) 바이러스 중 하나 또는 둘 이상이 포함될 수 있다.

[0029] 또한, 상기 바이러스 분류 시스템(100)은 하나의 서버 컴퓨터로 구현되거나, 둘 이상의 서버 컴퓨터가 연동되어 구현될 수도 있으며, 상기 서버 컴퓨터의 예로는 서버 컴퓨팅 디바이스, 개인용 컴퓨터, 미니 컴퓨터, 및/또는 메인프레임 컴퓨터를 포함할 수 있지만 이러한 것으로만 한정되는 것은 아니고, 상기 서버 컴퓨터는 분산형 시스템일 수 있으며, 서버 컴퓨터의 동작들은 하나 이상의 프로세서들 상에서 동시에 그리고/또는 순차적으로 실행될 수 있다.

[0030] 나아가, 상기 바이러스 분류 시스템(100)은 클라우드 시스템을 이용하여 구현되거나 전용 하드웨어 등을 사용하여 별도의 장치로 구현되는 등 매우 다양한 형태로 구현하는 것이 가능하다.

[0031] 또한, 상기 게놈 서열 입력 장치(20)에는 분석 대상 바이러스의 게놈 서열을 전송할 수 있는 PC(personal computer), 노트북 컴퓨터(laptop computer), 랩탑 컴퓨터(laptop computer), 스마트폰(smart phone), 태블릿(tablet) 등 단말이거나, 상기 바이러스에 대한 분석을 통해 게놈 서열을 생성하는 모듈 등을 포함하여 구성하는 등 다양한 장치들이 폭넓게 포함될 수 있다.

[0032] 또한, 상기 통신 네트워크(30)로서는 유선 네트워크와 무선 네트워크를 포함할 수 있으며, 구체적으로, 근거리 통신망 (LAN: Local Area Network), 도시권 통신망 (MAN: Metropolitan Area Network), 광역 통신망 (WAN: Wide Area Network) 등의 다양한 통신망을 포함할 수 있다. 또한, 상기 통신 네트워크(30)는 공지의 월드 와이드 웹(WWW: World Wide Web)을 포함할 수도 있다. 그러나, 본 발명에서 통신 네트워크(30)는 상기 열거된 네트워크에 국한되지 않고, 공지의 무선 데이터 네트워크나 공지의 전화 네트워크 또는 공지의 유무선 텔레비전 네트워크를 적어도 일부로 포함할 수도 있다.

[0033] 또한, 도 2에서는 본 발명의 일 실시예에 따른 바이러스 분류 방법의 순서도를 도시하고 있다.

[0034] 도 2에서 볼 수 있는 바와 같이, 본 발명의 일 실시예에 따른 바이러스 분류 방법은, 바이러스의 게놈 서열을 기반으로 바이러스의 종류를 분류하는 방법으로서, 바이러스 분류 시스템(100)이, 분류 대상 바이러스의 게놈 서열 데이터를 수집하는 게놈 서열 데이터 수집 단계(S110), 상기 게놈 서열 데이터에서 특징점(feature)을 추출하는 특징점 추출 단계(S120) 및 추출된 상기 특징점을 미리 학습된 분류기로 입력하여 상기 바이러스의 종류를 분류하는 바이러스 분류 단계(S130)를 포함하여 구성될 수 있다.

[0035] 여기서, 상기 바이러스에는 코로나(COVID-19), 사스(SARS), 중동호흡기증후군(MERS), 에볼라(Ebola) 바이러스 중 하나 또는 둘 이상이 포함될 수 있다.

[0036] 또한, 상기 특징점 추출 단계(S120)에서는, 상기 게놈 서열 데이터에서 뉴클레오티드(nucleotide)의 빈도(frequency)를 특징점으로 추출할 수 있다.

[0037] 또한, 상기 특징점 추출 단계(S120)에서는, 상기 게놈 서열 데이터에서 트리-뉴클레오티드(tri-nucleotides)의 빈도(frequency)를 특징점으로 추출할 수 있다.

[0038] 또한, 상기 특징점 추출 단계(S120)에서는, 상기 게놈 서열 데이터에서 아미노산의 조성(composition of amino acids)을 특징점으로 추출할 수 있다.

[0039] 또한, 상기 특징점 추출 단계(S120)에서는, 상기 게놈 서열 데이터에서 뉴클레오티드 트리플렛의 서열(sequence of nucleotide triplets)을 특징점으로 추출할 수 있다.

[0040] 또한, 상기 특징점 추출 단계(S120)에서는, 상기 게놈 서열 데이터에서 코딩 서열(coding sequence, CDS)을 특징점으로 추출할 수 있다.

[0041] 또한, 상기 특징점 추출 단계(S120)에서는, 상기 게놈 서열 데이터에서 바이러스 종류간 정렬 유사도(alignment similarity)를 특징점으로 추출할 수 있다.

[0042] 또한, 상기 특징점 추출 단계(S120)에서는, 상기 게놈 서열 데이터에서 바이러스 종류간 게놈 서열의 차이를 보여주는 도트 플롯(Dot Plot) 이미지를 특징점으로 추출할 수 있다.

[0043] 또한, 상기 바이러스 분류 단계(S130)에서, 상기 분류기는, 훈련 데이터를 이용하여 뉴클레오티드(nucleotide)의 빈도(frequency), 트리-뉴클레오티드(tri-nucleotides)의 빈도(frequency), 아미노산의 조성(composition of amino acids), 뉴클레오티드 트리플렛의 서열(sequence of nucleotide triplets), 코딩 서열(coding sequence, CDS), 바이러스 종류간 정렬 유사도(alignment similarity), 바이러스 종류간 게놈 서열의 차이를 보여주는 도트 플롯(Dot Plot) 이미지 중 하나 이상을 이용하여 학습된 지원 벡터 머신(Support Vector Machine)을 기반으로 구성될 수 있다.

[0044] 이에 따라, 본 발명의 일 실시예에 따른 바이러스 분류 방법에서는, COVID-19 및 유사 바이러스에 대한 게놈 서열 분석을 기반으로 인공 지능을 이용하여 COVID-19 및 유사 바이러스를 분류할 수 있게 된다.

[0045] 이하, 본 발명의 일 실시예에 따른 인공 지능을 이용한 게놈 서열 분석 기반 COVID-19 및 유사 바이러스의 분류 방법, 장치 및 시스템에 대한 예시적인 실시 형태들을 첨부된 도면을 참조하여 보다 자세하게 설명한다.

[0046] 이때, 본 발명에서는 COVID-19(코로나 바이러스)와 사스(SARS), 중동호흡기증후군(MERS), 에볼라(Ebola)와 같은 유사 바이러스의 유전체 염기서열 분석을 수행하는 인공 지능을 이용한 게놈 서열 분석 기반 바이러스 분류 방법, 장치 및 시스템을 개시한다.

[0047] 본 발명의 일 실시예에 따른 인공 지능을 이용한 게놈 서열 분석 기반 바이러스 분류 방법, 장치 및 시스템은 다양한 바이러스의 게놈 서열에서 중요한 정보를 얻는 데 도움이 될 수 있다. 이를 위하여, 본 발명에서는 코로나-19 및 기타 유전체 염기서열에 대하여 염기 조성 및 빈도, 트리뉴클레오티드(tri-nucleotide) 조성, 아미노산 개수, 염기 서열 간의 정렬, DNA 유사성 정보 등의 기본 정보를 추출하여 데이터에 대한 비교 분석을 수행하고, 이를 기반으로 다양한 시각화 방법을 적용하여 바이러스의 게놈 시퀀스를 분석하며, 나아가 머신 러닝 기반 분류기 지원 벡터 머신(SVM)을 적용하여 상이한 게놈 시퀀스를 분류하게 된다.

[0048] 이를 위하여, 본 발명에서는 각 바이러스의 게놈 서열에 대한 데이터 세트는 공개 액세스가 가능한 온라인 데이터 센터 저장소에서 수집하였고, 이를 통해 본 발명의 일 실시예에 따른 인공 지능을 이용한 게놈 서열 분석 기반 바이러스 분류 방법, 장치 및 시스템에서는 COVID-19의 경우 97%, SARS의 경우 96%, MERS 및 에볼라의 게놈 서열에 대해 각각 95%의 높은 정확도로 분류 결과를 도출할 수 있었다.

[0049] 보다 구체적으로, 본 발명의 일 실시예에 따른 인공 지능을 이용한 게놈 서열 분석 기반 바이러스 분류 방법, 장치 및 시스템에서는 유전체 염기 서열 분석 에서 개선된 인공 지능 기술에 기반하여 SARS, MERS, 에볼라 등 코로나19와 유사 바이러스의 유전체 염기서열 분석을 위한 시스템을 제공하는 것을 목적으로 한다. 이를 위하여, 본 발명에서는 이러한 바이러스의 게놈 염기서열의 기본 패턴을 연구하기 위해 비교 분석을 수행하고 분류를 위해 기계 학습 알고리즘을 추가로 활용하였다. 이와 관련하여, 본 발명의 목적을 달성하기 위한 기술적 기여로서 다음과 같은 내용들이 포함될 수 있다.

[0050] i) 본 발명에서는 코로나-19, 사스, 메르스, 에볼라의 유전자 염기서열 분석을 위한 인공지능 기반 시스템을 제시하며, ii) 게놈 시퀀스에서 흥미로운 패턴을 찾기 위해 다양한 유형의 데이터 분석 및 시각화 기술을 사용하여 비교 분석을 수행할 수 있고, iii) 바이러스 분류를 위해 기계 학습 분류기(SVM)를 적용하여 서로 다른 게놈 시퀀스를 분류하고, iv) SVM 분류기의 결과를 다른 기계 학습 알고리즘과 비교하여 검증하였다.

[0051] 보다 구체적으로, 본 발명에서는 코로나-19, 사스, 메르스, 에볼라의 유전체 염기서열 분석 및 분류를 위한 인공지능 기반 시스템을 제시한다. 이와 관련하여, 도 3에서는 본 발명의 일 실시예에 따른 바이러스 분석 방법 및 시스템의 전반적인 동작 프로세스를 도시하고 있다.

[0052] 이때, 본 발명에서는 다양한 유형의 해석 및 시각화 기술을 사용하여 비교 데이터 분석을 수행하여, DNA의 뉴클레오타이드 빈도, 트리뉴클레오타이드 조성, 가장 안정적인 DNA를 갖는 게놈 서열을 나타내는 GC 백분율, 아미노산 수, 상이한 게놈 서열 간의 유사성 또는 정렬 등 이러한 바이러스의 게놈 내부 세부 정보를 도출하며, 게놈 서열의 길이를 포함하여 뉴클레오타이드에 대한 시각화를 수행하였다.

[0053] 또한, 본 발명에서는 다양한 유형의 그래프를 사용하여 게놈 서열 정보에 대한 세부 정보를 제공했으며, 나아가

다른 게놈 서열을 분류하기 위해 기계 학습 기반 분류기를 적용하여 바이러스에 대한 분류를 수행하였다.

[0054] 이를 위하여, 먼저 GenBank의 .fasta 및 .gb 파일 확장자에서 서로 다른 바이러스의 게놈 서열을 수집하였고, 이때 4가지 유형의 바이러스 모두에 대해 총 300개의 서로 다른 게놈 코딩 서열을 사용하였다. 여기서 GenBank는 생물학적 주석을 추가로 촉진하는 뉴클레오티드 서열의 친숙한 온라인 오픈 액세스 데이터베이스 중 하나로서, NCBI(National Center for Biotechnology Information)에서 지원하며, GenBank는 약 18개월마다 시퀀스 파일이 증가하면서 지난 20년 동안 기하급수적으로 발전하면서, 전 세계의 연구 분석가가 게놈 서열의 특정 바이러스 구조와 기능을 즉시 평가할 수 있도록 지원하고 있다. 이때, 온라인 데이터베이스에서 수집한 바이러스에 대한 게놈 염기서열 데이터는 백신, 항바이러스 약물 및 진단 테스트 등을 수행하기 위해 필수적인 데이터라고 할 수 있다.

[0055] 또한, 본 발명의 일 실시예로서, 게놈 서열의 비교 분석을 위해 Biopython을 사용할 수 있다. 게놈 서열 분석은 질병 발생을 연구하는 데 필수적인 도구가 되는 경우가 많으며, 본 발명에서는 코로나-19와 SARS, MERS, 에볼라 게놈을 포함한 기타 바이러스가 사용될 수 있다.

[0056] 보다 구체적으로, 본 발명에서는 DNA 서열을 읽어 들이는 것으로부터 분석을 시작할 수 있다. 본 발명에서는 이를 통해 염기 정보 또는 게놈 서열의 길이를 추출했는데, 예를 들어, COVID-19 게놈 서열의 길이는 29,903개, SARS 게놈 서열은 2975개, MERS 게놈 서열은 30,119개, 에볼라 게놈 서열 길이는 18,959개로 추출되었다. 본 발명의 일 실시예로서, 각 유전체 염기서열 DNA의 염기를 시각화한 결과는 도 4에 도시되어 있다. DNA는 세포의 핵(여기서 핵DNA라고 함)에서 발견되는 유기체의 유전 물질로서, DNA 정보는 도 4와 같이 Adenine-A, Guanine-G, Cytosine-C, Thymine-T의 4가지 화학 염기로 구성된 코드로 저장될 수 있다.

[0057] 이때, DNA의 뉴클레오티드는 다른 색상으로 표시되었는데, 예를 들어 게놈 서열의 처음 300개 뉴클레오티드를 표시했으며, 도 4에서 볼 수 있는 바와 같이, 뉴클레오티드의 분포가 다양하다는 것을 확인할 수 있다.

[0058] 도 5에서는 서로 다른 게놈 서열의 DNA에 있는 각 뉴클레오티드의 구성을 보여주고 있다. 뉴클레오타이드 A와 T의 빈도가 C와 G의 빈도보다 높은 것을 알 수 있으며, 이때 뉴클레오타이드 쌍으로 인해 COVID-19, SARS, MERS에서 가장 많이 존재하는 뉴클레오타이드는 T인 반면, 에볼라에 존재하는 뉴클레오티드는 A라는 것을 확인할 수 있다. A와 T는 각 게놈 서열에서 가장 많이 존재하는 뉴클레오티드로 확인되는 반면, 다른 뉴클레오티드의 분포는 달라질 수 있다. 또한, 본 발명에서는 COVID-19와 SARS에서 Cytosine-Gua-9(GC)와 Adenine-Thymine(AT)의 양 사이에 큰 차이가 있음을 확인하였다. 이에 따라, 뉴클레오티드 빈도를 사용하여 각 게놈 서열에 대한 GC%를 결정할 수 있으며, 이는 DNA의 안정성에 관한 중요한 정보를 제공할 수 있다. 또한, GC의 양은 AT와 거의 비슷하므로 에볼라 바이러스 게놈 서열의 GC 백분율이 가장 높은 GC 백분율을 갖는 것으로 추정되는 45.50%, 즉 가장 안정적인 DNA를 가지고 있다고 가정할 수 있다. 다음으로, 사스(SARS)와 메르스(MERS) 바이러스 염기서열에서는 각각 40.76%, 41.76%로 GC 비율이 가장 낮고, DNA가 가장 불안정한 염기서열은 코로나-19(37.97%)로 파악되며, 상기 정보는 DNA의 안정성이 변화에 저항하는 데 필수적이기 때문에 중요한 의미를 가질 수 있다.

[0059] 다음으로, 도 6a 내지 도 6d에서는 tri-nucleotide 조성을 도시하고 있으며, 생물정보학에서는 생물학적 서열에 포함된 범위의 하위 서열에 해당한다. 상기 tri-nucleotide 조성은 주로 컴퓨터 유전체학 및 염기서열 분석(즉, A, T, G, C)을 포함하는 염기서열 분석에 적용되며, 이종 유전자 표현을 증가시키고 메타게놈 샘플에서 종을 구별하며 약독화 백신을 생산하는 데 도움이 될 수 있다.

[0060] 또한, 게놈 서열에는 바이러스에 대한 모든 암호화된 기본 정보가 포함되어 있으며, 유전 정보를 이해하는 것은 치료법과 백신을 얻는 열쇠가 될 수 있다. 이때, 유전자 발현의 과정에서 유전자의 데이터는 작동하는 유전자 산물을 합성하는 데 사용되고, 합성되는 산물은 종종 단백질이 된다. 따라서, 본 발명에서는 DNA가 전령 RNA(mRNA)로 복사되는 전사를 수행하고 번역을 사용하여 mRNA가 아미노산으로 번역되며, 이는 한 코드(뉴클레오티드 A T C G 서열)에서 다른 코드(아미노산 서열)로의 번역에 해당하게 된다. 이때, 모든 아미노산 서열이 단백질은 아니고, 20개 이상의 아미노산 코드를 갖는 서열만이 기능성 단백질에 해당할 수 있다.

[0061] 이에 대하여, 도 7에서는 본 발명의 일 실시예에 따라 시각화된 각 아미노산의 빈도를 예시하고 있으며, 아미노산의 분포는 게놈 서열에 따라 다양함을 알 수 있다. 또한, 도 7에서는 모든 서열에서 류신(L)의 조성이 높은 반면, 최소 아미노산의 조성은 다양함을 보여주며, 두 가지 우세한 아미노산은 류신(L)과 세린(S)이고, 이때 도 7에서 y축의 단위 개수는 1000으로 작성되었다.

[0062] 또한, 도 8a 내지 도 8d에서는 COVID-19, SARS, MERS 및 에볼라 게놈의 오픈 리딩 프레임(ORF)을 예시하고 있다. 이때, DNA를 아미노산으로 변환하여 단백질 서열을 도출할 수 있게 되고, 이에 대하여 번역될 수 있는 리

딩 프레임(reading frame)의 일부인 오픈 리딩 프레임(open reading frame, ORF)-즉, 아미노산으로 바뀔 때 노-스톱 코돈(no-stop codons)을 운반하는 DNA 분자의 일부-을 사용하였다. 여기서, 유전자 코드는 이중 가닥 DNA 분자가 역방향으로 3개, 순방향으로 3개와 같이 6개의 가능한 판독 프레임 중 하나로 해석될 수 있음을 결정하는 3개의 염기 파트너의 조합으로 DNA 서열을 보여줄 수 있다. 긴 리딩 프레임(long reading frame)은 유전자의 허용 가능한 구성 요소로서, 일반적으로 기점 코돈(보통 AUG)으로 시작하여 종결 코돈(보통 UAA, UAG 또는 UGA)에서 종결될 수 있으며, 본 발명에서는 (Genbank 형식)을 사용하여 전사 및 게놈 서열로의 번역을 사용하였다. 도 8a 내지 도 8d에서는 게놈 시퀀스의 ORF와 x축에서 GC% 함량을 보여주는 그래프를 도시하고 있다. 여기서, 코딩 영역이 게놈 서열에서 어디에 있는지 이해하는 것이 중요하며, 이때 COVID-19와 SARS는 주로 ORF1ab, ORF3a, E, M, S, N에서 더 유사함을 알 수 있으며, MERS와 에볼라의 ORF는 COVID-19와 약간 다르다는 점을 확인할 수 있다.

[0063] 또한, CDS(Coding Sequence)로도 인식되는 유전자 코딩 영역은 단백질을 코딩하는 유전자 DNA 또는 RNA의 일부로서, 게놈 서열에서 코딩 영역 CDS를 찾는 것은 유전자의 기능적 주석을 위한 필수 단계에 해당하며. 본 발명의 일 실시예로서 도 9a 내지 도 9d에서 볼 수 있는 바와 같이 CDS 그래프를 도시하였다. 이때, CDS는 단백질의 아미노산 서열과 유사한 뉴클레오티드 서열로도 알려져 있으며, 일반적인 CDS는 ATG로 시작하여 정지 코돈으로 종결하게 된다. 도 9a 내지 도 9d의 코드는 코딩 영역 CDS를 빨간색으로 강조해 표시하고 있으며, 주요 CDS는 ORF1ab, ORF3a, S 단백질, M 단백질 및 N 단백질을 포함하는 그림 6에서 이미 발견된 ORF 중 하나이고, 코로나19와 사스의 DNA 구조는 거의 동일한 반면 메르스와 에볼라의 DNA 구조는 약간 다르다는 것을 확인할 수 있다.

[0064] 또한, 도 10에서는 COVID-19와 다른 게놈 서열의 정렬 유사성을 도시하고 있다. 이때, 4가지 유형의 DNA 염기서열 간의 유사성을 분석하기 위해 염기서열 정렬 방법을 사용하였으며, 서열 정렬은 둘 이상의 서열(DNA, RNA 또는 단백질 서열)을 특정 순서로 제공하여 이들 사이의 유사 영역을 인식하는데 도움이 될 수 있다. 여기서, 유사한 영역을 인식하면 종 사이에 어떤 특징이 보존되는지, 여러 종들이 유전적으로 어느 정도 가깝고 어떻게 종이 성장하는지 등과 같은 정보를 이해할 수 있으며, 페어와이즈(pairwise) 시퀀스 정렬은 두 개의 시퀀스만을 연관시키고 가장 신뢰할 수 있는 실행 가능한 시퀀스 정렬을 제공하여, 페어와이즈(pairwise) 시퀀스 정렬은 결과 시퀀스 정렬로부터 해석하고 판단하기 위한 쉽고 좋은 방법에 해당한다.

[0065] 또한, 본 발명에서는 두 개의 생물학적 유전체 염기서열 사이의 상관관계를 제공하고 이들 사이에서 가장 유사한 영역을 인식하기 위해 도 11과 같은 도트 플롯(Dot Plot)을 사용하였다. 이때, 가장 쉬운 방법은 염기서열이 동일한 곳에 점을 배치하는 것이고, x축에서 하나의 시퀀스를 조정하고 y축에서 다른 시퀀스를 조정하여 두 시퀀스를 비교하여, 두 시퀀스의 초과가 동시에 플롯과 유사하면 해당 위치에 점이 표시될 수 있다. 이는 시퀀스의 역 또는 직접 반복에 대해 시퀀스를 시각적으로 검사하는 데에도 유용하게 적용할 수 있으며, 또한 염기서열 복잡도가 낮은 영역, 유사한 영역 또는 영역, 복제된 염기서열, 게놈 염기서열의 재배열, RNA 구조 및 유전자 순서를 조사하는 데에도 사용될 수 있다.

[0066] 나아가, 본 발명의 일 실시예로서, 다른 게놈 시퀀스의 분류를 위해 앞서 도 3(b)에서 예시한 바와 같이 기계 학습 분류기를 학습할 수 있다. 종래에도 분류 목적으로 다양한 기계 학습 알고리즘을 활용하는 것이 가능하였으나, 본 발명에서는 회귀 및 분류 문제에 사용될 수 있는 효율적인 지도 머신 러닝 분류기인 SVM을 사용하는 것이 가능하다. 보다 구체적으로, 본 발명에서는 4가지 유형의 바이러스에 대한 다중 게놈 시퀀스를 사용하여 일부 전처리를 수행하며 클래스 레이블을 할당하고 SVM 분류기에 제공되는 특징점(feature)를 추출하게 된다. 또한, 본 발명에서는 수집된 게놈 코딩 시퀀스를 각각 80%와 20%의 비율로 훈련 및 테스트 샘플로 무작위로 분할하여 적용하였다. 이에 따라 본 발명의 일 실시예에 따른 분류기는 출력에서 도 3(b)와 같이 서로 다른 바이러스의 게놈 시퀀스를 분류하게 된다. 이때, SVM 분류기는 시퀀스 데이터를 4개의 클래스로 가장 정확하게 분리하는 특징 공간의 하이퍼-평면(hyper-plane)을 결정하게 되며, 데이터 포인트가 하이퍼-평면에서 멀어질수록 올바르게 분류될 가능성이 높아지고, 여기서 하이퍼-평면에 가장 가까운 데이터 포인트가 서포트 벡터(support vectors) 가 될 수 있다. 이때, 상기 서포트 벡터가 제거되면 하이퍼-평면의 위치가 변경될 수 있고, 따라서 서포트 벡터는 데이터 세트의 중요한 요소로 간주될 수 있으며, 상기 하이퍼-평면의 양쪽과 서포트 벡터 사이의 거리를 마진이라고 하면, 새로운 데이터를 올바르게 분류하기 위해 훈련 세트의 각 지점과 하이퍼-평면 사이의 마진이 가장 큰 하이퍼-평면을 선택하는 것이 중요한 목표가 될 수 있다. 나아가, SVM은 바이너리(binary)이므로 다중 클래스 문제는 여러 바이너리 분류 문제로 축소되어야 하며, 본 발명의 일 실시예로서 선형 SVM을 사용할 수 있다.

[0067] 보다 구체적으로, 본 발명에서 SVM 분류의 목적 함수는 아래 수학식 1과 같이 제공될 수 있다.

[0068] [수학식 1]

$$L_{(w)} = \sum_{i=1} |\max(0, 1 - y_i[w^T x_i + b]) + \lambda||w^2||.$$

[0069]

[0070] 또한, 아래에서는 본 발명의 일 실시예로서, 유전체 염기서열 분류에 사용되는 인공지능 기반 시스템의 성능 검증 결과를 제공하도록 한다. 통상의 경우 분류 정확도를 적용하여 시스템 성능을 평가하게 되나 알고리즘의 성능을 정확하게 평가하는 것만으로는 충분하지 않으므로, 본 발명에서는 상기 설명된 시스템의 성능을 검증하기 위해 다른 평가 메트릭을 적용하였고, 보다 구체적으로 아래와 같이 정의된 True Positive(TP), True Negative(TN), False Positive(FP) 및 False Negative(FN)와 같은 다양한 분류 메트릭(classification metrics)을 사용하였다.

[0071] (1) 정확하게 예측된 양성 클래스에 대한 참 양성(TP).

[0072] (2) 부정확하게 예측된 양성 클래스에 대한 거짓 양성(FP).

[0073] (3) 정확하게 예측된 네거티브 클래스에 대한 TN(True Negative).

[0074] (4) 부정확하게 예측된 음성 클래스에 대한 거짓 음성(FN).

[0075] 또한, 상기 메트릭을 사용하여 아래와 같은 파라미터가 산출될 수 있다.

[0076] 먼저, 분류 정확도(Accuracy)는 전체 입력 샘플 수와 올바른 예측의 비율로서, 모든 클래스에 대해 동일한 수의 샘플이 있는 경우에는 정상 동작하게 된다. 분류 정확도는 아래 수학식 2와 같이 평가될 수 있다.

[0077] [수학식 2]

$$\text{Accuracy} = \frac{TP}{TP + TN + FP + FN}$$

[0078]

[0079] 또한, 정밀도(Precision)은 특정 클래스에 속할 것으로 예상되는 전체 샘플 중 중요한 샘플(True Positive)의 비율로부터 아래 수학식 3과 같이 결정될 수 있다.

[0080] [수학식 3]

$$\text{Precision} = \frac{TP}{TP + FP}$$

[0081]

[0082] 또한, 재현율(Recall)은 아래 수학식 4와 같이 클래스에 실제로 속하는 모든 샘플에 대하여 클래스에 속할 것으로 예측되는 샘플의 비율로 산출될 수 있다.

[0083] [수학식 4]

$$\text{Recall} = \frac{TP}{TP + FN}$$

[0084]

[0085] 또한, 민감도(Sensitivity)라고도 하는 True Positive Rate(TPR)는 양성으로 올바르게 분류된 양성 샘플의 비율로부터 계산될 수 있으며, 아래 수학식 5와 같이 정의될 수 있다.

[0086] [수학식 5]

$$TPR = \frac{TP}{TP + FN}$$

[0087]

[0088] 또한, True Negative rate(TNR)은 특이도(Specificity)라고도 하며, 음성으로 올바르게 분류된 음성 샘플의 비율로서, 아래 수학식 6과 같이 주어진다.

[0089] [수학식 6]

$$TNR = \frac{TN}{TN + FP}$$

[0090]

[0091] 또한, False positive rate(FPR)은 양성으로 잘못 분류된 음성 샘플의 수를 나타내며, 아래 수학식 7과 같이 정의될 수 있다.

[0092] [수학식 7]

$$FPR = \frac{FP}{FP + TN}$$

[0093]

[0094] 또한, F1-Score는 Recall과 Precision 사이의 조화 평균으로 정의될 수 있으며, 분류기가 얼마나 정확한지(얼마나 많은 샘플이 올바르게 분류되었는지)와 견고성을 주로 결정하게 된다. 높은 정밀도와 낮은 재현율 값을 가지는 경우 매우 정확하지만 분류하기 어려운 많은 샘플을 놓칠 수 있으며, F1 Score의 값이 높을수록 알고리즘의 성능이 더 우수할 수 있다. 통상적으로 재현율과 정밀도 사이의 균형을 유지하는 방향으로 구성하게 되며, 아래 수학식 8과 같이 산출될 수 있다.

[0095] [수학식 8]

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

[0096]

[0097] 또한, 도 12에서는 본 발명의 일 실시예에 따른 바이러스 분류 방법, 장치 및 시스템에서 산출된 정밀도, 재현율, 정확도 및 F1 점수를 도시하고 있다. 여기서, 본 발명에 따른 SVM 분류기가 COVID-19의 경우 97%, SARS의 경우 96%, 메르스, 에볼라의 경우 각각 95%의 우수한 분류 정확도 결과를 보여주는 것을 확인할 수 있다. 또한, Precision, Recall 및 F1-Score는 COVID-19의 경우 96%, 77%, 96%, SARS의 경우 96%, 74%, 96%, MERS의 경우 95%, 74%, 95% 및 에볼라의 경우 95%, 74% 95%의 결과를 보였다.

[0098] 이어서, 상기 평가 매개변수를 사용하여 도 13과 같이 ROC 곡선을 도시하였다. 도 13의 ROC 곡선은 정의된 컷오프 값에 대한 TPR 대 위양성 비율 FPR 값을 사용하여 플로팅될 수 있고, ROC 곡선이 높을수록(즉, SVM이 선 y = 1에 더 가까울수록) 적합도가 우수하다고 할 수 있는데, 모든 알고리즘에서 95%에 가깝거나 보나 높은 우수한 결과를 보여주나, 특히 SVM의 성능이 모든 알고리즘 중에서 가장 우수함을 알 수 있습니다. 이에 대응하는 다른 기계 학습 알고리즘의 비교 결과는 도 14에 도시되어 있다.

[0099] 또한, 도 15에서는 본 발명의 일 실시예에 따른 바이러스 분류 시스템(100)의 블록도를 도시하고 있다. 이때, 본 발명의 일 실시예에 따른 바이러스 분류 시스템(100)에 대해서는 앞서 설명한 본 발명의 일 실시예에 따른 바이러스 분류 방법에 대한 설명을 참조하여 통상의 기술자가 용이하게 구현 가능하므로, 아래에서는 자세한 설명은 생략하고 본 발명의 일 실시예에 따른 바이러스 분류 시스템(100)의 주요 구성에 대해서 간략하게 살핀다.

[0100] 이때, 도 15에서 볼 수 있는 바와 같이, 본 발명의 일 실시예에 따른 바이러스 분류 시스템(100)은, 바이러스의 게놈 서열을 기반으로 바이러스의 종류를 분류하는 바이러스 분류 시스템으로서, 분류 대상 바이러스의 게놈 서열 데이터를 수집하는 게놈 서열 데이터 수집부(110), 상기 게놈 서열 데이터에서 특징점(feature)을 추출하는 특징점 추출부(120) 및 추출된 상기 특징점을 미리 학습된 분류기로 입력하여 상기 바이러스의 종류를 분류하는 바이러스 분류부(130)를 포함하여 구성될 수 있다.

[0101] 이때, 상기 바이러스에는 코로나(COVID-19), 사스(SARS), 중동호흡기증후군(MERS), 에볼라(Ebola) 바이러스 중 하나 또는 둘 이상이 포함될 수 있다.

[0102] 이에 따라, 본 발명의 일 실시예에 따른 인공 지능을 이용한 게놈 서열 분석 기반 COVID-19 및 유사 바이러스의 분류 방법, 장치 및 시스템에서는, COVID-19 및 유사 바이러스에 대한 게놈 서열 분석을 기반으로 인공 지능을 이용하여 COVID-19 및 유사 바이러스를 분류할 수 있게 된다.

[0103] 이상의 설명은 본 발명의 기술 사상을 예시적으로 설명한 것에 불과한 것으로서, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자라면 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 다양한 수정 및 변형이 가능할 것이다. 따라서 본 발명에 기재된 실시예들은 본 발명의 기술 사상을 한정하기 위한 것이 아니라 설명하기 위한 것이고, 이러한 실시예에 한정되는 것은 아니다. 본 발명의 보호 범위는 아래의 청구범위에 의해서 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 발명의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

[0104] 20, 20a, 20b : 게놈 서열 입력 장치

30 : 통신 네트워크

100 : 바이러스 분류 시스템

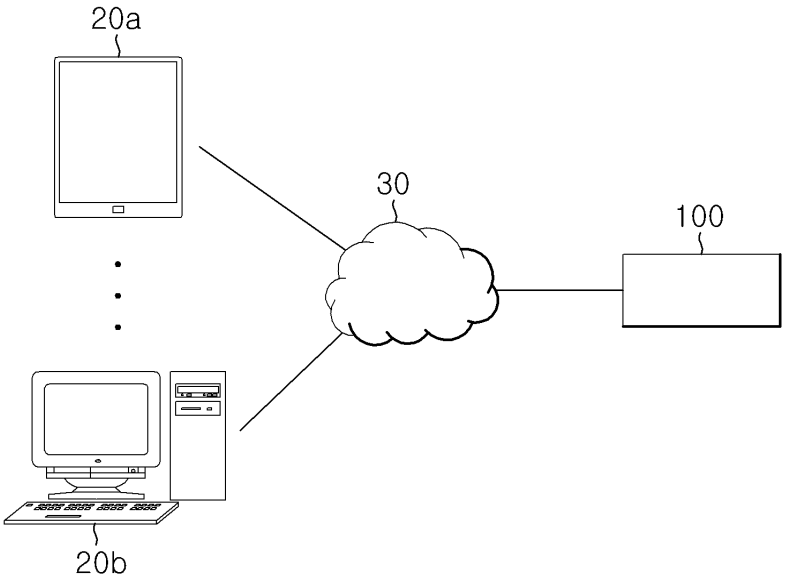
110 : 게놈 서열 데이터 수집부

120 : 특징점 추출부

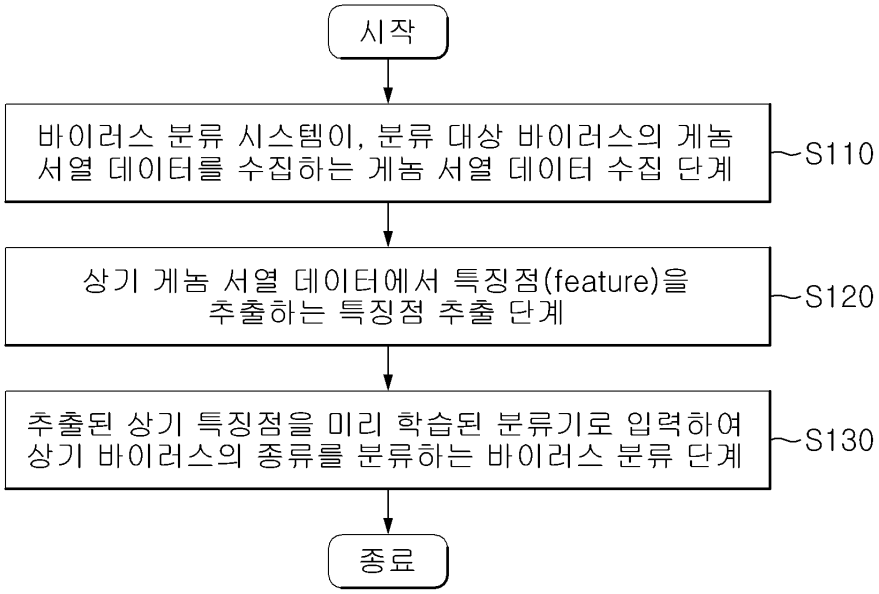
130 : 바이러스 분류부

도면

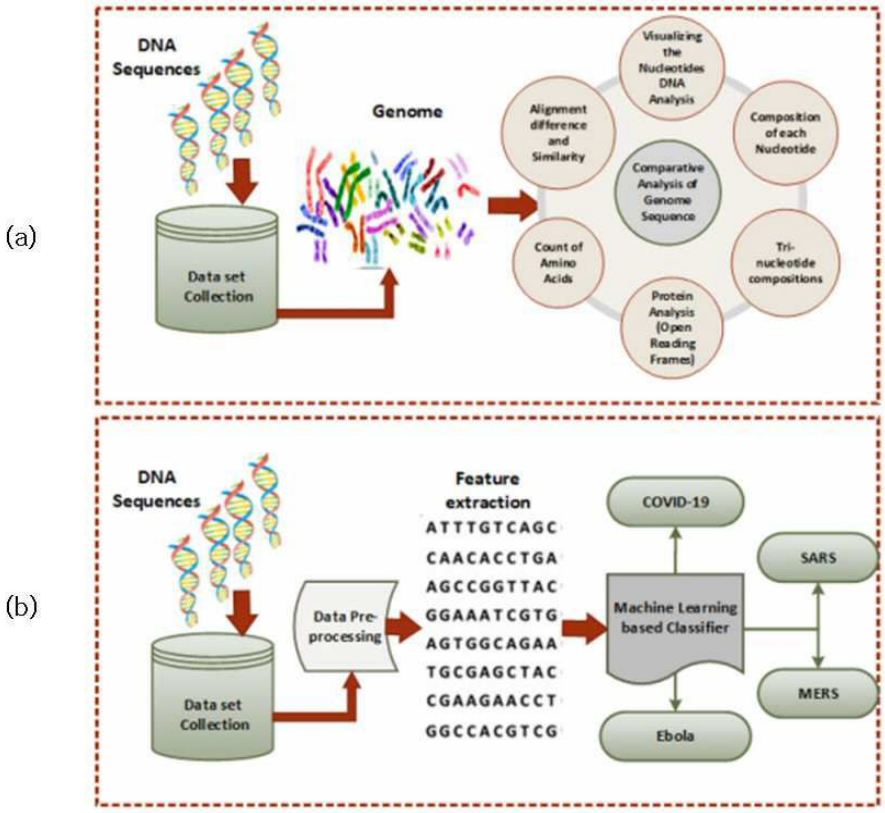
도면1


20a
30
100
20b

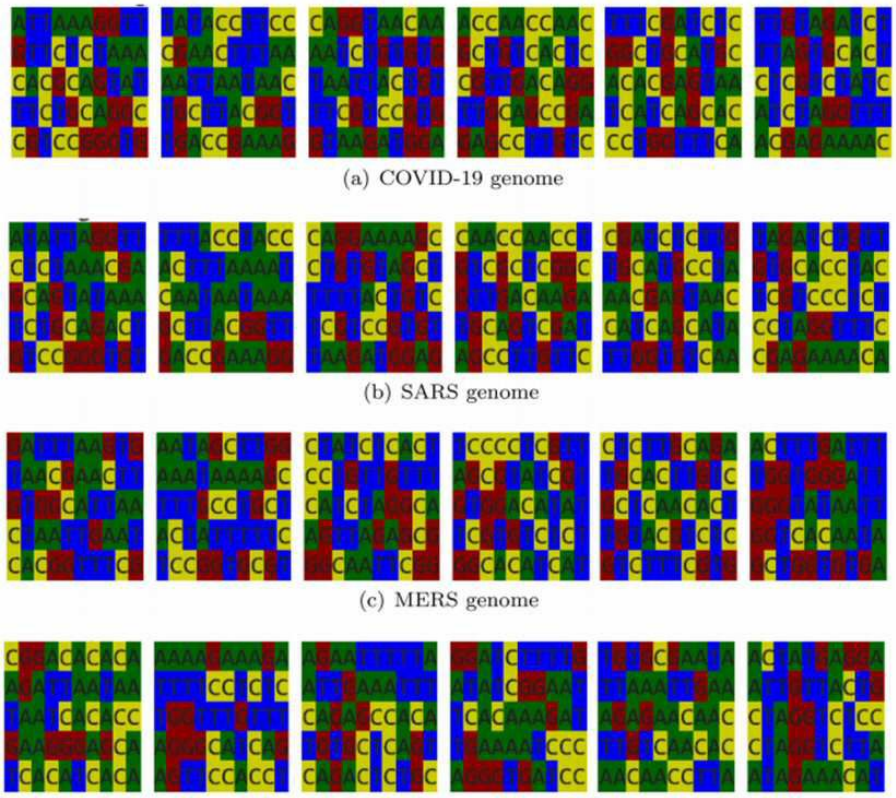
도면2


시작
바이러스 분류 시스템이, 분류 대상 바이러스의 게놈 서열 데이터를 수집하는 게놈 서열 데이터 수집 단계
S110
상기 게놈 서열 데이터에서 특징점(feature)을 추출하는 특징점 추출 단계
S120
추출된 상기 특징점을 미리 학습된 분류기로 입력하여 상기 바이러스의 종류를 분류하는 바이러스 분류 단계
S130
종료

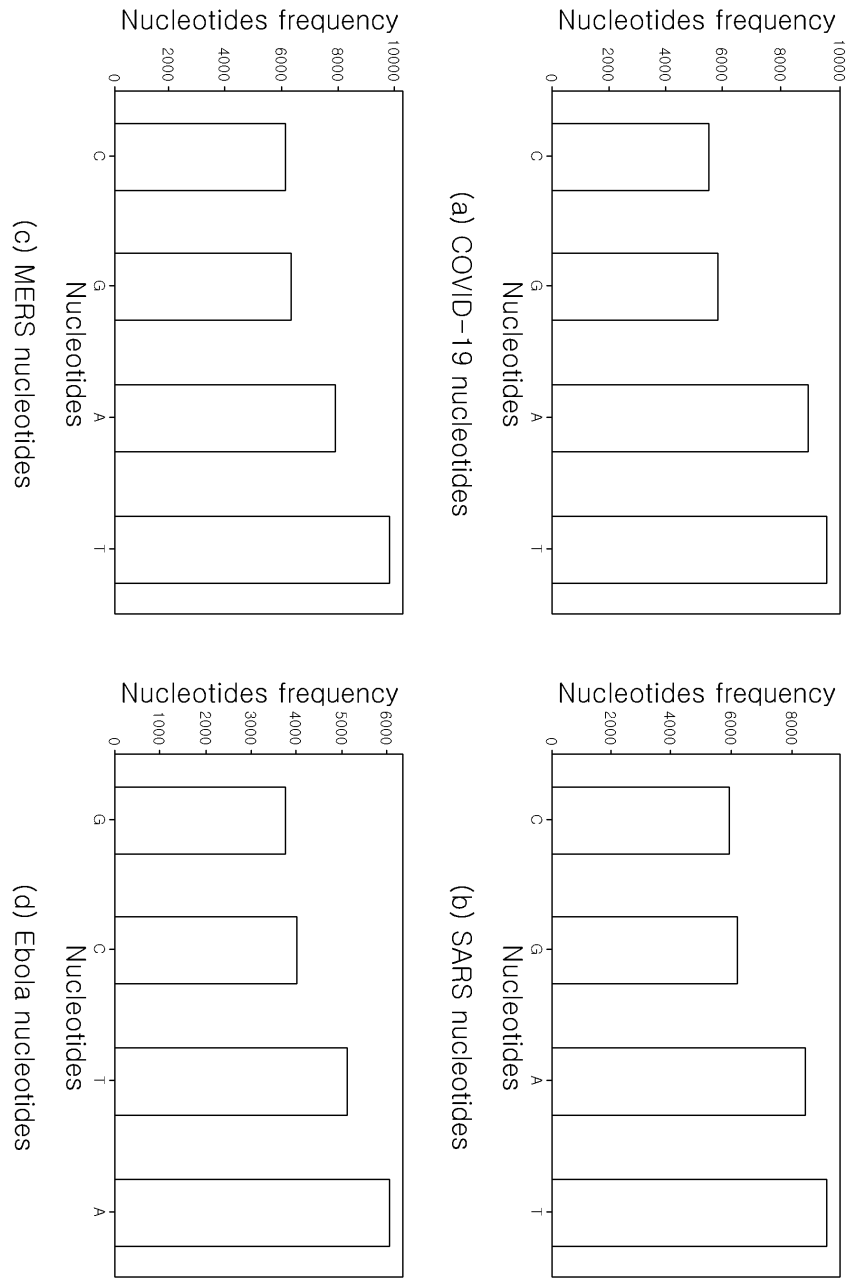
도면3


(a)
DNA Sequences
Genome
Data set Collection
Visualizing the Nucleotides DNA Analysis
Composition of each Nucleotide
Comparative Analysis of Genome Sequence
Tri-nucleotide compositions
Protein Analysis (Open Reading Frames)
Count of Amino Acids
Alignment difference and Similarity
(b)
DNA Sequences
Data set Collection
Data Pre-processing
Feature extraction
ATTTGTCAGC
CAACACCTGA
AGCCGGTTAC
GGAAATCGTG
AGTGGCAGAA
TGCGAGCTAC
CGAAGAACCT
GGCCACGTCG
Machine Learning based Classifier
COVID-19
SARS
MERS
Ebola

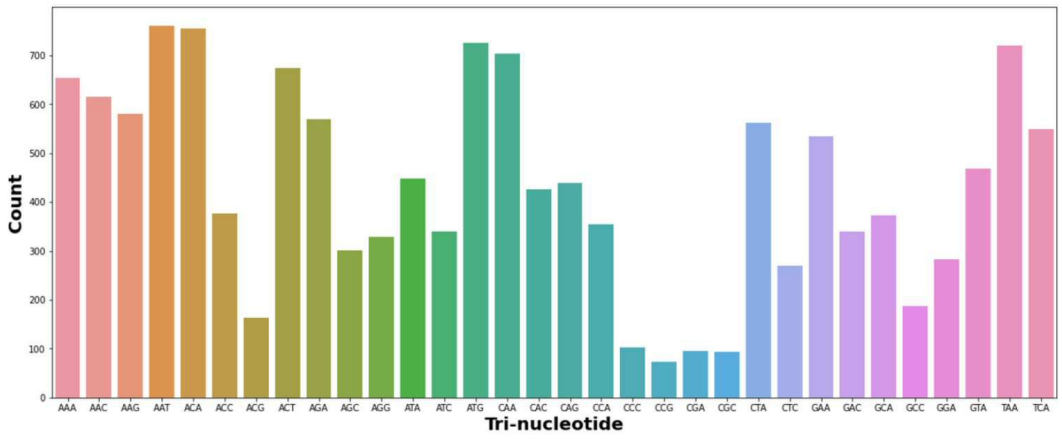
도면4


(a) COVID-19 genome
(b) SARS genome
(c) MERS genome

도면5

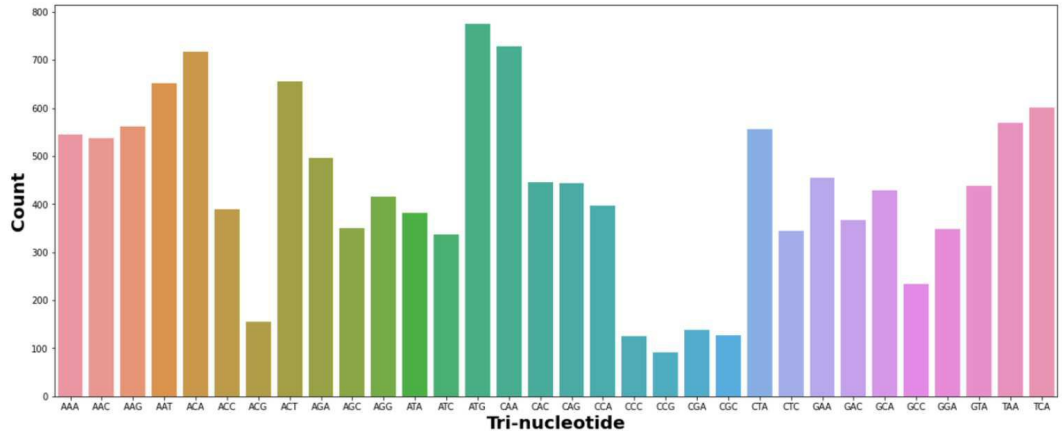

Nucleotides frequency
0
2000
4000
6000
8000
10000
C
G
A
T
Nucleotides
(a) COVID-19 nucleotides
(b) SARS nucleotides
(c) MERS nucleotides
(d) Ebola nucleotides

도면6a



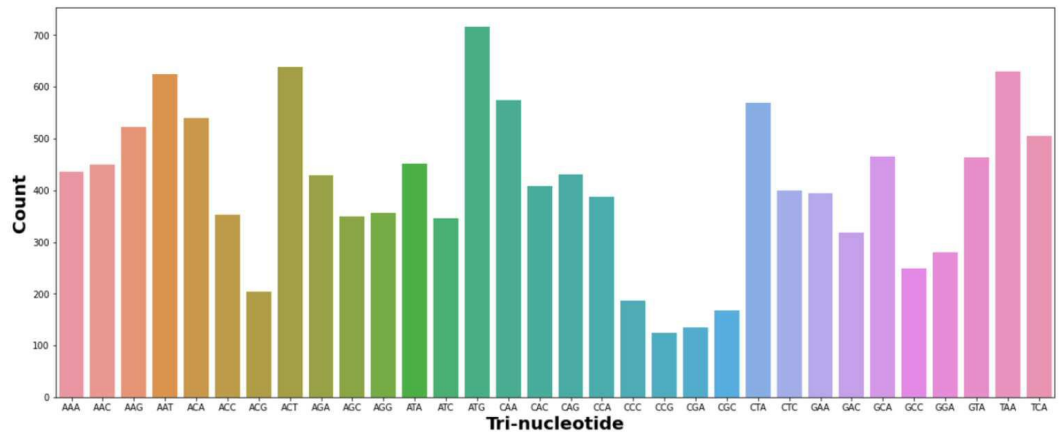
COVID-19 Tri-Nucleotides

도면6b



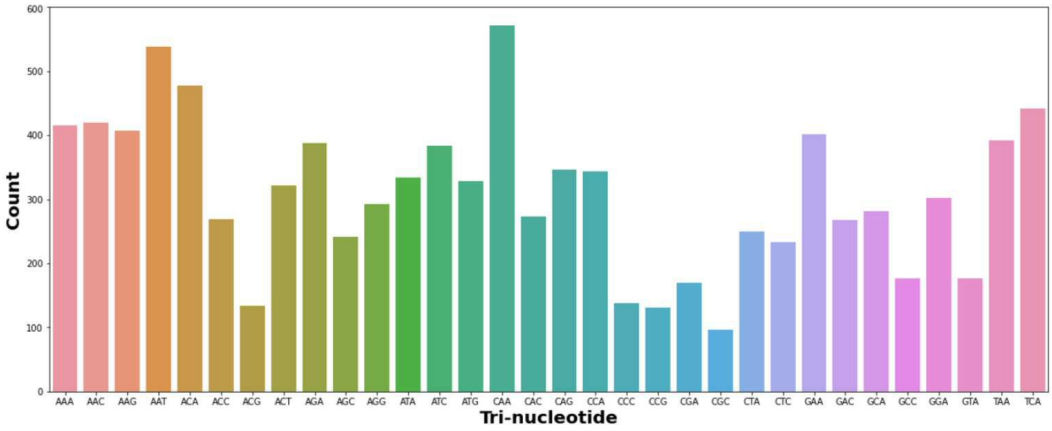
SARS Tri-Nucleotides

도면6c



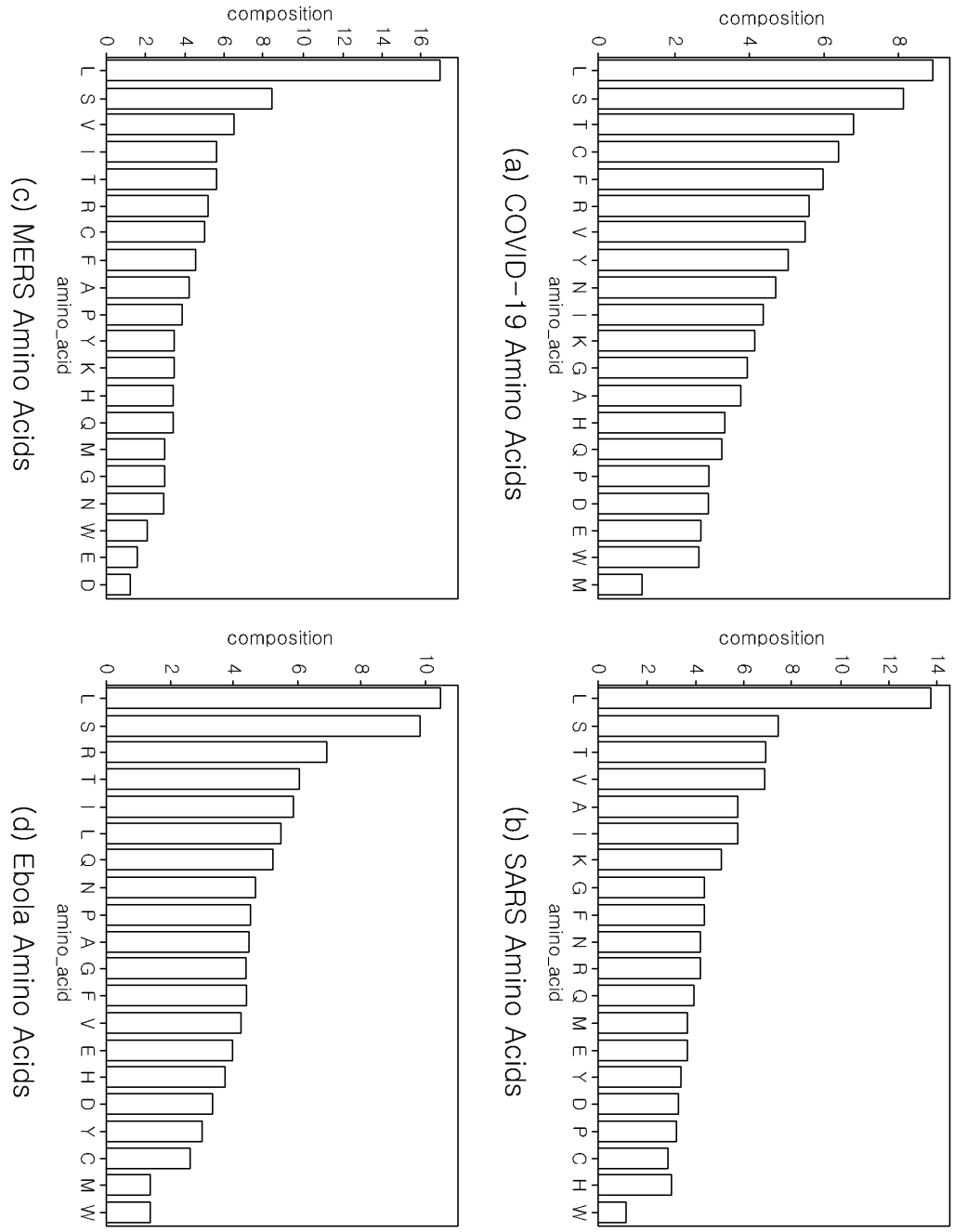
MERS Tri-Nucleotides

도면6d

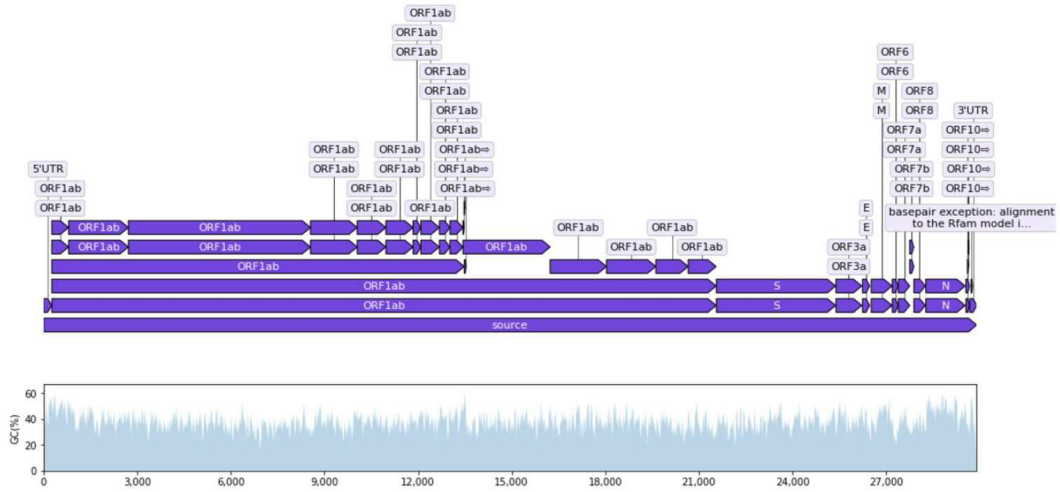


Ebola Tri-Nucleotides

도면7

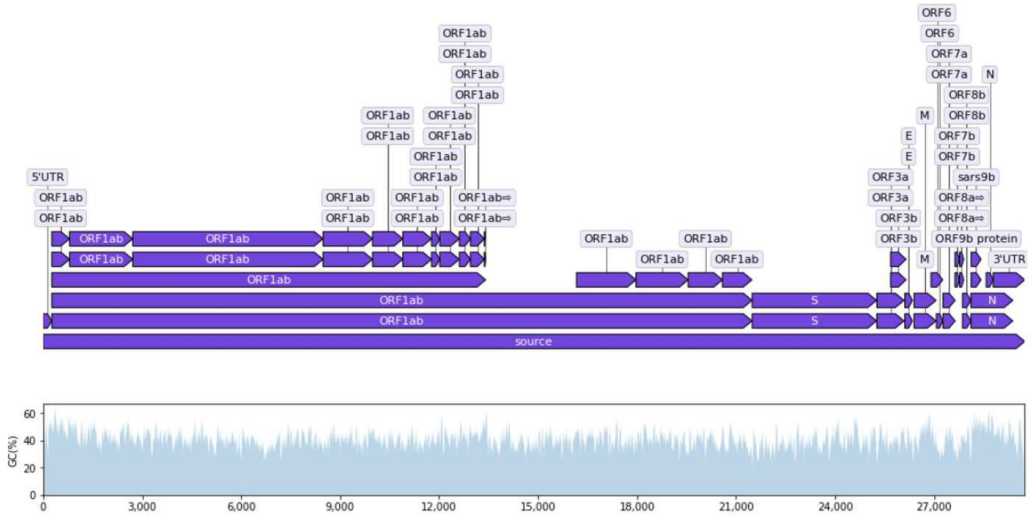

composition
amino_acid
L S T C F R V Y N I K G A H Q P D E W M
(a) COVID-19 Amino Acids
composition
amino_acid
L S T V A I K G F N R Q M E Y D P C H W
(b) SARS Amino Acids
composition
amino_acid
L S V I T R C F A P Y K H Q M G N W E D
(c) MERS Amino Acids
composition
amino_acid
L S R T I L Q N P A G F V E H D Y C M W
(d) Ebola Amino Acids

도면8a



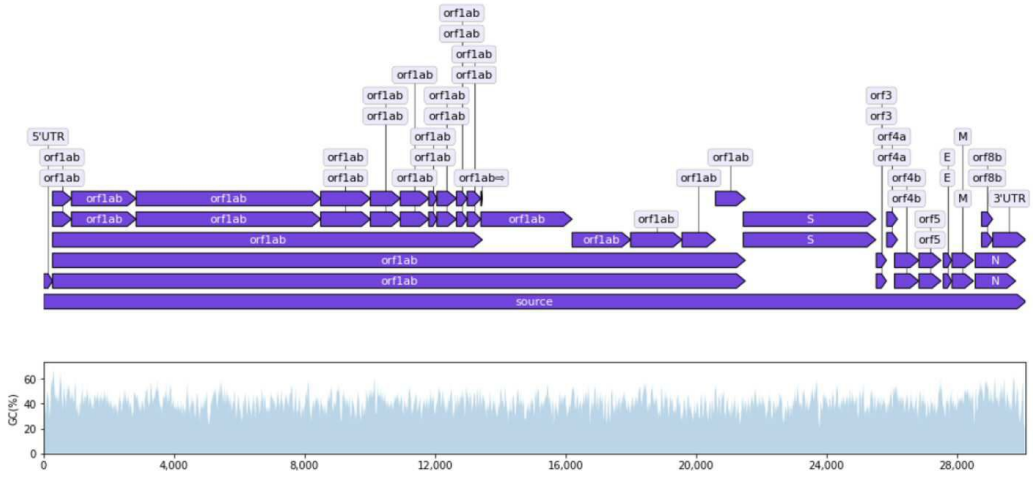
COVID-19 Open Reading Frames

도면8b



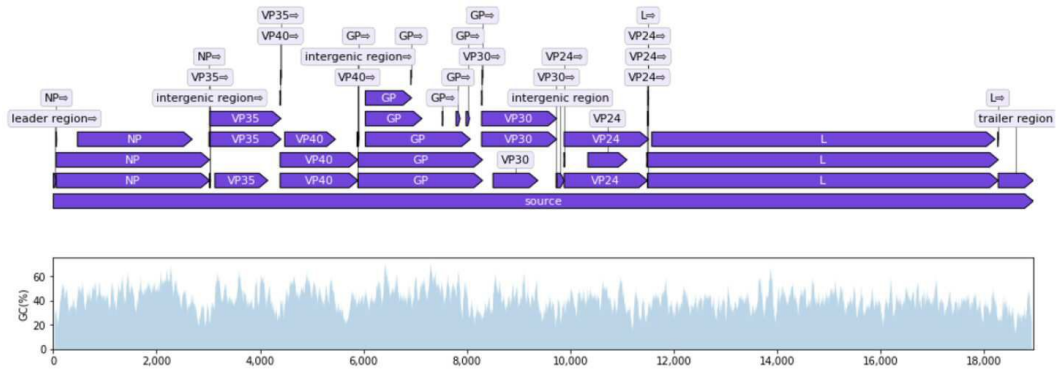
SARS Open Reading Frames

도면8c



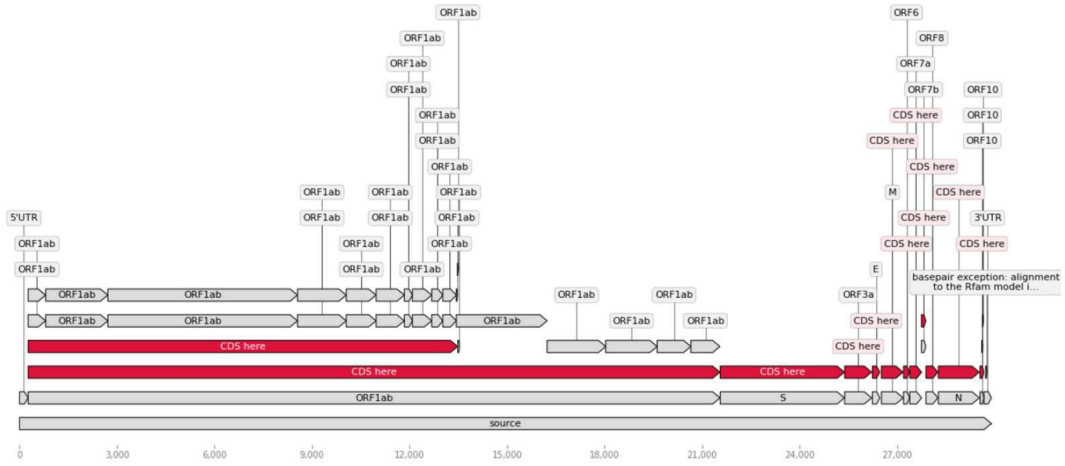
MERS Open Reading Frames

도면8d



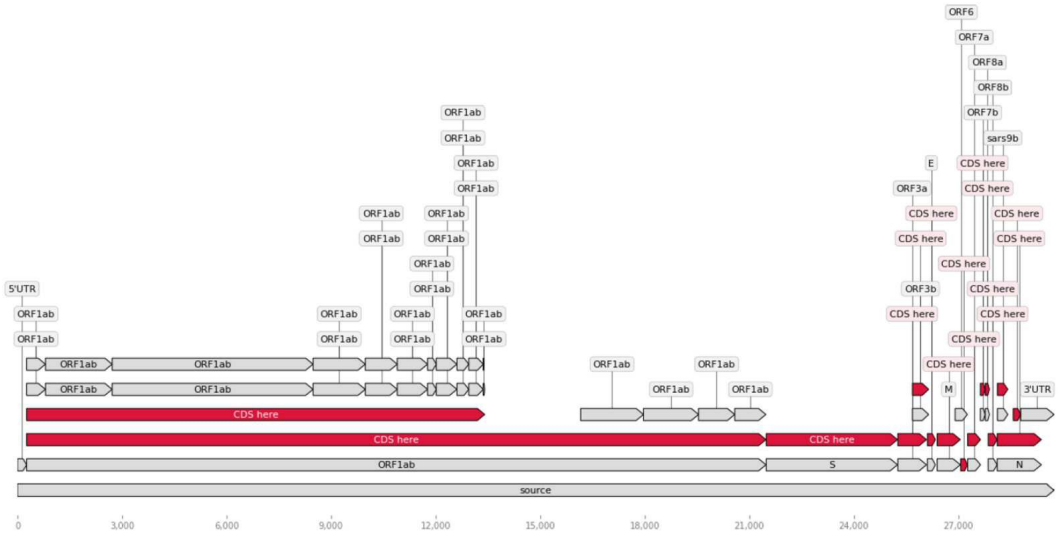
Ebola Open Reading Frames

도면9a



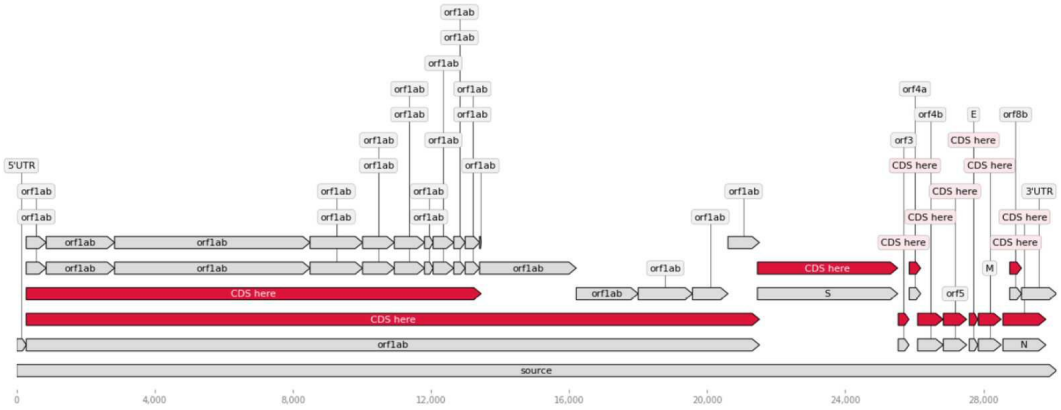
COVID-19 CDS

도면9b



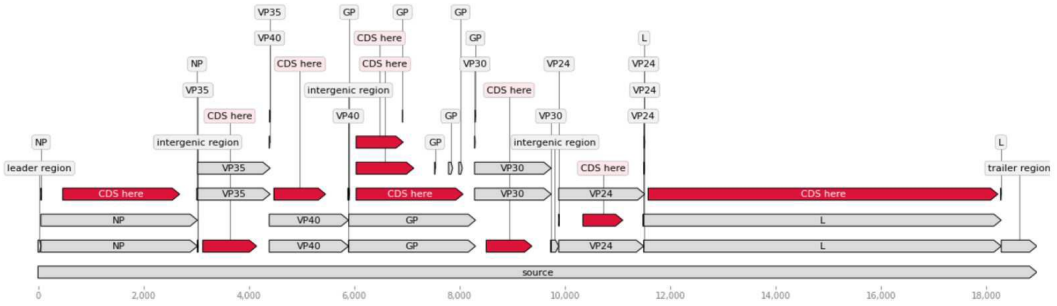
SARS CDS

도면9c



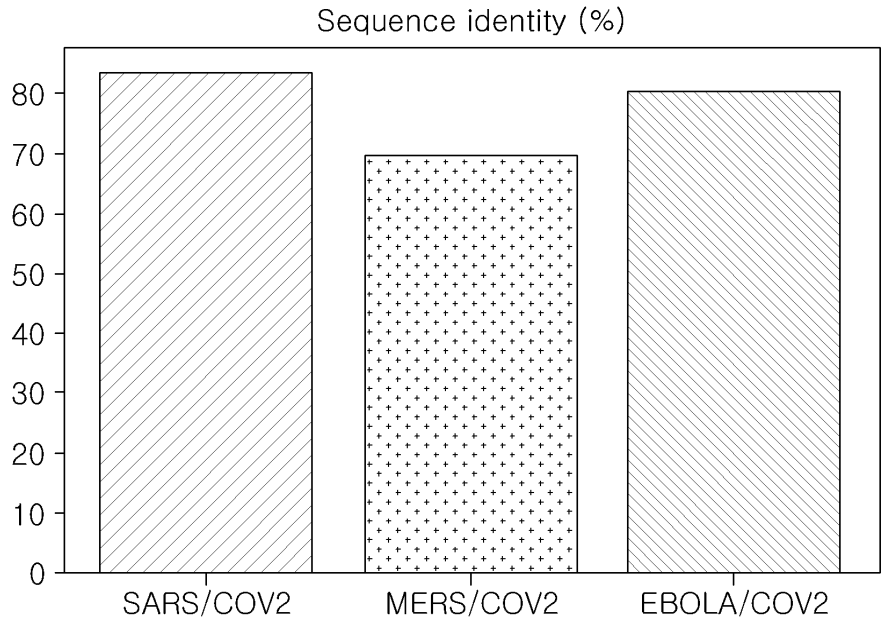
MERS CDS

도면9d

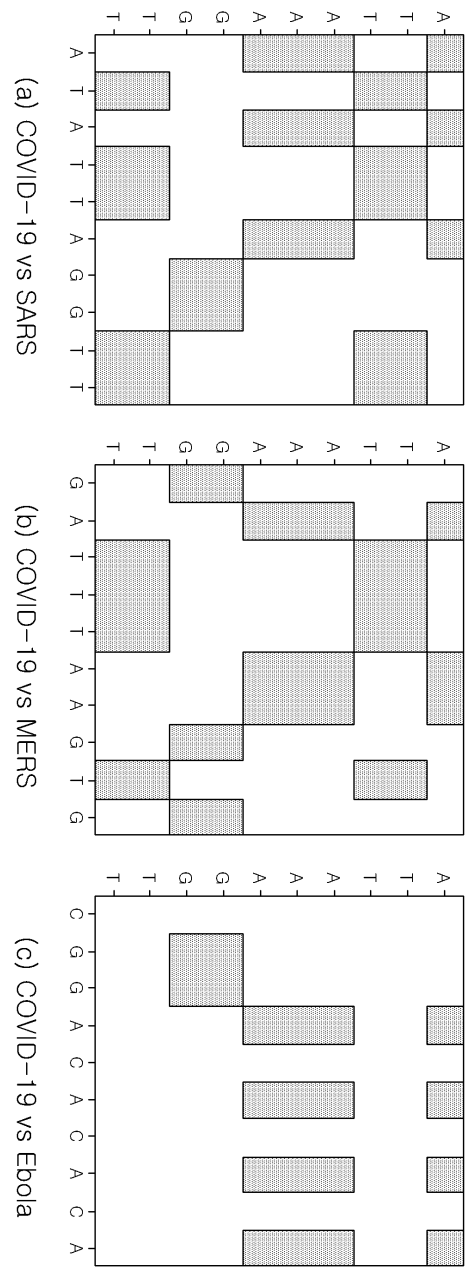


Ebola CDS

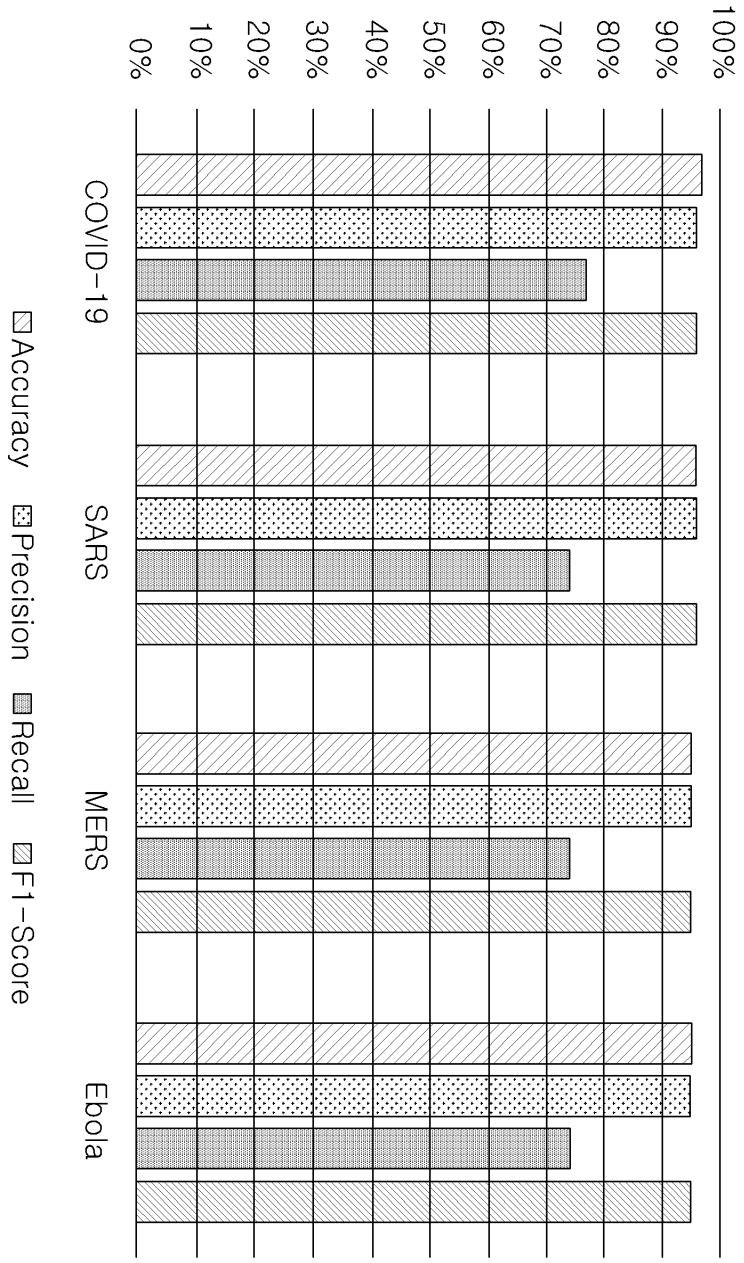
도면10


Sequence identity (%)
80
70
60
50
40
30
20
10
0
SARS/COV2
MERS/COV2
EBOLA/COV2

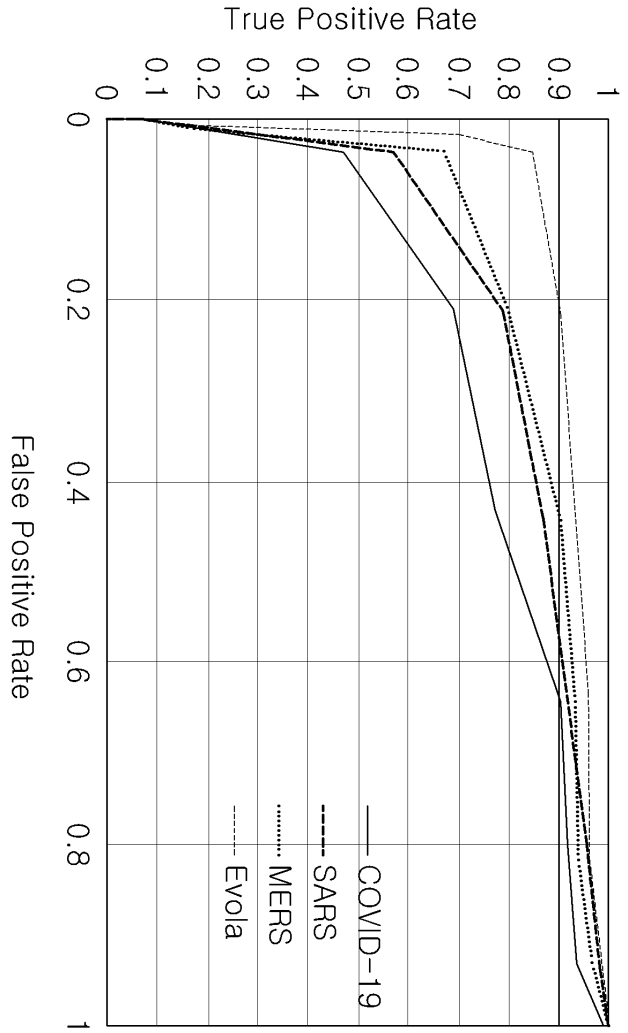
도면11



도면12


0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
COVID-19
SARS
MERS
Ebola
Accuracy
Precision
Recall
F1-Score

도면13


True Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
False Positive Rate
0
0.2
0.4
0.6
0.8
1
COVID-19
SARS
MERS
Evola

도면14

S.no	Algorithm	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)
1	Extreme gradient boost	92	77	94	92
2	Random forest	94	76	93	96
3	Naive Bayes	94	70	92	96
4	K-nearest neighbour	95	77	94	95
5	Logistic regression	95	70	94	96
6	Decision tree	95	77	95	96
7	Support vector machine	97	77	97	97

도면15

