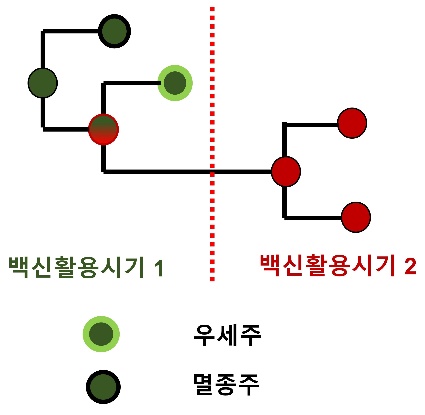
계절 인플루엔자 H3N2 우세주 예측 모델

1. 파일 구성
2. **Train 데이터 – 2000년부터 2019년까지 600개 H3N2 인플루엔자 Strain 구성**
3. 메타 데이터

* Access key로 활용 가능한 ID 및 각 Strain의 이름이 있습니다.
* 해당 바이러스가 유행했을 때 백신으로 활용된 백신주의 이름과 분류를 위해 사용된 Vaccine code가 있습니다.
* 해당 바이러스가 수집된 날짜가 연도, 월, 일의 형식으로 분류되어 있습니다.
* 위 모델의 **중요한 종속 변수인 우세주 여부(Dom)이 Binary code 형식 (0/1)로 Labeling 되어 있습니다.** 이는 인플루엔자 **핵심 유전체인 HA 유전체를 기준으로 차기 백신주 활용시기에 해당 바이러스가 후손을 만들었는지 (우세주 - 1) 실패했는지 (멸종주 – 0), 여부를 통해서 결정하였습니다.**



1. Nonsynonymous genetic distance from vaccine strain

* Access key로 활용 가능한 ID 및 각 Strain의 이름이 있습니다.
* 각 8개의 유전체별 계통수 (Phylogeny)를 활용하여, 해당 바이러스가 유행했을 때 백신으로 활용된 백신주로부터 유전거리를 측정하였습니다.
* 해당 파일은 핵산 변이 중, **아미노산의 염기서열이 바뀌는 변이** (Nonsynonymous genetic mutation) 누적 여부를 유전거리로 측정하였습니다.
* 각각의 Column은 독립변수로서 8개의 유전체별 다른 유전거리를 나타냅니다.

1. Synonymous genetic distance from vaccine strain

* Access key로 활용 가능한 ID 및 각 Strain 들의 이름이 있습니다.
* 각 8개의 유전체별 계통수 (Phylogeny)를 활용하여, 해당 바이러스가 유행했을 때 백신으로 활용된 백신주로부터 유전거리를 측정하였습니다.
* 해당 파일은 핵산 변이 중, 아미노산의 **염기서열이 바뀌지 않는 변이** (Synonymous genetic mutation) 누적 여부를 유전거리로 측정하였습니다.
* 각각의 Column은 독립변수로서 8개의 유전체별 다른 유전거리를 나타내었습니다.

텍스트, 스크린샷, 도표, 번호이(가) 표시된 사진

자동 생성된 설명

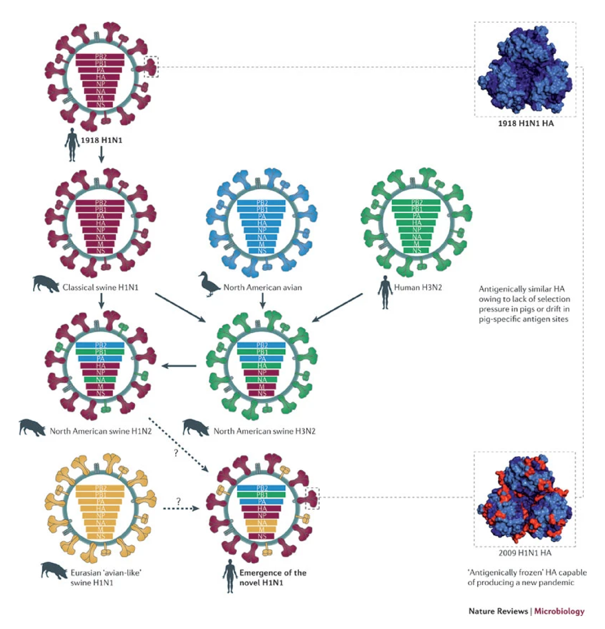
1. Other genetic changes

* Access key로 활용 가능한 ID 및 각 Strain의 이름이 있습니다.
* 해당 연구가 다양한 유전체 중 HA 변이를 중심으로 한만큼 HA 유전체와 연관된 다른 핵심 유전자 변이를 주로 포함시켰습니다.
* 우선 HA 유전체에 항체가 Binding 하는 Receptor binding domain (RBD)와 그 주변 15Å거리까지 단백질 위치에 각 백신주와 비교한 아미노산 변이 개수를 독립변수로 포함시켰습니다 (HA\_RBD, HA\_15A). <대조군 모델 독립변수>
* 우선 HA 유전체에 핵심 변이위치에 대한 주요 연구 중 2013년 Science지에 발표된 “Substitutions Near the Receptor Binding Site Determine Major Antigenic Change During Influenza Virus Evolution”에서 선정한 아미노산 부위에 각 백신주와 비교 시 아미노산 변이의 개수를 독립변수로 포함시켰습니다(HA\_Koel). <대조군 모델 독립변수>

스크린샷, 클립아트, 픽셀이(가) 표시된 사진

자동 생성된 설명

* 해당 파일에는 앞선 백신주로부터의 면역회피를 위한 변이누적 (Antigenic drift)외 다른 유전적 변이에 대한 정보를 독립변수로 제공하였습니다.
* Segment화 되어있는 인플루엔자 유전체에 변이를 누적하는 또 다른 방법인 Reassortment (재편성)을 통해, HA 중심, 다른 유전체 Segment가 삽입되었는지 여부를 GiRaF(Graph-incompatibility-based Reassortment Finder)로 측정하여 확인, 결과를 독립변수로 포함시켰습니다. 이는 Binary code 형식 (0 - 없음/1 - 있음)로 Labeling 되어있습니다 (HA\_유전체이름\_Reassort).



* 이는 전체 8개의 유전체중 단 한 번의 Reassortment가 측정되었는가 (All\_Reassort) 혹은 HA와의 Reassortment의 합(HA\_Reassort\_Sum)과 같은 변형된 독립변수도 제공하였습니다.

1. **Test 데이터 – 2000년부터 2024년까지 724개 H3N2 인플루엔자 Strain 구성**

위 연구는 아래 그림과 같이, 20년간의 유전적 변이의 특징을 활용하여 차기 우세주를 예측하는 4번의 모델 Cross-validation을 수행하였습니다.

스크린샷, 텍스트, 다채로움, 폰트이(가) 표시된 사진

자동 생성된 설명

해당 연구를 위해 2019년부터 2024년까지 124개의 H3N2 인플루엔자 8개 유전체 정보를 추가로 수집하여, 계통수를 추정하여 백신주부터의 유전거리를 측정하였습니다.

1. 메타 데이터

* Access key로 활용 가능한 ID 및 각 Strain의 이름이 있습니다.
* 해당 바이러스가 유행했을 때 백신으로 활용된 백신주의 이름과 분류를 위해 사용된 Vaccine code가 있습니다.
* 해당 바이러스가 수집된 날짜가 연도, 월, 일의 형식으로 분류되어 있습니다.
* 위 모델의 **중요한 종속 변수인 우세주 여부(Dom)이 Binary code 형식 (0/1)로 Labeling 되어 있습니다.** 이는 인플루엔자 **핵심 유전체인 HA 유전체를 기준으로 차기 백신주 활용시기에 해당 바이러스가 후손을 만들었는지 (우세주 - 1) 실패했는지 (멸종주 – 0), 여부를 통해서 결정하였습니다.**

1. Nonsynonymous genetic distance from vaccine strain

* Access key로 활용 가능한 ID 및 각 Strain 들의 이름이 있습니다.
* 각 8개의 유전체별 계통수 (Phylogeny)를 활용하여, 해당 바이러스가 유행했을 때 백신으로 활용된 백신주로부터 유전거리를 측정하였습니다.
* 해당 파일은 핵산 변이 중, **아미노산의 염기서열이 바뀌는 변이** (Nonsynonymous genetic mutation) 누적 여부를 유전거리로 측정하였습니다.
* 각각의 Column은 독립변수로서 8개의 유전체별 다른 유전거리를 나타냅니다.

1. Synonymous genetic distance from vaccine strain

* Access key로 활용 가능한 ID 및 각 Strain 들의 이름이 있습니다.
* 각 8개의 유전체별 계통수 (Phylogeny)를 활용하여, 해당 바이러스가 유행했을 때 백신으로 활용된 백신주로부터 유전거리를 측정하였습니다.
* 해당 파일은 핵산 변이 중, 아미노산의 **염기서열이 바뀌지 않는 변이** (Synonymous genetic mutation) 누적 여부를 유전거리로 측정하였습니다.
* 각각의 Column은 독립변수로서 8개의 유전체별 다른 유전거리를 나타내었습니다.