INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD
R&D SHOWCASE 2021
Technology, Social Impact

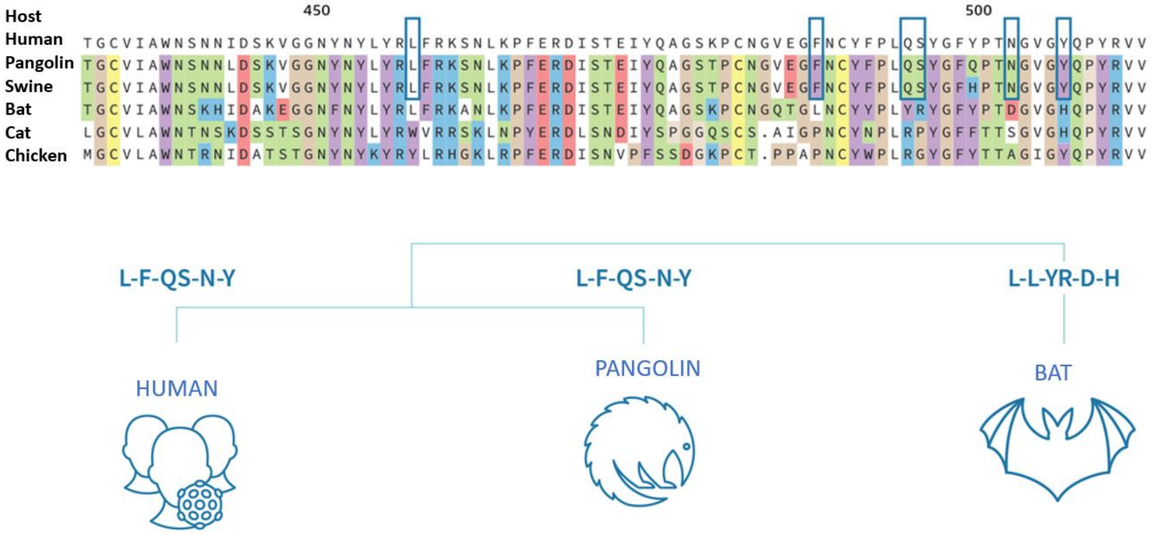
PREHOST: Host Prediction of virus using genome sequences.

ABSTRACT

Coronavirus, an RNA virus is capable of causing Respiratory Tract Infections in humans and birds. The ability of Coronavirus to transmit has introduced a challenge to predict the host of coronavirus which is essential to control the future pandemics. To evaluate their intimidating remark to humans, we aimed to extrapolate the potential hosts of viruses belonging to Coronaviridae family using machine learning algorithms. Both the random forest (RF) model and the KNN model achieved high accuracies while training them on a data consisting of 32k genome sequences (99.81% and 99.76% respectively). The results indicate that machine learning algorithms alone can be used to predict hosts of coronavirus.

OBJECTIVE

To develop a host prediction system that is capable of predicting the host of the virus belonging to the coronaviridae family, given a genomic sequence.



METHOD

- Downloaded the genome sequences from Viral Pathogen Database.
- Data was preprocessed, where the duplicate sequences were eliminated.
- Implemented SMOTE to overcome the problem of class imbalance.
- Embeddings generated of the genome sequences were fed as an input to the model to predict the host of the virus.

Authors:Anusha Chaturvedi, Kushal Borkar, Prof. Vinod P.K, Prof. Deva Priyakumar

Research Center Name:HAI