Establishing Influenza A viral mutation database through literature and data mining

Ambreen Kedwaii¹, Pavan Kumar Attaluri², and Guoqing Lu³

¹Department of Pathology and Microbiology, University of Nebraska Medical Center
²Department of Computer Science and Engineering, University of Nebraska at Omaha
³Department of Biology, University of Nebraska at Omaha

Understanding virus evolution is critical for influenza surveillance and vaccine strain selection. Genetic mutation and genome re-assortment are two major evolutionary mechanisms which result in new viral strains with distinct genotypes. The goal for this project is to identify biologically significant mutations that will assist in annotating influenza A sequences. Two approaches were used, including information retrieval and data mining. Informative mutation sites were collected manually from literature as compared with those predicted by the decision tree approach.

In this preliminary study, around 250 sites were found in hemagglutinin of human influenza A viruses from about 50 journal papers, whereas 68 positions were discovered by decision tree, with a total of 36 mutations common in two datasets. A relational database was built to retrieve, manage and use this valuable information. In addition, these mutation sites were used to build a Hidden Markov Model (HMM) model and a web tool was developed to use this model to identify informative mutations of a given sequence. Biological meanings of the predicted mutations will be provided as well. We are in the process to collect more mutation sites to enrich our database and extend web functions for the annotation of influenza A genotypes.

The database and web functions are available at [http://glee.ist.unomaha.edu/~pattaluri/mutation/](http://glee.ist.unomaha.edu/~pattaluri/mutation/).