Low-dimensional clustering reveals new influenza strains before they become dominant

Jiankui He[1] and Michael W. Deem[1,2]

***Short Abstract* — Influenza has been circulating in the human population and has caused three pandemics in the last century (1918 H1N1, 1957 H2N2, 1968 H3N2). The newly appeared 2009 A(H1N1) is classified by the World Health Organization (WHO) as the fourth pandemic. Influenza has a high evolution rate, which makes vaccine design challenging. We here consider an approach for early detection of new dominant strains. By clustering the 2009 A(H1N1) sequence data, we found two main clusters. We then define a metric to detect the emergence of dominant strains. We show on historical H3N2 data that this method is able to find a cluster around an incipient dominant strain before it becomes dominant. For example, for H3N2 as of 30 March 2009, we see the cluster for the new A/British Columbia/RV1222/2009 strain. This strain detection tool would appear to be useful for annual influenza vaccine selection.**

***Keywords* — Influenza, H1N1, vaccine, clustering.**

I. Purpose

THE influenza viruses are hyper-mutating viruses Influenza escapes the human immune system by continual antigenic drift [1]. Usually, there is a dominant influenza strain circulating in the population for each flu season. The flu vaccine is most effective when it matches this dominant circulating strain. Due to frequent evolution of the antigenic regions of the influenza virus, the composition of the flu vaccine is typically modified annually. However, since the influenza strains used in the flu vaccine are decided 6 months before the flu season, a mismatch between the vaccine strain and dominant circulating strain may happen if there are significant evolutionary dynamics which results in a decrease in vaccine efficacy and more infections in the population. Such a situation arose for the H3N2 virus in the 2009-2010 flu season, when the A/British Columbia/RV1222/2009 was detected after the H3N2 vaccine component was selected [2]. Early accurate prediction of the dominant circulating strain is an essential and important task in influenza research. Here, we present a low-dimensional clustering method that can correctly detect the cluster containing next dominant strain before it becomes dominant.

[1]Department of Physics & Astronomy and [2]Department of Bioengineering, Rice University, E-mail: Jiankui.he@rice.edu

II. Methods

A. Multidimensional scaling

Hemagglutinin (HA) protein sequences are downloaded from Influenza Virus Resources database. First, we do multialignment to HA1 proteins. Second, protein distance between any two proteins is calculated as the number of amino acids difference divided by the length of protein. Third, the multidimensional scaling produces a protein distance map. In this map, each point represents a flu strain. The Euclidean distance between two points in the map approximates the protein distance between these two flu strains.

B. Detecting new strains

We predict a dominant strain may arise in a cluster in the future if this cluster satisfies the following two criteria: 1, this cluster can be detected by kernel density estimation, and is separated from the cluster that contains the current dominant strain; 2, p_{epitope} of the new cluster with regard to the current dominant strain is much larger than with the centroid of the new cluster [3].

C. Results

We can detect A/British Columbia/RV1222/2009 earlier using the data as of 03/30/2009. We can also correctly detect A/Fujian/411/2002 as a dominant strain in 2003-2004 season, when vaccination strain mismatches the dominant. We recommend A/Texas/05/2009(H1N1) as H1N1 vaccination strain and A/British Columbia/RV1222/2009 as H3N2 vaccination strain in 2010-2011 season.

III. Conclusion

Our method can be used for vaccine selection. It seems that the best time to detect a new dominant strain is at the earliest appearance of immune-escape new strains, when these new strains begin to form a cluster.

Rapid evolution of influenza under mutation, recombination and genetic drift is experienced as a cluster, rather than individually [1,4]

References

[1] Smith DJ, et al. *Science* 2004;305:371–376.
[2] ProMed. 2009 May 5. archive no. 20090505.1679.
[3] Gupta V, Earl DJ, Deem M *Vaccine* 2006;24:3881–3888.
[4] Plotkin JB, Dushoff J, Levin SA. *PNAS* 2002;99:6263–6268.