Mathematics, Statistics, and Computer Science **@ UIC**

Statistics and Data Science Seminar

An ensemble distance measure of k-mer and Natural Vector for the phylogenetic analysis of multiple-segmented viruses

Hsin-Hsiung Huang (University of Central Florida)

Abstract: The Natural Vector combined with Hausdorff distance has been successfully applied for classifying and clustering multiple-segmented viruses. Additionally, k-mer methods also yield promising results for global genome comparison. It is not known whether combining these two approaches can lead to more accurate results. The author proposes a method of combining the Hausdorff distances of the 5-mer counting vectors and natural vectors which achieves the best classification without cutting off any sample. Using the proposed method to predict the taxonomic labels for the 2,363 NCBI reference viral genomes dataset, the accuracy rates are 96.95%, 94.37%, 99.41% and 93.82% for the Baltimore, family, subfamily, and genus labels, respectively. We further applied the proposed method to 48 isolates of the influenza A H7N9 viruses which have eight complete segments of nucleotide sequences. The single-linkage clustering trees and the statistical hypothesis testing results all indicate that the proposed ensemble distance measure can cluster viruses well using all of their segments of genome sequences.

Wednesday, September 30 at 4:00 PM in SEO 636