UNIVERSITY of GUELPH

COLLEGE of ENGINEERING AND PHYSICAL SCIENCES

SCHOOL OF COMPUTER SCIENCE

Qualifying Exam

Monday September 14th, 2020 at 2:45pm on Zoom (To view please contact Dr. Joe Sawada at jsawada@uoguelph.ca)

Pathogenicity Classification of Viruses using Deep Learning Algorithms

Akshay Chadha

Chair: Dr. Joe Sawada
Advisory Member: Dr. Rozita Dara
Advisory Member: Dr. Zvonimir Poljak [PopMed]
Non-Advisory Member: Dr. Luiza Antonie
Non-Advisory Member: Dr. Olaf Berke [PopMed]

ABSTRACT:

Over the past decades, numerous viral epidemic outbreaks across the globe related to Avian Influenza virus (AI), Porcine Reproductive and Respiratory Syndrome virus (PRRSV) and other viruses have resulted in large scale economic losses and death of livestock. Despite the availability of high dimensional viral genetic sequences, in-silico genotypic to phenotypic relationship establishment has remained a challenging problem for a number of reasons including availability of a small number of relevant genetic sequences, the lack of domain expertise about genetic virulence of different viral diseases and differential phenotypic impact of a single viral strain among different hosts. The virulence encoding information present within limited vocabulary based high dimensional genetic sequences makes establishing genetic-phenotypic relationships an interesting task but difficult task.

Traditional hand designed bioinformatics solutions failed to generalize for different diseases. The recent success of deep learning architectures such as convolutional neural networks, recurrent neural network and autoencoders on high dimensional datasets helped it achieve state of the art accuracy over traditional machine learning classifiers in numerous application domains including bioinformatics applications. Convolutional architectures achieved high accuracy in numerous genomic sequence classification tasks and recent work helped to improve the interpretive capabilities of convolutional networks which helped to gain novel insights on biological processes. Keeping this in mind, in this thesis, I propose to develop a deep learning approach for pathogenicity classification of small and complex viruses (e.g., PRRSV).

The long-term goals of my work proposed in this thesis are: (1) To classify small, complex and high dimensional viral genomic sequence dataset labelled based on phenotypic outcomes observed in form of clinical impact with high accuracy. (2) Identify the possible contribution of genotypic-phenotypic relationship establishing motifs present in labelled viral genomic sequence dataset. The main objectives of my thesis are: (1) to establish pathogenicity classification of small and complex viral genomic sequence datasets of AI (H5Nx haemagglutinin gene) and PRRS (open reading frame-5) using machine learning & deep learning techniques; (2) to establish the role and importance of sequence motifs in pathogenicity classification of viral genomic sequences; (3) adoption of ensemble and hybrid classifier approaches, including the demographic data, to increase the robustness and accuracy in pathogenicity classification tasks for small and complex datasets; and (4) Evaluation of supervised and unsupervised deep representation learning approaches to evaluate the effectiveness of using knowledge obtained from one viral dataset to another for pathogenicity classification.

To validate the first and second objective, as a preliminary study I analyzed the efficacy of using machine learning and deep learning algorithms for pathogenicity classification of H5Nx Avian Influenza (AI) Hemagglutinin (HA) gene sequences, for which the pathogenicity markers were established in literature. I used a convolutional neural network for pathogenicity classification of H5Nx AI HA gene sequences. The network achieved 99.20% mean accuracy over 10-fold cross validation and successfully identified positions of motifs which are known to participate in pathogenicity classification of HA gene sequences. After validating the findings on pilot H5Nx AI genomic dataset, I obtained preliminary results for virulence impact classification of PRRSV open reading frame-5 genomic sequences. In future, I propose: (1) to improve the classifiers' performance by using ensemble techniques and developing hybrid neural networks to overcome the challenges related with complex datasets; (2) to use semi-supervised learning to overcome challenges related with small dataset; and (3) to study the use of representation learning techniques for improving generalization capabilities of viral datasets with limited availability of training samples.