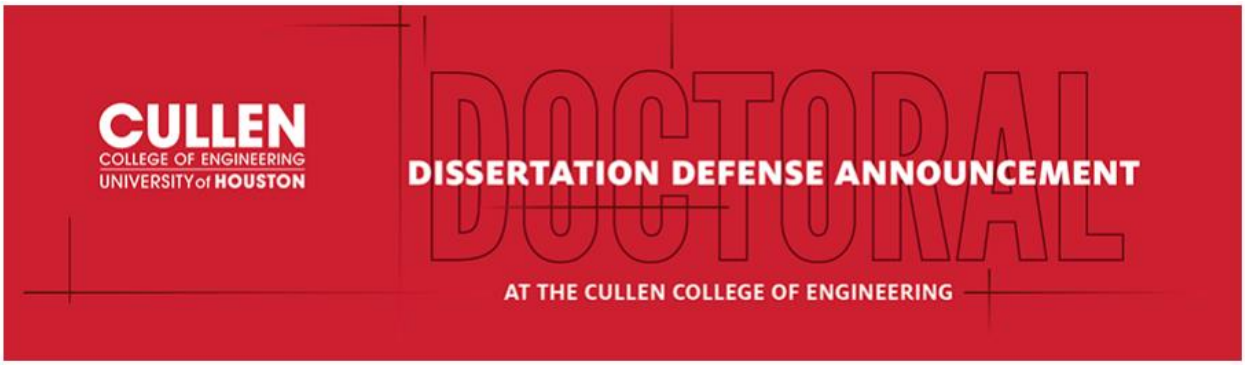

Machine Learning and Network-driven Approaches for Risk Genes Discovery

Shiva Afshar
April 28, 2023; 10:00 AM - 12:00 PM (CDT)
Location: E214 Engr. Bldg. 2

Committee Chair:
Ying Lin, Ph.D.

Committee Members:
Gino Lim, Ph.D. | May Feng, Ph.D. | Wenjiang Fu, Ph.D. | Shizhong Han, Ph.D.

Abstract

There is strong evidence that many complex diseases such as mental disorders are linked to genetic variants. Thus, it is crucial to identify the causal genes or variants contributing to disease onset to advance the understanding of disease pathology and inform better treatment. Due to the complexity of the human genome only a few causal genes or variants have been identified. To enhance the disease-associated risk genes discovery in complex diseases, this thesis aims to develop network-based and machine learning methods for explicitly capturing the cell-type-specific gene interactions and integrating these interactions with existing gene-disease association evidence for risk gene prioritization.

To achieve this goal, this thesis proposed several innovative techniques. Firstly, a multimodal deep learning model was developed to integrate multi-source and multi-structure data including single-cell gene expression and global gene interactions for predicting cell-type specific gene networks.

The effectiveness of the proposed method was demonstrated by comparing its prediction performance with baseline models and downstream analysis for risk gene discovery. Secondly, a supervised machine learning approach was employed to integrate various genomic features and cell-type specific gene networks' topological information to prioritize disease risk genes. The method was employed to prioritize autism risk genes and results demonstrated that our gene ranking system provides a useful resource for prioritizing autism candidate genes. Thirdly, an unsupervised ensemble learning model was developed to combine the multisource correlated disease-gene association scores for risk gene discovery. The results of artificial and real datasets demonstrated that the proposed method can efficiently integrate individual scores to an ensembled score without the need of ground truth data.

The supervised methods proposed in this thesis can be applied to different complex human disease risk genes discovery problems and be effective to find more novel disease-associated risk genes using the ground truth information. Furthermore, some of the models developed in this thesis are unsupervised methods which are applicable to the problems that the ground truth information is very limited or is not available.