Oral Defense Announcement

University of Missouri – St. Louis Graduate School

An oral examination in defense of the dissertation for the degree
Doctor of Philosophy in Mathematical & Computational Sciences with an emphasis in Computer Science

Kenneth P. Smith, Jr.

M.A. in Mathematics, December 2020, University of Missouri-St. Louis
M.S. in Electrical Engineering, May 2013, Missouri University of Science and Technology
B.S. in Physics, May 2011, Missouri University of Science and Technology

Finding Combinatorial Patterns in Real Valued Omics Data

Date: April 18, 2024
Time: 1:00 p.m. to 3:00 p.m.
Place: 304 Express Scripts Hall

Abstract

Precision medicine is a healthcare approach which tailors disease prevention and treatment to an individual, based on their genetics, environment, lifestyle, and physiological state. These factors interact to produce biological changes that can be measured to produce data called omics, and include genomics, lipidomics, and proteomics. Despite the abundance of omics data and analysis techniques, researchers still struggle to identify biological findings that replicate across data sets and translate into clinical applications. In this dissertation, we employ combinatorial optimization techniques to improve upon three steps in the precision medicine analysis pipeline: 1) data cleaning, 2) community detection, and 3) feature selection. Real-world data sets are often messy, containing missing values, erroneous measurements, and repeated entries. To better prepare these data sets for analysis, we introduce a data cleaning suite called Mr. Clean that removes missing elements, while maximizing the amount of valid data kept. Within Mr. Clean, we include exact approaches and high quality greedy methods. After the data has been cleaned, community detection and feature selection can be used to identify biologically significant patterns. In biology, networks are used to represent several relationships, including gene co-expression, gene regulation, and signaling pathways. Community detection is used to identify analytes in these networks that exhibit combinatorial relationships. To better identify communities of biological relevance, we further develop the Sieve objective function, which is tailored to large, sparse networks. Additionally, we develop an exact algorithm and approximation algorithm for community detection based on maximizing Sieve. In addition to inferring combinatorial relationships from a network, we can directly search for combinations of features associated with a disease. Identifying the most useful combination of features from the billions or trillions of possibilities is called feature selection. Using both greedy and exact algorithms, we develop a feature selection technique based on the Youden J Statistic. We demonstrate the usefulness of this technique for selecting features for general classification and subtype identification. Finally, we report on biological insights gleaned from our approach, particularly in the areas of Alzheimer's Disease and COVID-19.

Defense of Dissertation Committee

Sharlee Climer, Ph.D.
Badri Adhikari, Ph.D.
Cezary Janikow, Ph.D.
Daniel Jacobson, Ph.D.