Time & Location: July 6, 2017 at 11:00 AM in HEC 438
Title: DATA REPRESENTATION IN MACHINE LEARNING METHODS WITH ITS APPLICATION TO COMPILATION OPTIMIZATION AND EPITOPE PREDICTION

In this dissertation we explore the application of machine learning algorithms to compilation phase order optimization, and epitope prediction. The common thread running through these two disparate domains is the type of data being dealt with. In both problem domains we are dealing with discrete/categorical data, with its representation playing a significant role in the performance of classification algorithms.

We first present a neuroevolutionary approach which orders optimization phases to generate compiled programs with performance superior to those compiled using LLVM’s -O3 optimization level. Performance improvements calculated as the speed of the compiled program’s execution ranged from 27% improvement for the ccbench program, to 40.8% for bzip2.

This dissertation then explores the problem domain of epitope prediction. This problem domain deals with text data, where protein sequences are presented as a sequence of amino acids. DRREP system is presented, which demonstrates how an ensemble of extreme learning machines can be used with string kernels to produce state of the art epitope prediction results. DRREP was tested on the SARS subsequence, the HIV, Pellequer, Antijen datasets, and the standard SEQ194 test dataset. AUC improvements achieved over the state of the art ranged from 3% to 8%.

We then present the SEEP epitope classifier, which is an SMV ensemble based classifier which uses contjoint triad feature representation, and produces state of the art classification results. SEEP leverages the domain specific knowledge based protein sequence encoding developed within the protein-protein interaction research domain. Using an ensemble of SVMs, and a sliding window based pre and post processing pipeline, SEEP achieves an AUC of 91.2 on the standard SEQ194 test dataset, a 24% improvement over the state of the art.

Finally, this dissertation concludes by formulating a new approach for distributed representation of 3D biological data through the process of embedding. Analogously to word embedding, we develop a system that uses atomic and residue coordinates to generate distributed representation of residues. Preliminary results are presented where the Residue Surface Vectors, distributed representations of residues, are used to predict conformational epitopes and protein-protein interactions, with promising proficiency. The generation of such 3D BioVectors, and the proposed methodology, opens the door for substantial future improvements, and application domains.

Major: Computer Science

Educational Career:
Bachelor's of Mathematics, BS, 2007, University of California, Santa Cruz
Master's of Computer Science, MS, 2014, University of Central Florida

Committee in Charge:
Shaojie Zhang, Chair, Computer Science
Damian Dechev, Co-Chair, Computer Science
Gary Leavens, Computer Science
Avelino Gonzalez, Electrical & Computer Engineering
Degui Zhi, School of Biomedical Informatics, University of Texas Health Science Center

Approved for distribution by Shaojie Zhang, Committee Chair, on June 21, 2017.
The public is welcome to attend.