UNIVERSITY of GUELPH

COLLEGE of ENGINEERING AND PHYSICAL SCIENCES

SCHOOL OF COMPUTER SCIENCE

PhD.CSCI Seminar 1

Friday May 8, 2020 at 9AM on Zoom (Please contact Dr. Stefan C Kremer to join at skremer@uoguelph.ca)

Computational Tools for Protein Domain Comparison and Classification

Angela Murcia Rios

Advisor: Dr. Stefan C Kremer

Co-Advisor: Dr. Steffen Graether [Molecular and Cellular Biology]

Advisory Committee: Dr. Ying "Tiffany" He

Advisory Committee: Dr. Vladimir Ladizhansky [Physics]

ABSTRACT:

Proteins are the most diverse kind of macromolecules. They can range in size and have a wide variety of uses within a biological cell. Every protein is made up of a specific amino acid sequence. These sequences are mostly determined by the in vitro translation of coding sequences derived from gene prediction. Coding sequences are regions found in DNA or RNA that determine the pattern of amino acids in a protein. The UniProt database is known as the central hub of protein sequences where it currently stores roughly 560,000 sequences.

Every protein adopts a distinct three-dimensional shape that determines its function. The structures of proteins can be experimentally determined where their structural information is usually stored in a text format file known as a PDB file. The PDB file contains Cartesian coordinates of all atoms as well as other structural information of the protein. PDB files are stored in the Protein Data Bank (PDB) database, which currently holds over 160,000 protein structures and still continues to grow.

To understand proteins and their function, it is important to be able to categorize them into groups. Proteins are commonly sorted based on their similarities in amino acid sequence and structure. Based on these similarities, a protein's function and ancestral background can be inferred. Many categorization techniques have broken down amino acid sequences and protein structures into a smaller, more comparable part known as domains. Domains usually correspond to a specific protein function such as a binding site.

With the growing database of PDB files and protein sequences, it has become more important to find a way to automate the categorization of proteins. In this seminar, I will summarize some of the different methods available for protein domain classification into families, such as those of CATH and SCOP. I will also focus on the automated processes available within these databases for protein sequence and structural classification, specifically that of HMMER: a probabilistic inference method based on profile Hidden Markov Models.