Current Bioinformatics Research Projects

1. **Study the Mental Health Impact of COVID-19 in Canada with Combined Statistics and Machine Learning (ML)**

Summary: This project aims to explore the changes in human behavior and wellbeing, examine how these changes differ among people in different groups, identify important factors influencing mental health conditions, and make recommendations on effective strategies to help different individuals. We will integrate both statistics and machine learning methods to robustly analyze available survey datasets for Canadians. For ML, methods will be explored including Bayesian networks (BNs), Random Forest (RF) (available in both R (e.g. Random Forest version 4.6) and Python (e.g. Scikit-learn library), support vector machine (SVM), logistic regression, and naive Bayes.

Students’ tasks may include

* preprocessing the survey data, understanding the data properties, and applying Bayesian networks and probabilistic graphical models for the data analysis.
* applying various ML methods to the data analysis and develop ready-to-use tools integrating all the methods.

1. **Big-Data Processing Tools for Whole-Genome Population Data**

Summary: This project aims to develop and revise biological data processing and automate analysis tool/pipeline. Whole-genome data usually are in varied format depending on platforms and analysis software. For example, Single Nucleotide Polymorphisms (SNPs) data, which records changes of single DNA base-pairs, is typically in plain text but with different formats in storing the information. Common formats includes plink (.ped and .map file), csv, hmp, and vcf. Data format conversion becomes necessary in data analysis across different programs but there are limited tools available for them now. Currently, we have developed an initial version of the tool for data format conversion in Java and will need to test it out and make necessary changes for improvement.

Students’ tasks may include

* collect SNP dataset from online database and run testing on the existing version of the Java-based processing tool.
* Identify issues and limitations of the current tool and revise the tool for improvement.

1. **Comparison of Deep Learning and Principal Component Analysis for Dimensionality Reduction of scRNA-Seq Data**

Summary: This project aims to study different methods for dimension reduction on Single-cell RNA-seq (scRNA-seq) data and develop automate analysis tool/pipeline for such purpose. An scRNA-seq dataset contains the expression profiles of a large number of genes, where each gene corresponds to a dimension, and the expression profile of each cell corresponds to a data point in the high dimensional cell state space. The number of genes is much larger than the number of cells (e.g. 100,000 vs 1,000) which makes the dimension reduction on the genes necessary before analysis. Currently, we have investigated two deep learning methods and a statistical method (PCA) on two datasets and received mixed results. We will expand the study and investigate the way to achieve better results.

Students’ tasks may include

* collect additional scRNA-seq datasets from online database and test the two deep learning methods and a statistical method (PCA) on them.
* Explore the literature to find other methods that could be used for dimension reduction and test their usability on scRNA-seq data.
* Develop a tool/pipeline to automatedly run multiple methods on the same dataset to compare the performance.

1. **Feature Selection on Whole-genome SNP Data**

Summary: This project aims to study different feature selection methods to identify significant

Features from the whole-genome SNP data. Single Nucleotide Polymorphisms (SNPs) are changes of single DNA base-pairs. Genome-Wide Association Studies (GWAS) have served as primary methods for the past decade for identifying associations between genetic variants and traits or diseases (also known as phenotype). Out of all SNPs, the ones correlated with phenotypes are just a small subset, which means that the genotype data is sparse. In addition, there are millions of SNPs but only hundreds of samples. Such data brings a complex challenge to current GWAS methods as many of them fail to find the correlated SNPs on large complex data. By selecting a reduced number of SNPs (features) with significantly larger effects compared to other SNPs, researchers can apply existing methods on the most promising SNPs.

Currently, we have studied two methods based on The Least Absolute Shrinkage and Selection Operator (LASSO) technique called biglasso and AUTALASSO, respectively. The initial results show that biglasso works well on binary phenotypes and AUTALASSO on quantitate phenotypes. We will expand the study to further validate the findings and investigate other methods combined with the above two to achieve better results.

Students’ tasks may include

* collect additional SNP datasets from online database and test the two lasso-based methods to validate the findings.
* Develop a tool/pipeline to automatedly run multiple methods on the same dataset and compare the performance.
* Explore the literature to find other methods that could be used for feature selection, e.g. deep learning methods, and apply them on the SNP datasets to examine the performance.

1. **Phenotype Prediction using Whole-genome SNP Data**

Summary: This project builds upon the success of the feature selection. With the identified significant features (subsets of the whole SNPs), we aim to develop models to predict phenotypes. To get started, we will try to reproduce work from article “Interpretable genotype-to-phenotype classifiers with performance guarantees”. The paper is available at

<https://www.nature.com/articles/s41598-019-40561-2#rightslink>

and the code & data at <https://github.com/aldro61/kover2_paper>

tutorial: <https://aldro61.github.io/kover>

Students’ tasks:

* reproduce the work from the authors’ data
* test the above method on SNP data