GENERAL PUBLIC

STATISTICAL AND MACHINE LEARNING CHALLENGES IN CLASSIFICATION AND INTEGRATION OF MOLECULAR AND GENOMIC DATA

Modern life sciences routinely generate high-dimensional data to elucidate complex and systematic phenomena underlying molecular activities, genomes, disease, and others. As new measurement techniques are invented and advanced, scientists quantify an ever increasing amount of information. For example, instead of measuring the mRNA abundance for one candidate human gene under a given experimental study design, a molecular biologist can now utilize DNA microarray or high-throughput sequencing technology to obtain the mRNA abundance for every gene in the human genome. Analogously, we can now examine multiple levels of biological variations, by measuring DNA sequences, gene expression, and abundance of metabolites across a large number of individuals. Such high-throughput measurement techniques – which has enabled genome-wide analysis, functional genomics, and genomic medicine – require rigorous statistical methods and integrative approaches to realize their full potentials.

Motivated by critical challenges in genomics and metabolomics, this project will develop novel statistical and machine learning methods that can directly learn from multiple sources of molecular, biological, and genomic data.

Firstly, how to tell if a large number of categorical data are substantially different (or similar) using highly popular Tanimoto/Jaccard similarity coefficients? Categorical data include DNA sequences, metabolite absence/presence profiles, molecular/reaction fingerprints, and others. And, comparing such data is fundamental to identification of orthologs, molecular database search, and classification.

Secondly, how to identify robust membership of clusters when attempting to classify a large number of observations? Although clustering is routinely used in genomics and metabolomics, statistical evaluation of genes and metabolites in clusters are still an unsolved problem. Our proposed method will be able to tell if genes and metabolites in a given cluster are due to random chances (i.e., unstable). This procedure can evaluate and improve results of many classification algorithms.

Thirdly, how does lipid metabolism alter the progress of ovarian cancer? Methodologically, we would also investigate how to discover common latent structure from multiple types of biological datasets. Our preliminary analysis demonstrated that network of proteins involved in lipid metabolism may play an important role in ovarian cancer, of which Poland reports the fourth highest incidence rate in the world. Using expression levels of genes and metabolites taken from cancer patients, we will identify ovarian cancer sub-types and related biological pathways

As advocates of open science and open source movements, we will freely and widely release methods and algorithms developed during this grant phase as *R* packages for practical use in a wide range of applications. Furthermore, this SONATA grant, supporting a United States scientist in a premier Polish research university, will encourage interdisciplinary and international collaboration across biostatistics, informatics, and molecular biology.