MLGenSig: Machine Learning Methods for building of Integrated Genetic Signatures

The main scientific goal of this project is to develop a methodology for integrated genetic signatures based on data from divergent high-throughput techniques used in molecular biology. Integrated signatures base on ensembles of signatures for RNA-seq, DNA-seq, data as well for methylation profiles and protein expression microarrays.

The advent of high throughput methods allows to measure dozens of thousands or even millions features on different levels like DNA / RNA / protein. And nowadays in many large scale studies scientists use data from mRNA seq to assess the state of transcriptome, protein microarrays to asses the state of proteome and DNA-seq / bisulfide methylation to assess genome / methylome.

Genetic signatures are widely used in different applications, among others: for assessing genes that differentiate cells that are chemo resistant vs. cells that are not, assess the stage of cell pluripotency, define molecular cancer subtypes. For example, in database Molecular Signatures Database v5.0 one can find thousands of gene sets - genetic signatures for various conditions. There are signatures that characterize some cancer cells, pluripotent cells and other groups.

But they usually contain relatively small number of genes (around 100), results with them are hard to replicate and they are collection of features that were found significant when independently tested. In most cases signatures are derived from measurements of the same type. Like signatures based of expression of transcripts based on data from microarrays or RNA-seq, or methylation profile or DNA variation. We are proposing a very different approach. First we are going to use machine-learning techniques to create large collections of signatures. Such signatures base on ensembles of small sub-signatures, are more robust and usually have higher precision.

Then out of such signatures we are going to develop methodology for meta-signatures, that integrate information from different types of data (transcriptome, proteome, genome). Genetic signature may be used both to better understand what is different between two or more sets of cells or they may be used for scoring if a new sample has a given property (defined by it’s signature).