Quantifying predictive value of biological data types in machine learning models of cancer outcome

Samantha Jensen

If this project is selected I would definitely like to be a part of it.

1. Description of the project

Precision medicine is a growing movement toward utilizing molecular diagnostics to guide medical decisions. It is particularly useful when applied to cancer treatment, as knowing details about cancer stage, genetic pathology, and tumor type can inform life-saving decisions. Increasingly, physicians may use genetic, proteomic, epigenetic, and expression data to determine treatment strategy and even choose specific chemotherapy drugs[1]. Machine learning algorithms can parse large amounts of biological data and find patterns in order to categorize individuals. For example, a computer can now differentiate between acute lymphoblastic and acute myeloid leukemias based only on DNA microarray gene expression data. As the two leukemias have different treatment regimes and pathologies, this information is crucial to maximize efficacy and minimize toxicity of treatment[2]. After training with gene expression information categorized by tumor type, when presented with new unlabeled data the algorithm could quickly and accurately predict leukemia type. What previously took a hematopathologist's interpretation of the tumor's morphology, histochemistry, immunophenotyping, and cytogenetic analysis to distinguish can now be accurately determined by a computer with a single biological sample. It is obvious then why precision medicine is the darling of modern cancer treatment. But despite the hundreds of millions of dollars currently being invested into generation of data for this purpose, we have little understanding of which genomic data types are most useful in machine learning algorithm design. This project will comprehensively compare the performance of different types of molecular data in a number of different cancer types and with a variety of machine learning algorithms to determine the most accurate combinations. The scope of this benchmarking analysis will far exceed any that has been conducted to date and will provide researchers with empirical performance insights into which data types and algorithms provide the most value for informing cancer-treatment decisions.

2. What features the data set would include

In order to make this project doable in a semester I would probably only take a piece of the larger project I'm working on. This would mean that we would only do this analysis in one cancer type (probably BRCA2). Each data type has different features to train on, but for purposes of this proposal here is what the first five columns and row of one of the protein expression data files looks like:

tcga_barcode	14-3-3_beta-R-V	14-3-3_epsilon-M-C	14-3-3_zeta-R-V	4E-BP1-R-V
TCGA-3C-AALI	-0.0176	-0.0188	-0.0550	0.0094

3. How and from where would the data set be gathered and labeled

The Cancer Genome Atlas (TCGA), the national resource for the cancer research community, contains 2.5 petabytes of genomic data, including DNA mutations, mRNA, miRNA, and protein expression data, and information on epigenetic tagging for tens of millions of individuals. The data would need to be cleaned and prepared for use.

1	MUT	DNA somatic mutations
2	CNV	DNA copy number variations
3	mRNA	mRNA expression
4	miRNA	miRNA expression
5	RPPA	protein expression
6	MET	DNA methylation

Table 1: Different biological data types we will test to see which are most predictive of characters of interest.

1. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **13,** 8–17 (2015).
2. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286,** 531–537 (1999).