Scientific Abstract

SU2C Machine Learning for Cancer Convergence Research Team:

"Machine Learning for Cancer"

[This abstract was provided by the scientists when their application was accepted.]

Modeling approaches: To better understand why some patients with a high tumor burden fail respond to checkpoint inhibitors, the Friedman and Wherry groups have established a pipeline for evaluating immune health. In particular, they use a variety of molecular measures to assess T cell exhaustion or senescence before and after therapy. We will explore multi-level modeling approaches to these data. On the one hand, standard machine learning approaches will be applied to determine which variables, if any correlate with outcome. However, we expect that such approaches will have limited statistical power because the number of measured variables exceeds the number of patients by many orders of magnitude. To overcome this challenge, we will examine two other types of approaches.

First, we will explore a range of machine learning methods that incorporate dimensionality reduction. These methods included lasso and ridge regression, reduced rank regression, and deep generative models. Second, we will develop a novel class of methods that use prior biological knowledge to model groups of molecules that function in related pathways. These approaches take advantage of annotated pathways as well as unannotated relationships that can be derived from high-throughput protein- protein interactions.

Expanding the range of molecular data: We will explore ways to complement the existing systems immunology pipeline by collecting untargeted metabolomics and/or phosphoproteomics. The current pipeline combines several genome-wide methods including genomic sequencing, ATAC-Seq and RNA-Seq, with targeted proteomics to identify relevant pathways. In our experience untargeted metabolomics and phosphoproteomics are highly complementary to these methods, as the metabolomics and phosphoproteomics capture many post-translational regulatory phenomena that are not detectable by the other methods. We will work with Friedman and Wherry to determine which samples are most amenable to these methods. We will use methods that we have already developed (Omics Integrator and PIUMet) to analyze these data and incorporate the results into the predictive models of Aim 1A. The results may also be useful for choosing targets to study at the single-cell level through Multi-Ion Beam Imaging (MIBI) as described in the Friedman and Wherry proposal.

Natural Language Processing. Our goal is to support retrospective analysis of the past medical records. We hypothesize that given a large collection of clinical records, we can explore multiple questions related to disease progression and patients' response to different treatment regimes. The first step towards this goal is to translate information available in clinical records to a structured format (e.g., database) format that can be easily queried. Jointly with the physicians on the team, we will

identify a range of categories which are relevant for their analysis and relevant parts of patient records from which this information can be extracted. Based on our prior work on analyzing breast pathology reports (Yala et al, 2016), we will employ neural sequence-to-sequence models to train underlying extractors. To minimize the number of annotated examples needed for training, we will explore semi-supervised methods recently developed at MIT's NLP group (Zhang et al., 2017). To make system decisions transparent for physicians, we will develop a user interface that connects extraction decision to the input documents, highlighting the portion of the document that rationalizes system prediction. As our experience at MGH shows this design enables physicians to easily identify erroneous predictions. In turn, these corrections are utilized to retrain the system and continuously improve extraction accuracy.

Identifying pathways conferring resistance to natural killer cells: We have developed computational methods that are ideally suited to analyze the data generated by Mitsiades et al. regarding sensitivity to NK cells. In their preliminary work, they screened over 500 cell lines to determine their sensitivity to NK cells. By analyzing existing expression data from these cell lines, they were able to identify groups of genes whose expression correlated with resistance. In parallel, they also conducted a CRISPR screen to determine which genes confer resistance to NK cells in a single, highly sensitive cell line. They discovered little overlap between the genes associated with resistance in these two assays. While this result came as a great surprise to them, it is in fact completely expected. Back in 2009, we published a study comparing genetic and expression screens of this type, showing that they typically have ***less*** overlap than would be expected by chance. In that paper, we reported that genetic screens tend to identify response regulators, whereas mRNA profiling frequently detects effector molecules. We developed a computational method to search for causal pathways linking the master regulators to their downstream effects. In the intervening years, we have continued to expand on these approaches, which should be of immediate benefit in understanding the mechanisms of NK resistance and sensitivity. We will work with Mitsiades et al. to generate testable hypotheses from these models and to examine ways to incorporate other datasets available to us at the Broad Institute for many of the same cell lines.