

Acute Myeloid Leukemia Outcome Prediction Challenge



DREAM 9: An Acute Myeloid Leukemia Prediction Big Data Challenge

David Noren¹, Steven M. Kornblau², Chenyue W. Hu¹, Byron Long¹, Alex Bisberg¹, Raquel Norel³, Kahn Rhrissorrakrai³, Gustavo Stolovitzky³, Amina A. Qutub¹

¹Department of Bioengineering, Rice University, Houston, TX, USA

²Department of Leukemia, M.D. Anderson Cancer Center, Houston, TX, USA

³IBM T.J. Watson Research Center, Yorktown Heights, New York, USA

Demo of Algorithms & Clinical Visualization In 2014, there will be 18,860 new cases of acute myeloid leukemia (AML), and 10,460 deaths from AML. There is urgency in finding better treatments for this type of leukemia, as only about a quarter of the patients diagnosed with AML survive beyond 5 years. The goal of the 2014 DREAM 9 Acute Myeloid Leukemia (AML) Outcome Prediction Challenge is to harness the power of crowd-sourcing to speed the pace of analyzing a high-dimensional proteomics and clinical dataset for AML. The DREAM (Dialog for Reverse Engineering of Assessments & Methods) community consists of diverse computational researchers, biomedical scientists and clinicians who apply their skills to solve a biomedical problem. In this year's DREAM AML Outcome Challenge, participants worldwide compete to develop the best predictive models of AML clinical outcome based on clinical attributes and proteomics. Results of the Challenge include predictive clinical models that surpass current standards; new algorithms to visualize high-dimensional clinical outcome data; and insight into markers of AML and potential new cancer drug targets. In this short demo, we will present on some of the methods behind this crowd-sourced biomedical data challenge.

Methods. In June 2014, DREAM 9 participants were provided a dataset of 190 AML patients seen at M.D. Anderson Cancer Center, and treated with ARA-C therapy. The dataset includes 40 clinical correlates and the expression level of 231 proteins probed by RPPA protein array analysis. This AML dataset provides information that will enable researchers for the first time to link protein signaling with mutation status and cytogenetic categories – offering DREAM Challenge participants the potential to surpass existing methods in identifying drug targets and tailoring therapies for cancer patient subpopulations. Challenge participants were posed three questions based on this data: to predict which AML patients will be primarily resistant to therapy and which patients will have complete remission; to predict remission duration; and to predict overall survival. Baseline predictive models of AML outcome (relapse, remission duration and overall survival duration) were provided participants by the scientific organizers. Each week, teams predict outcomes for 100 representative patients whose outcome was withheld, based on their choice of clinical and proteomic features. These predictions are scored against the test data using two statistical comparisons for each Challenge question. In addition to the development of data

analytics methods, new visualization tools were introduced for the first time in this DREAM Challenge to help participants navigate the clinical data fields and explore patterns in the original protein levels (Figure 1). We will demonstrate the use of these tools on the AML dataset.

Results: The best algorithms are in the process of being developed and scored for the Challenge, which finishes September 15th. The results of the top-scoring algorithms – either separately or averaged – will provide insight into the main factors determining AML outcome, both with and without proteomic data included. Baseline statistical models with no parameter optimization were already provided to the competing teams. These models considered all or some of the data (clinical correlates and RPPA protein levels). The four model types consisted of logistic regression, Random Forest, decision tree with adaptive boosting and support vector machine. Median and mode imputation was used to replace missing patient data values. Area under the ROC (receiver operating characteristic) curve was used to assess the models’ ability to predict patient outcome. In this demo, we will briefly introduce and show the performance of these diverse models on the clinical data.

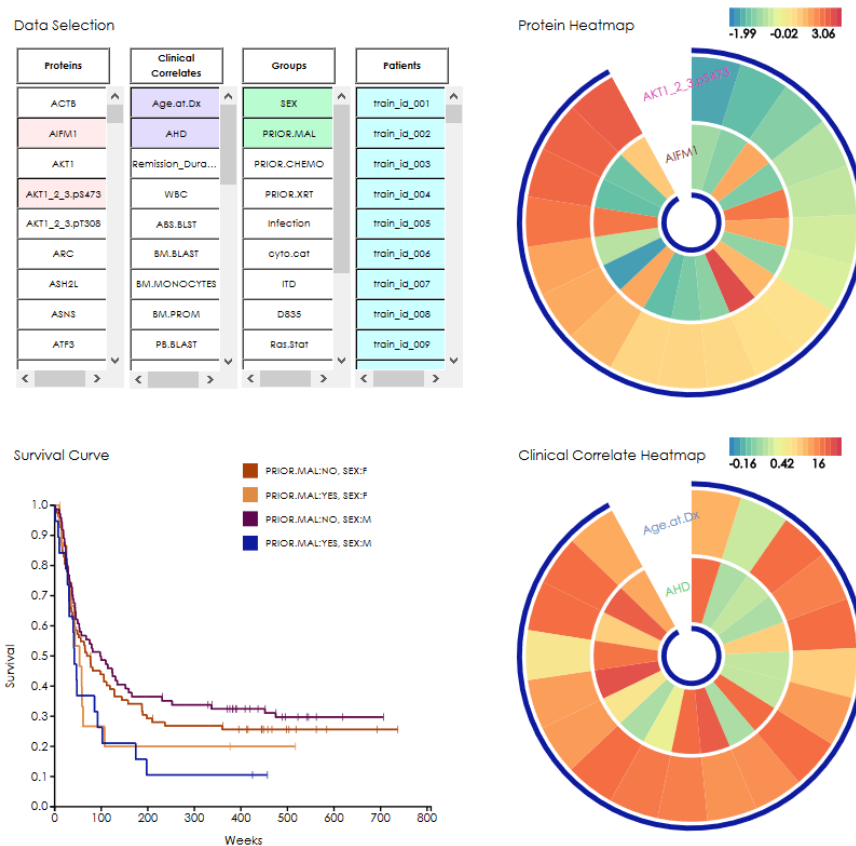


Figure 1. Acute Myeloid Leukemia Outcome Prediction DREAM Challenge Data. New visualization tools provide DREAM participants the leukemia dataset in a web-based format, which users can interactive with. In this demo, we will briefly describe the top performing algorithms used in the DREAM Challenge and showcase patient outcomes and proteomic signatures using the visualization tools. <https://www.synapse.org/#!/Synapse:syn2455683>