CWRU Data Analytics Boot Camp

Spring 2020

Final Project Proposal

Group 2: Frankie Wong, Ben Snyder, Debra Fenty, Chris Bock

Protein Signatures in an Olink Simulated Dataset

Introduction

“Olink Proteomics is a Swedish company (with a US-based laboratory located in Watertown, MA) dedicated to innovation, quality, rigor and transparency, providing outstanding products and services for human protein biomarker discovery. Our groundbreaking Olink panels for precision proteomics help scientists make research decisions more quickly and confidently through robust, multiplex biomarker analysis. Our high quality multiplex immunoassay panels help bring new insights into disease processes, improve disease detection, and contribute to a better understanding of biology.” (source: <http://www.olink.com/>)

Olink uses their proprietary Proximity Extension Assay (PEA) to detect over 1,100 unique proteins in human biological fluids, typically plasma or serum, where the output of the assay is reported in a relative protein concentration unit called NPX. NPX is a log2 transformation of DNA copy number generated from a quantitative polymerase chain reaction (qPCR) measurement. The NPX value is a direct translation of DNA copy number to amount in the biological sample. We will use a simulated dataset that was created by Olink to train other, experienced Data Scientists on how to analyze PEA generated data.

Proposed Project

The simulated study that we have received contains two individual datasets -- that can be analyzed separately or together via a set of ‘bridge’ samples -- and an annotation file for each respective dataset. Each dataset has the same number of unique subjects, but the second dataset contains the ‘bridge’ samples that were analyzed in the first dataset. Each annotation file contains the following data: SampleID, Subject, Treatment, Site, Time, Project (indicating the respective dataset).

Dataset Details:

* 1. Clinical variables
     1. Three timepoints (Baseline; Week 6; Week 12)
     2. Site information (from sites A-E)
     3. Response status (Responder or Non-responder)
  2. The first dataset has 52 subjects, and 156 overall samples
  3. The second dataset has another 52 subjects, but also has 16 ‘bridge’ samples from the first run (and thus has 172 samples)

We have been tasked to identify some or all of the following via analysis of the data:

* Identify what factors or groupings have significant effects on protein expression
  + Site-to-site variations
  + Time points
  + Batch-to-batch variations
* Use Random Forest to find a classifier for response
  + Test and train on one dataset
  + Does the model give favorable results using the second dataset?
* Determine whether the two datasets can be compared together
* (Maybe) Do the two datasets agree with one another in their results.
* (If possible, can you combine the two datasets and look at the analysis again)

Data Sources:

|  |  |  |
| --- | --- | --- |
| **Source** | **Format** | **Description** |
| NPX\_data1 | Excel Sheet | 160 total samples (52 patients x 3 timepoints + 4 controls) |
| NPX\_data2 | Excel Sheet | 176 total samples (52 patients x 3 timepoints + 16 ‘bridge’ samples + 4 controls) |
| NPX\_data1\_info | Excel Sheet | Sample metadata for dataset1 |
| NPX\_data2\_info | Excel Sheet | Sample metadata for dataset2 |

Initial Assignments:

* Frankie Wong: Initial data cleaning and review, web scrape for UniProt ID information
* Ben Snyder: Initial data cleanup and visualizations
* Debra Fenty: Web page design for final data visualizations and “about” pages
* Chris Bock: Project proposal, obtain datasets, setup Trello board, initial machine learning analysis

Tasks will be updated on a Trello board as the project continues