CS171 final project proposal

Background and Motivation

High-throughput screening (HTS) has emerged as one of the most important techniques that connect basic and translational medical research (Persidis, 1998), and advances in HTS technologies (Janzen, 2014) have steadily led to increase in number of studied targets. A multitude of high-throughput screens, which are routinely performed by researches in both academia and in industry generate extensive and complex datasets.

In industry settings, those datasets are handled by commercial (e.g. ActivityBase, Columbus, StarDrop, SCOPE, Genedata Screener) or proprietary data management and analysis software. These software solutions often use certain algorithms for automated assay evaluation and hits annotation. This automation often leads to misinterpretation of data due to the error and noise emerged during experiments.

Project Objectives

In this project, I proposed that human-guided hit selection is better than automated hit selection. By utilizing data visualization and interactions, I will build a tool that better present the data and help screener to select the lead drug candidates.

Data

Both screening datasets and small molecule meta data will be retrieved from public available datasets of ICCB (http://iccb.med.harvard.edu/). A screening data set will consists of the assay readouts of small molecules. It is in 384-well plate format. The meta data of small molecules consists of structure, chemical properties (for example molecular weight), drug like properties (for example PSA) and the structure fingerprint for structure alignment.

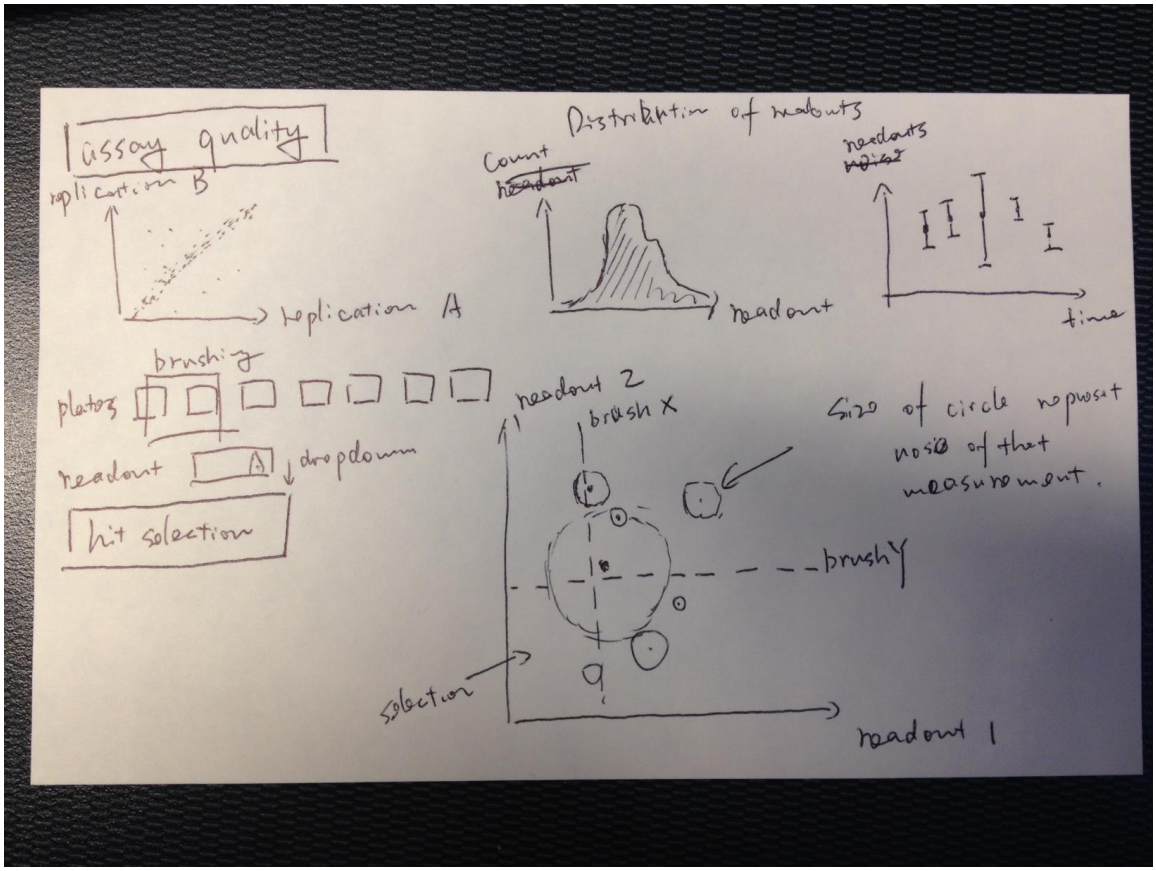
Data processing

To process the screening data, we need to read the datasets in 384-well plate format and convert it to a table format. A row of table will like this:

Library Plate-number Well-number Well-type Readout1 Readout2 Readout3…

Then we will link each screening data to meta data and create a new dataset. This part will be implemented using python scripts.

Visualization



a) Visualization of assay qualities. We will implement visualizations just to evaluate the quality of assay: how much noise of each run, how consistent it is and so on. In this part the visualization will not be interactive.

b) Visualization for hit selections. There are 3 factors for hit selection, the first is the noise of that experiment run; the second is the readout; the third is the property of this compound. We will use interactive visualization to better present those data and enable user to make decision based from various aspects of data.

Must-Have Features

Data wrangling, Visualization of assay qualities, Visualization for hit selections.

Optional Features

Visualization of compound clustering: if there is enough time, we will implement a feature - clustering the molecules base on their properties like structure, charge, library and so on.

Project Schedule

Week 1: Data wrangling
Week 2: Prototypes of visualizations
Week 3: Fine tuning visualizations.
Week 4: Optional Features.