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 academia, researchers usually choose to process their data manually. The processing procedure is often error-prone and lack of consistency. Some researchers use tools like excel or some databases. With these very limited tools, it is very hard to connect chemicals with the assay readout. For example, after finding a chemical that has decent activity, researchers has to use other tools to retrieve the information about this chemical like logp, molecular weight or structure.

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

Inspired by the idea of human-based computation, I propose that human-guided hit selection is better than automated hit selection. By utilizing data visualization tools and computing tools, I will build a web-based system that can fully displayed each important aspect of screening dataset, allow researcher to play with their data interactively and eventually help screener to better select active chemicals.

A high-throughput screening dataset usually consists of several parts: A) library, id, plate number and well number. These data links a position of an assay-plate to a certain chemical. B) Readouts. Readout measures the biological activity of a certain chemical. Readout can also be other properties of chemicals. C) Chemical information. This part usually consists of chemical structures, molecular weight, logp, links to chemical databases.

The first goal of this project is to enable researchers to directly connect biological activities to chemical properties like structure or logp. An experience chemist can easily tell if a chemical is promising or not. This feature will enable researchers to make better call on whether to choose a chemical for next step.

The second goal of this project is to present the dataset: the distribution of one type of readouts, correlation between 2 readouts, the deviation of positive controls and negative controls. These presentations are important for researcher to know their data. For example, in the distribution, if all chemicals obey Gaussian distribution, it means the experiments was good, otherwise, it is an indication of systems errors. Another example is that if the deviations of controls are too large, it decreases the probability of finding active chemicals from that dataset (P value increases).

The third goal of this project, if there is enough time, is to cluster chemicals based on their similarities. For example, if one chemical is active, it will be useful for the researcher to know how similar chemical behave in their experiments. This often provides very insightful information about the structure-activity-relation (SAR) of a chemical.

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.

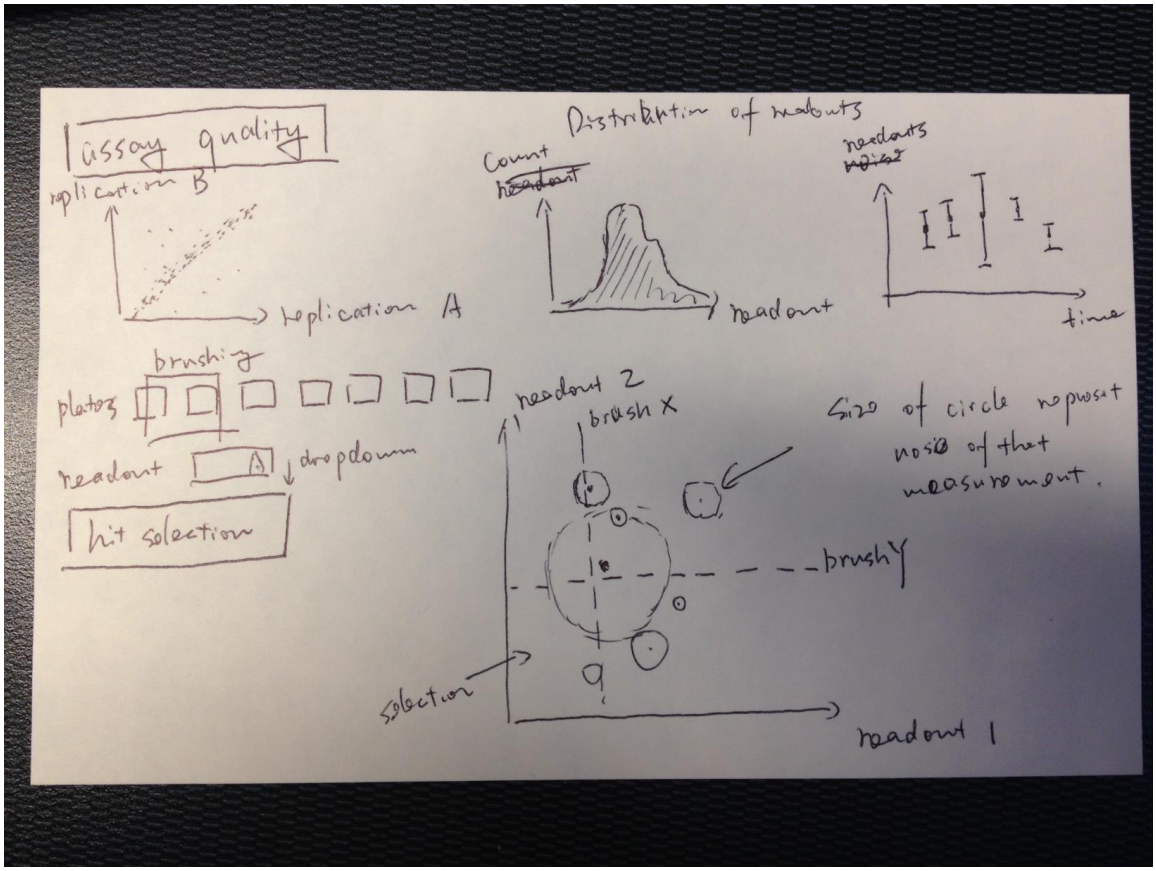
Data processing

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) Assay plates selector. Draw rectangles that represent each assay plates in the dataset. Within each plate, also present the distribution of positive controls and negative controls (boxplot). Allow users to select and unselect plates by clicking.

b) Distribution of one readout. I will use D3 area to present the distribution of readouts. Brushing feature will also be included for user to selection a readout domain. And this selection will be reflected in Scatter plot and Chemical Cloud.

c) Scatter plot of assay readouts. The scatter plot will include 2 drop down boxes allow users to selection which 2 readouts to use. Each chemicals will be color/sized coded for their chemical properties like molecular weight or logp.

d) Chemical Cloud. It is a svg that displays all selected chemical structures. The size of a chemical encodes its activity.

e) Cluster. It is similar to Chemical Cloud but with links the represent similarities.

Must-Have Features

Data wrangling
Plate selector
Distribution plot
Scatter plot
Chemical Cloud
Interactive Part: Plate selector and Brushing of Distribution plot.

Optional Features

Visualization of compound clustering: if there is enough time, we will implement a feature - clustering the molecules base on structure. The clustering is based on a fast similarity search algorithm using chemical fingerprints.

Project Schedule

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