Pharming Detection Spring 2019

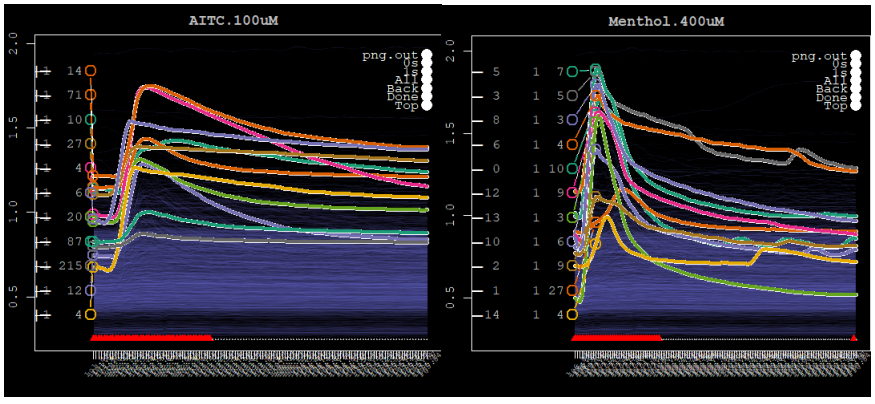
Lee S. Leavitt

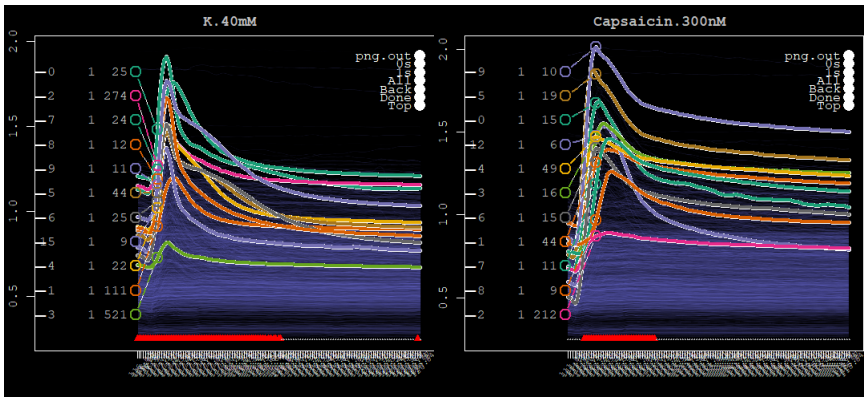
April 28, 2019

1 Introduction

Constellation Pharmacology is an assay that looks at a dissociated tissue of interest. In this case the dorsal root ganglion (DRG), the an organ of the body responsible for all physical sensations (heat, cold, pain, proprioception, light touch, heavy touch). The tissue is;

1. Gently broken up until the cells in this tissue (neurons) are isolated from one another.
2. The cells, ≈ 5000, are plated so that each cell is seperate from one another.
3. The are loaded with a dye fura2, a dye which allows for the measure of calcium into and out of the cell.
4. Varieties of natural products are applied to the cell, which produce unique response in subsets of the neurons while a video is collected. The natural products applied are;
   (a) Allyl isothiocynate (AITC) the main compenent in mustard oil, and wasabi.
   (b) Menthol, the compound found in peppermint which evokes a cooling sensations.
   (c) Capsaicin, the spicy component of chili peppers.
   (d) Potassium (K^+) at a concetration of 40mM, which activates a variety of voltage gated ion channels on the surface of these cells.





Then using an open source software, CellProfiler, cells are identified based on whether they are loaded with the fura2 dye. Each cell is considered a region of interest (ROI). Within each frame, each ROIs mean intensity is collected. Each intenisty is compiled into a n, by d matrix where $n = time$, and $d = cell$. The traces then undergo a significat cleaning step;

1. Padding: 3 points are added to the end of each trace since the last natural product application does not allow the trace baseline to return to the bottom (for example Capsaicin), three points are added allowing a better representation of the trace.

2. Despiking and smoothing: The traces vary in the ammount of general noise. In addition to this, since this assays uses fluorescent imaging many dust particles are very fluorescent. To obtian an accurate reading of peak height these aberant values are taken out.

3. Baseline Correction: The baseline is then corrected. This means the trace is corrected to correct for upward baseline drift, moves everything to a zero base.

4. Normalization: K 40mM and Capsaicin induce a maximal response. This maximal response is then considered 1, where the minimum value is now considered 0.

The scientists then use a in house application to correct an automatic binary scoring. The images above cooresponding to each natural product application shows the diversity of responses. This process is aided by hierarchal clustering. This techniques allows to scientist to view and obttain a view of all cells, while correcting the scoring.

Statistics are generated for each natural product application. The statistics below show the statistics generated for the feature space.

1. snr: Signal to Noise Ratio provide a measure of peak sharpness. This statistic has proven to be most useful for menthol, and least significant for Capsaicin since the Capsaicin response does not return to baseline.

2. tot: Sum of all point within a window region of 4 minutes. General measure of area.

3. max: maximun value within the region of application.

4. ph.a.r: Peak height to area ratio.

5. wm: where does the maximun value occur.

2 Data Processing

This was the most time consuming and difficult aspect of the project. Since we have not dedicated time to generating a database, each experiment is stored in an R list of dataframes, and lists. In addition to this each scientist spells each natural product differently, applies each natural product different orders. Significant time went into building an flexible application to allow for growing a significant feature and label space. After a significant ammount of data wrangling 58,301 instances were collected.

3 Support Vector Machine

The goal of this experiment was to determine whether a response could be detected at a significant rate. Initially the e1071 package was used to produce the SVM. This package produced fantastic results. After this, radial svm was attempted to see if this could further improve the model. Suprsingly radial SVM takes significantly longer than linear. Where the majority of the time comes from is for the tuning of the C and σ value. Due to this extensive ammount of time required to tune the radial SVM, the *caret* package was invoked and used for tuning. What is most nice about the caret package is the ability to paralleize the code. with e1071 my processors utilization was $\approx 20\%$. After parallelization my processors were at 100% and was able to cross validate.

1. Linear vs Radial SVM: The first experimentor included a single experimentors data which is 30,000 instances. For linear SVM it was cross Validated (10 fold, repeated 4 times), and 12 C values were tested for each natural product.

SVM Method	AITC	Menthol	Capsaicin	K.40mM
linear	92%	92.94%	97.8%	97.9%
radial	93%	93.6%	98%	98.5%

2. The W vector produced from the linear SVM also provides significant insight into what feature has the greatest effect on the fit. As can be observed below the most important feature for defining the hyperplane is the signal to noise ratio (snr), whereas a majority of the other features do not help in defining the hyperplane.

”	snr	tot	max	ph.a.r	wm
AITC	4.49	0.42	-0.42	-0.33	-0.17
Menthol	2.93	0.04	0.06	-0.52	-0.26
Capsaicin	5.41	-0.07	0.39	-0.06	-0.56
K.40mM	5.13	-0.55	0.77	-0.00	-0.81

4 Future Directions

1. It would be nice to move our data into a database format. This would allows for a much faster data collection and testing method. Currently each experiment is $\approx 500 - 1,500MB$ Only a subset of this information needs to be accessed as well, thus a significant ammount fo time was spend troublshooting the loading features and label space into the support vector machine.

2. Auditing: The next series of our application that needs to be developed is a way to view the support vectors. These cells should have a some significant value for helping us tune the statistics collected, along with improving the data cleaning protocol. I have already developed a UI within the R environment while works in rapid way for displaying the data.

3. Our data scientist has developed a way of obtianing more advanced statsitics by fitting polynomials to the trace shapes. This would be the next set of experiments i would like to run, although this will take a very long time to complete, not to mention the painstakinly difficult process of organizing our messy data.

4. Now that SVM's have proven to been very useful for us during these simple experiment, more advanced effect detection would be interesting to observe and continue development on. In the figure below is an example of the more advanced effects our group would like to detect. a compound pl14a.10uM is applied, and this application causes the cells to have an amplified response to this compound. THis effect is repeated three times. Unfortuneatly this example is the cleanest as others have significantly more suddle effects.