Data story

Elena Tortosa

3/8/2019

Unraveling Protein-Signaling Networks from Multiparameter Single-Cell Data

Introduction

Cells, as part of complex multicellular organisms, need to communicate among themselves and work together to make an organism alive. Communication between cells triggers signaling cascades inside the cell that regulate specific cellular functions such as cell division, differentiation, migration or death. This communication is initiated with a stimulus, an extracellular messenger that binds to a receptor located in the external boundary of the cell that initiates intracellular cascades, activating a series of targets, which in turn will lead to the desired cellular response. Events in a signaling cascade occur in a series, much like a current flow in a river. Interactions that occur before a certain point are defined as upstream events whereas events after that point are called downstream events.

The main components of the intracellular cascades are the proteins (biomolecules that perform a vast array of functions within the cell). Interestingly, signaling cascades induce protein modifications such as the addition of a phosphate group (named as protein phosphorylation). Phosphorylation is one of the main events that happen in the transduction of the signal cascade and this modification is associated with the activation or inactivation of the protein, leading to the activation of downstream targets or to the complete block of the downstream cascade.

Signaling cascades can be very complicated. Among others, complicating elements in these processes are that 1) a single pathway can branch off toward different endpoints and 2) signals from two or more different cell-surface receptors merge to activate the same response in the cell. Although the identification of signaling cascades is being quite challenging, the ability to manipulate them by using activators and inhibitors of certain components have been particularly useful. Thus, inhibition of a certain protein blocks the cascade at certain point. Upstream components won't be affected whereas downstream effector will be also inhibited.

Deciphering signaling cascades is essential to properly understand cellular responses and their potential dysregulation in disease. By mapping the network of connections between proteins and grouping proteins into different pathways, researches are trying to identify disease-associated proteins and analyze how they are connected to the disease. For pharmaceutical companies and health services as hospitals, unraveling these signaling cascades is an important challenge that has the potential to provide clinically actionable insights for disease diagnosis, prognosis and treatment.

It's important to keep in mind that an average human body contains approximately 37.2 trillion cells and one cell type is estimated to contain around 100,000 or more different proteins. Additionally, as mentioned before, proteins rarely work alone and they establish partnerships with other proteins to perform their many functions, establishing really dense and complex networks within the cell. These numbers give us an idea about the dimensions of the challenge of trying to identify protein signaling networks.

"If you break a gene coding a protein that goes into a complex, then that complex is dysfunctional in some way and that gives rise to a condition or disease," says Edward Marcotte, a systems biologist at the University of Texas at Austin.

What important fields and information does the data set have?

This project relies on the simultaneous measurement of multiple phosphorylated protein /phospholipid components in thousands of individual primary human immune system cells after a series of stimulatory cues and inhibitory interventions. The data obtained will be used to profile the effects of each condition on the intracellular signaling network and to elucidate different pathways and the ordering of connections between components (upstream vs downstream components).

The technique used in the measurements, the intracellular multicolor flow cytometry, allows more quantitative simultaneous observations of multiple proteins in many thousands of individual cells. Because each cell is treated as an independent observation, flow cytometric data provide a statistically large sample that enable us to predict pathway structure.

The authors measured 11 phosphorylated proteins/ phospholipids (PKC, PKA, P38, pjnk, praf, pmek, Erk , pAkt, pPLC-gamma, PIP2 and PIP3) in nine different conditions (roughly 700 to 900 single cell measurements).

Each independent sample in this data set consists of quantitative amounts of each of the 11 phosphorylated molecules, simultaneously measured from single cells.

What are the limitations i.e. what are some questions that you cannot answer with this data set?

The main limitation is that, from the estimated 100,000 or more different proteins that a single cell can contain, only 11 phosphorylated proteins/phospholipids have been measured in this dataset. Therefore, the information we get from each treatment we have applied to the cells is very limited. Additionally, measurements are done in just one cell type so predictions done here can be hardly apply to other cell types.

What kind of cleaning and wrangling did you need to do?

All data was provided in 14 different excel sheets. They were named with a number (df#) followed by the name of the perturbation + treatment each table contains and saved as .csv. There were no missing data and I just had to unify column names.

Measurements are obtained from two different perturbations ("general perturbation" : GP1 and GP2) so I added a new column called GP to classify the data depending on the general perturbation is applied: GP = 1 for GP1 and GP = 2 for GP2.

At the same time, these perturbations are combined with different treatments (protein activators or inhibitors) so I added a new column called "treatment_num" to each table to classify the data depending on the treatment applied :

0 <- no treatment, 1 <- Akt_inh1, 2 <- PKC_inh, 3 <- PIP2_inh, 4 <- MEK_inh, 5 <- Akt_inh2, 6 <- PKC_act and 7 <- PKA_act.

I reorder the columns to have the general perturbation and treatment names first, and after that all the measurements done.

Any preliminary exploration you've performed and your initial findings

I've checked characteristic of my dataset as dimensions and data type for each attribute, and I've calculated basic statistics as median, IQR and skewness. All the skew values are positive suggesting that my data is skew to the right and does not follow a normal distribution.

I've done some histograms for some of the proteins in all conditions in order to visualize the distribution of the data and, using the Shapiro test, I've checked if the data follows a normal distribution. I've grouped the data by treatment (treatment_num) and protein (variable_name) and applied the test to each single group. Since the p Values are less that the significant level of 0.05, we can reject the null hypothesis and confirm that the tested sample do not follow a normal distribution.

I've applied the Kruskal-Wallis test as a non-parametric method for testing whether samples originate from the same distribution. I tested if, for each single protein, there are statistically significant differences between the different treatments. As the p-values are less than the significance level 0.05, we can conclude that there are significant differences between the treatment groups for all the proteins.

I also calculated pairwise multiple comparisons between different treatments, for each single protein, according to the Dunn's-Test, but so far it's not working (probably, due to problems with the code).

I've also looked at correlation between the different proteins among the different treatments and visualized these correlations with a matrix of correlation plots.

In order to simplify the data, I've calculated the median for each protein and for each treatment, and I've performed a row-wise normalization of the proteins, so that we normalize the proteins across each experiment and we can compare experiments.

I've plotted the medians for each protein in each condition using a line plot. X axis represent all treatments (from 0 to 7) and Y axis represent the relative amount of protein. Each line represents the median of each single protein for each treatment.

I've also clustered the proteins and treatments based on their medians. First, I've calculated distance between experiments and protein in rows and clustered the data based on these distances. Afterwards, I've checked the clustering by plotting dendrograms. Finally, I've plotted a heatmap with all the medians of all the proteins for each condition, together with the dendrograms. From this heatmap, we can see that, although treatments affect differentially to each protein, treatment 3 (inhibition of PIP2) seems to increase the levels of phosphorylation of most of the proteins. Additionally, proteins like PKA and PIP3 are the ones that change more among all the treatments.

I've done few boxplots for each protein in each condition, comparing the two perturbations (GP1 and GP2), and we can see that there are not many differences between these two perturbations.

Based on these findings, what approach are you going to take? How has your approach changed from what you initially proposed, if applicable?

I would like to explore in more detail how correlations between the different proteins change among the treatments.