Exploring Similarities and differences between molecular aberrations between tumor types

By: Khan Inan

Topic/Question:

The main question I want to explore looking at this dataset of tumor gene expression is whether or not there are any themes or commonalities present across the 12 different tumor types. Exploring this question can be very important in possibly determining if any currently existing cancer treatments can be extended to other kinds tumors that are genetically similar enough to respond positively.

Hypothesis:

My hypothesis is that after I analyze this dataset, I will certainly find categories of tumors and even specific sample cases that are similar enough to be looked into further. Obviously, I dont have enough information to make a statement on how these similar cases would respond to the same treatment, but it still is useful information because possible trends and themes can be discovered across the cancer/tumor samples and the mutations/aberrations that causes them

Data Sources:

The biggest source of information that will be useful to me, besides the dataset itself, is definitely the Pan-cancer analysis project. The goal of this project and its associated team is to assemble a useful and coherent standard for TCGA data across all the tumor types. Their analysis of the data is online for reference as well and the papers/interpretations of this project will be very useful in my study

Methods:

I am absolutely certain that I will incorporate pearson correlation and distance analysis into my project somehow. The method that I am leaning towards is separating the data set into groups based on the 12 different kinds of tumors specified by the TCGA, and from there I would need to conduct my analysis and calculation and determine the groups that are the most similar and find out how or why this is the case.

Expected Results:

The biggest preconceived notion I have going into this project is that cancers that are from the same general locations on the human body will have the most similar gene expressions and molecular aberrations. This however, is just an assumption and I am sure I will find some surprising results of tumor categories that seem like they should not be related, but somehow are in accordance with the data. The main goal at the end of the day to discover themes across cancer types and sub-types and break down boundaries that are currently in place within cancer therapies and treatments. Essentially, I am hoping that my project and the larger Pan-Cancer project is able to affect clinical decision making in a positive way.

Potential problems and solutions:

The biggest problem that I fear running into is that none of the cancer/tumor groups will have any significant amount of similarity. In that case I would need to determine a threshold that would make sense for determining similarly, for use within the pearson correlation and distance calculations. On the other hand, if all the groups of cancers/tumors are more similar to each other then I expected, the same thing would need to be done and I would only select the most similar groups and purge the rest. The biggest challenge in my project would definitely be splitting the data into the separate categories, especially if the dataset is unorganized and mixed together. The Pan-Cancer project used numerous different metrics to split their data, however the data set that I have focuses on gene expression only

The image below is an image that I found to be interesting in regards to the workflow of the Pan-Cancer analysis project