TEAM 22

Topic Question

Are there any diseases that correlate more often to cancerous genes than non-cancerous ones? How do the chemicals associated with these specific diseases affect the activity of the genes?

While the basics of cancer are well known, details regarding the spread of the disease are still under investigation. We want to investigate some of the effects of cancerous genes in a molecular level. The human body is an overwhelmingly complex dynamical system, which makes our task difficult. Because there are no isolated systems, every molecular component interacts with a myriad of other proteins and chemicals. By analyzing some of these interactions, we can gain a better insight into how cancerous genes affect the rest of this network. To answer this question we looked at the *tcga*, *toxicogenomics_chemicals* and *toxicogenomics_diseases* data sets, which are essential to our analysis.

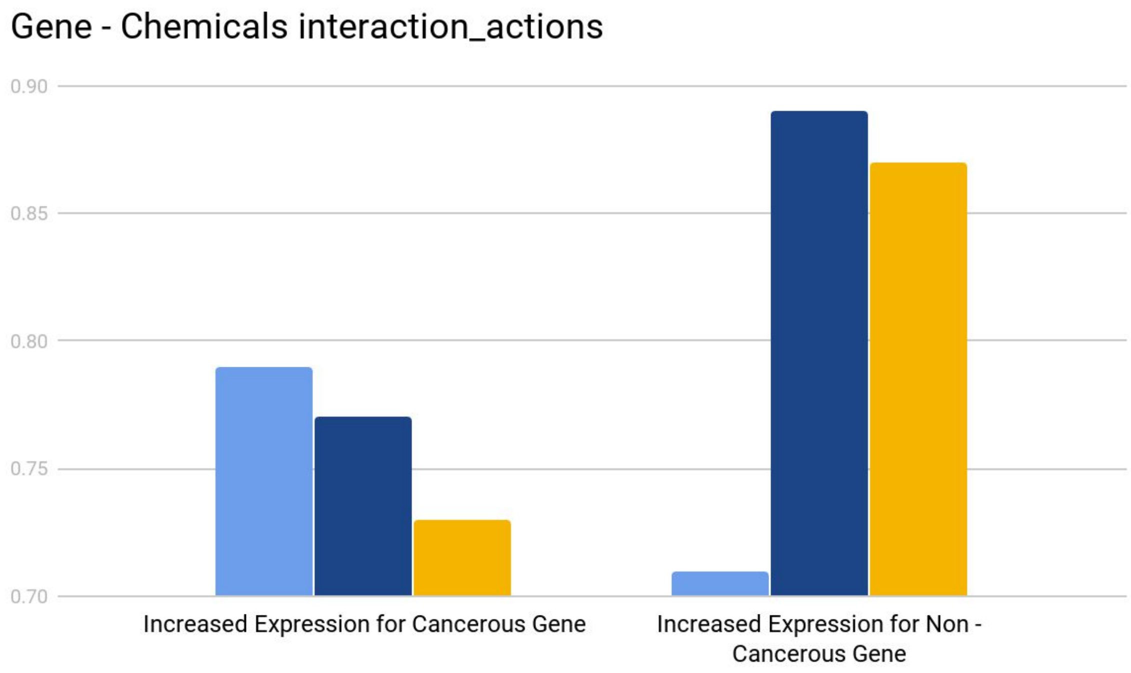
Non-technical Executive Summary

Our key finding is identifying the diseases that most often correlate with cancerous genes and non-cancerous genes respectively. These are represented in the table below.

Cancerous	*Non-cancerous*
Necrosis	Inflammation
Inflammation	Necrosis
Drug-Induced Liver Injury	Drug-Induced Liver Injury
Neoplasm Invasiveness	Edema
Kidney Diseases	Kidney Diseases
Edema	Neoplasm Invasiveness
Weight loss	Weight Loss
Prenatal Exposure Delayed Effects	Hyperplasia
Hyperplasia	Long QT Syndrome
Hypertension	Pain

These results show that we cannot guess whether genes are cancerous or not based on the other diseases that they correlate with. This might be the case because both cancerous and noncancerous genes play key roles in cellular development and thus cause similar diseases.

Further we look at the chemicals associated with these diseases and how they, in turn, affect the cancerous genes. The interaction actions for the top 3 cancerous genes and top 3 non-cancerous genes are shown below.



While most of the interactions increase gene expression we notice that the non-cancerous interactions have have higher levels of gene expressions compared to cancerous ones. Thus there are more gene suppression interactions (percentage-wise) between cancerous genes and their disease-specific chemicals.

There are several aspects we need to know in order to understand this result. By definition, a cancerous cell is one that has undergone a mutation and spreads faster than other benign ones. While we might expect gene expression to be significantly higher in cancerous genes, we can understand the higher suppression against cancerous genes as an immune response to the malignant gene. Thus, the interactions with various chemicals could infer the recognition of the malignant nature of the cancerous genes in opposed to the non-cancerous one.

Technical Executive Summary

Our approach was broken down into several steps:

1. Exploratory data analysis

In the exploratory data analysis step, we assessed the provided *toxicogenomics_diseases, TCGA,* and *toxicogenomics_chemicals* datasets. We looked especially for missing values and computed distributions as well as summary statistics for each of the three datasets.

2. Analyzing toxicogenomics chemicals with cancerous gene IDs

Next, we merged the *toxicogenomics_chemicals* dataframe with the *TCGA* dataframe on the gene_id. The resulting dataframe is the subset of *toxicogenomics_chemicals* that has a cancerous gene_id. From this result, we were able to compare the numbers and proportions of the interaction_actions before and after we considered the subset of those chemicals that are associated with cancerous genes.

3. Analyzing toxicogenomics diseases with cancerous gene IDs

Furthermore, we merged the *toxicogenomics_diseases* dataframe with the *TCGA* dataframe on the gene_id. The resulting dataframe is the subset of *toxicogenomics_diseases* that has a cancerous gene_id. We were able to find the unique diseases that occurred most frequently in this resulting dataframe (the number of gene ids associated with each disease). We were also able to find the diseases, that carry a cancerous gene, with the highest inference scores.

4. Analyzing the toxicogenomics chemicals with chemical names that are part of the toxicogenomics diseases with cancerous gene IDs

In this step, we took the inference chemical values from the joined *toxicogenomics_diseases* dataframe and merged it with the *toxicogenomics_chemicals* dataframe such that the resulting dataframe is a subset of the *toxicogenomics_chemicals* dataframe. This subset is all the rows that have a cancerous gene and matches with a disease in the *toxicogenomics_chemicals* dataframe.

5. Analyzing the distribution of interaction_actions for the most frequent genes in the result from the above step

We look at the resulting merged dataframe from the last step and seek out the three most occuring gene ids. We then look at, for each of these gene ids, the distribution of the interaction_actions.

6. Analyzing the subset of the *toxicogenomics_diseases* table that does not have a cancerous gene

We now merge the genes that are not cancerous with the *toxicogenomics_diseases* dataframe. In this result, we can analyze the diseases, that do not carry a cancerous gene, with the highest inference scores.

7. Analyzing the toxicogenomics chemicals with chemical names that are part of the toxicogenomics diseases without cancerous gene IDs
8. Analyzing the distribution of interaction_actions for the most frequent genes in the result from the above step

In steps 7 and 8, we repeat a similar analysis on the merged dataframe result as steps 4 and 5. Specifically, we look at the three most occuring gene ids in the merged dataframe result, and we look at the distribution of interaction_actions (this time for the toxicogenomics chemicals with chemical names that are part of the toxicogenomics diseases without cancerous gene IDs).