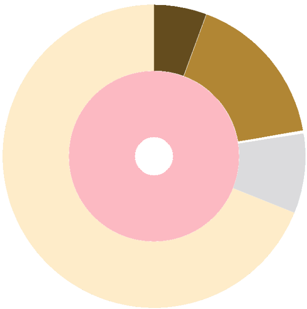

Visualising Cancer

with a focus on breast cancer

Kirtan Padh
Luis Medina
Sharbatanu Chatterjee

Processbook for COM-480 : Data Visualisation

22 December 2017

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Contents

List of Figures

List of Tables

Abstract

Cancer has been the second leading cause of death globally for a long time now, and was responsible for 8.8 million deaths in 2015. Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute estimates that approximately 35.5% of men and women will be diagnosed with cancer of any site at some point during their lifetime, based on 2012-2014 data. This makes it extremely important and interesting to develop tools for early detection of cancer.

We attempt to use the power of data visualization and the growth in the accessibility of genomic data available in the last few years to aid the detection of breast cancer.

Acknowledgements

We'd like to thank Dr. Kirrel Benzi for introducing us to a lot of different exciting ways that we can learn and do data visualizations, and also for allowing us to work on this project.

We'd like to thank also Rachel Jeitziner for explaining patiently to us all the theory behing this visualization and providing extremely helpful guidance on what we can do to make this tool useful to cancer researchers like herself. We look forward to continued collaboration with her in developing this tool to make it better.

1 Overview

This project envisages a clinic of the future, led on the one hand by data scientists with skills in data analysis and data visualisation and on the other hand by doctors and clinicians with research scientists providing the valuable back-end. We show, in our chosen example of breast cancer (Breast Invasive Carcinoma) a small peek at what is possible in the present and show a roadmap for future bettering of the human condition.

2 Motivation

Nearly 1 in 6 deaths in the world is due to cancer, making it a major problem for human potential. Traditional therapies for cancer depend often on how early in the stage of cancer the diagnostics can be done. Also, more often that not, the effectiveness of the cancer treatment depends on the particular subtype of cancer as well as the genetic makeup of the cancer patient. With the growth of large data sets and methods of analysis in biology, it is becoming increasingly easy to develop and use data analysis and visualisation methods to extract relevant insight from them which would help doctors and clinicians tackle the problem of early and personalised diagnosis as well as treatment.

The recent example of Dr. Shirley Pepke[1] shows her personal struggle against cancer and the need to '*tailor her treatment to her particular cancer*' and served as a motivating factor to explore this further. For Dr. Pepke, it was a bit less difficult to access the biological and data science research community than it is for people in regions of the world where access to information, technology and healthcare is not universal. Approximately 70% of deaths from cancer occur in low- and middle-income countries. The possibility to mitigate troubles in such places served as a strong motivating factor.

3 Target Audience

We believe that doctors and clinicians would be able to use our visualisation of data from a database of previous patients and genomes (in our case, from The Cancer Genomic Atlas (TCGA)) along with a visualization of genes thought to be

[1] https://cancergenome.nih.gov/researchhighlights/tcgainaction/researcher-studies-own-cancer

playing a role in the development of said cancer to compare and pinpoint to more helpful information about the patient. We also think the research community will be able to think more about metrics of distance between patients by looking at our visualisation, hence coming up with better ways to help patients. It may be a long shot but Dr. Pepke's example shows one might find out correlated genes and personalise healthcare. Quoting her, '*Just because something is not informative for most patients, does not mean it won't be for an individual patient...Combining patient-specific features with database information can make a difference*'. We wish that all unfortunate patients of cancer be included in our intended target audience and receive such individualised care.

4 Dataset

The dataset that we use is the publicly available dataset 'The Cancer Genome Atlas' (TCGA), thanks to the information provided by Rachel Jeitziner, Doctoral Student, EPFL. The Cancer Genome Atlas (TCGA) is a collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) that has generated comprehensive, multi-dimensional maps of the key genomic changes in 33 types of cancer. We will be focusing on one type of cancer, Breast Invasive Carcinoma.

4.1 Science behind the dataset

The word 'cancer genomics' refers to the study of sequences of 'letters' or nucleotides that make up the human DNA. The human DNA constitutes the genetic identity of an individual person and although it varies from person to person, there is a vast amount of similarity within the human population. Small changes in the sequences of nucleotides causes a cascading effect on the organism that carries this genetic material. It may mess up with the proteins created as an end product of reading the code written in the letters represented by the nucleotide. This may sometimes lead to cancer and hence scientists involved with the TCGA have tried to successfully obtain enough data about such patients to do a meaningful data analysis.

4.2 Description of the dataset

- The first raw file used consisted of the `BRCA.rnaseqv.txt` file which is publicly available at the website of TCGA, consisting of RNA expression rates (which is another way of checking the expressions of particular genes), 298.4 MB in size.

- The second is a json file used consisted of the details of the patients which were collected from the clinics, called `clinical.project-TCGA-BRCA.json` which is about 2 MB.
- The third is a file consisting of the distances between the patients as calculated by Rachel et. al and is 20.8 MB in size

4.3 Previous Work

On this dataset, the TCGA website itself mentioned the visualizations worked on at Xena. The kind of work at Xena depended on knowing specifically which kind of genes and which kinds of cancers to look at and required a lot of specific domain expertise to understand the meaning of the data. The screenshot in Figure 1 shows the way one can select the number of samples based on some of the features and we found a lack of intuition, functionality and aesthetic appeal in here.

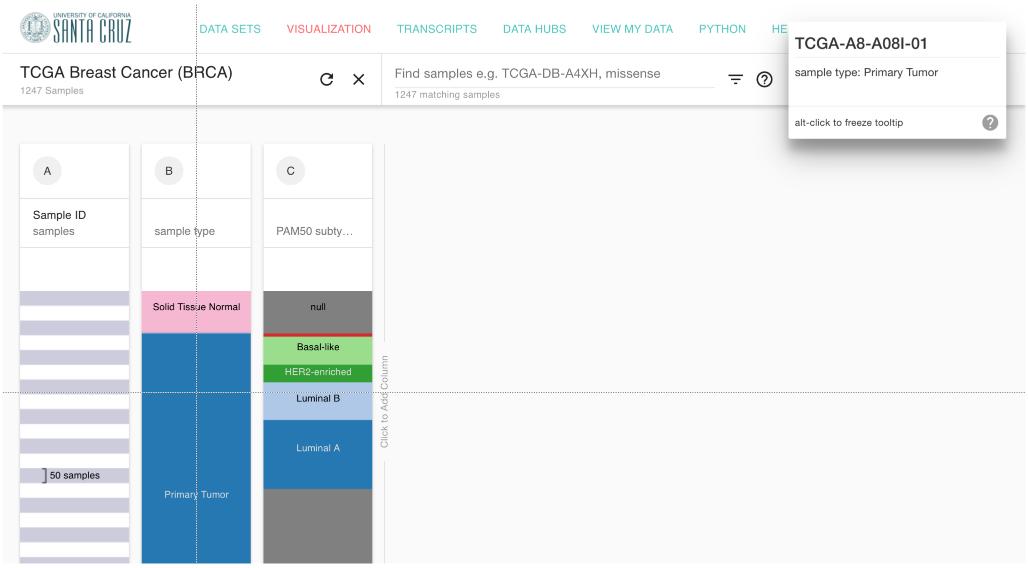

Figure 1: Visualisation in the Xena Browser

We were inspired by the examples shown in class about graphs, especially where we could zoom in and out, select edges and filter out those which were deemed necessary to give a spatial clue about the positions and importances of each character in the graph under consideration. That led us to first think of our initial proposal of making a zoomable graph which was filterable and which connected itself with graphs of genetic correlations, thus making the entire point of finding out patterns given a new patient's data, easy to view and navigate. But before that we will speak about the initial data cleaning.

4.4 Cleaning the data

We used Python code to clean our dataset to make it proper and readable by our `javascript` code. The cleaning consisted of :

- Removing data of the patients that are not in all of the three datasets
- Getting the references for the number of the patients from the dataset having information about them and connecting them across the three datasets
- Dividing the number of patients based on those who have a primary solid tumor, those who have a metastatic tumor and the people who do not have a cancerous cell but a solid normal tissue.

After cleaning the dataset in this way, we created a new single `DataFrame` in `pandas` by combining the information contained in the datasets, so that it would be easily used by the code that we write to visualize the dependence between the patients. We also have another `DataFrame` which consists of the list of all the patients and their expression levels of all the genes in the genome, which may be involved with the cancer they have, or not. We also calculated the correlation matrix between each gene deemed to be important by using this dataset. The total number of identified genes was 20531 and we thus had a pretty large correlation matrix.

4.5 Data Exploration

We did some simple data analysis to have a look at what the data looks like. Like we mentioned in the previous subsection, we need to clean the data and to reformat the original datasets because it was difficult to read them directly by using python. The main files that we have are the one that contains the correlation distance of the 1093 patients and the other about the demographic and clinical data of each of the 1093 patients. Since most of the clinical information contained in the clinical dataset is regarding the morphology and type of cancer, and since we know that we were dealing with breast cancer, we decided to work only with the demographic data in which we tried to see the proportion of the people with specific demographic information and this is actually done by the sunburst that we created.

5 Visualisation

Having seen the various aspects of the data, we decided on several ways to see how we can render the graphs. We will mention our initial decisions and why we decided to go against it in the following points :

- We decided to have a way to show the filtering amongst the different features of the patient without the graph, just to give an idea about the scales of data points that we have. We thought that the graph would be too overwhelming for that, so dropped the idea of using that as the first point of entry into exploring the dataset.
- We hence focused on using a sunburst for the interactive visualisation. We used the `Dexjs` library but found it to not have as much of an interaction as we would have liked, so we dropped the idea of using that as well after trying it out.
- For our graphs, we tried using `graph-tool` but having interactivity there was difficult. So we decided against using it. We used Cytoscape instead.
- We had initially thought of clustering together nodes by the filters they fall under, but not only did their positioning in space seem a bit arbitrary and not well defined, but also technically, it was difficult for us to implement.
- Dynamically recalculating and reconnecting the graph was difficult, it being another reason that clustering was discarded. The graph has 1000 nodes and several thousand edges and was therefore difficult to render as a collapsible graph in a meaningful layout, as in 2. Furthermore, the introduction of bezier curves made the computations extremely cumbersome considering the large number of nodes and edges.

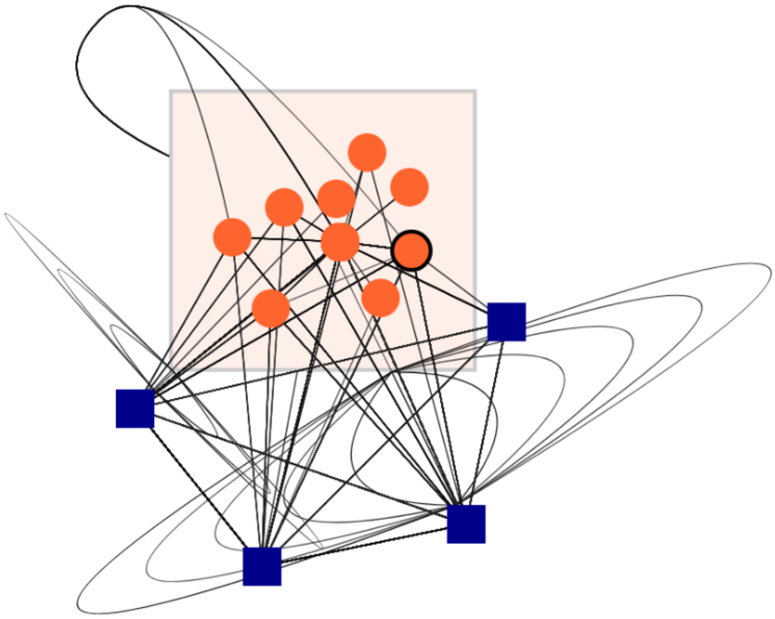

Figure 2: Bad collapsible visualization

- The entire correlation matrix of genes would be taking a long time to calculate and render. We tried with a smaller network of genes, but none of these genes had enough expression to allow us to work that into the graph visualisation. The correlation matrix turned out to have a size of 8GB, and the resulting graph would have had 25000 nodes and millions of edges. This made it a technological challenge for us and we decided against visualising it in the end.

During the actual visualisation of the graph, we thought a lot about visually appealing arrangements. The first thing that we did is put the nodes in a quadrilateral order, after putting a threshold so that not all the edges are described. This gave some idea about what kind of elements were not grouped with each other, and what kind were isolated, as shown in figure 3. We thought that there were too many isolated nodes for this to be useful, and the arrangement was a bit rigid, so we tried a second one, as shown in figure 4. These figures give us an idea about the importance of the threshold in deciding what the visualisation can represent.

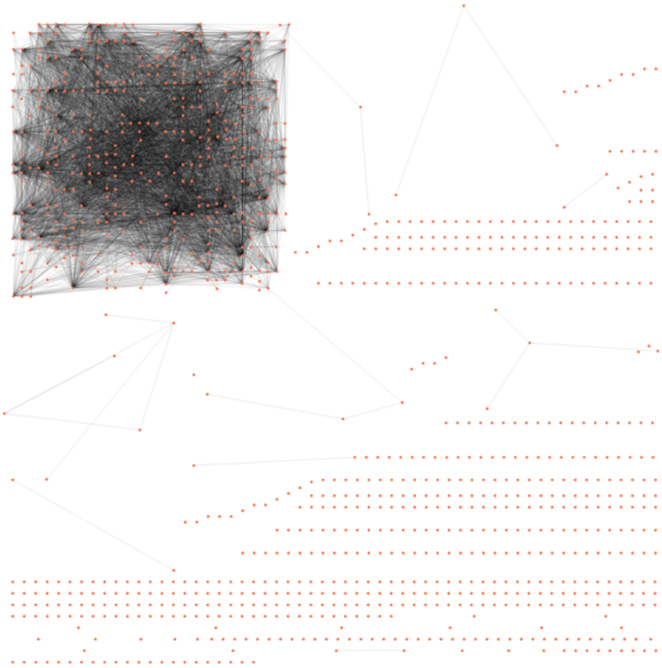

Figure 3: Bad Visualisation - 1

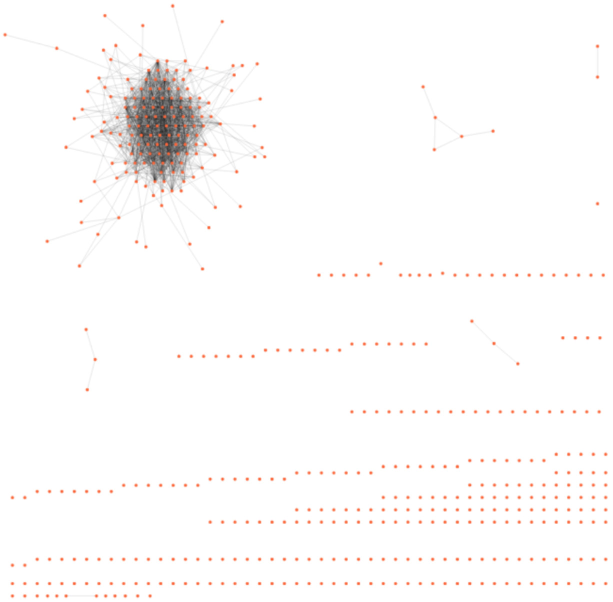

Figure 4: Bad Visualisation - 2

Then we thought of removing away the elements which are isolated as the lack of connections tells us about the lack of data about that portion of the population that leads to them being isolated. We see the effect of doing that in the figure 5. A suitable filter and selection is shown in figure 6

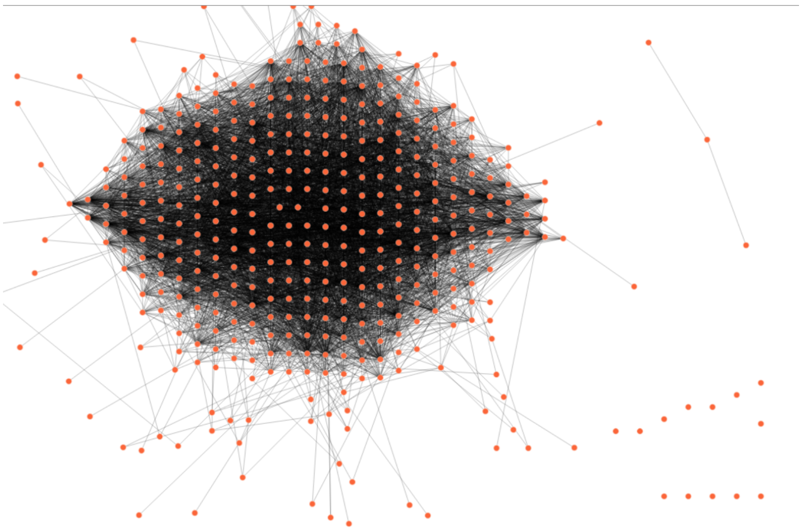

Figure 5: Test Visualisation with removed isolated nodes

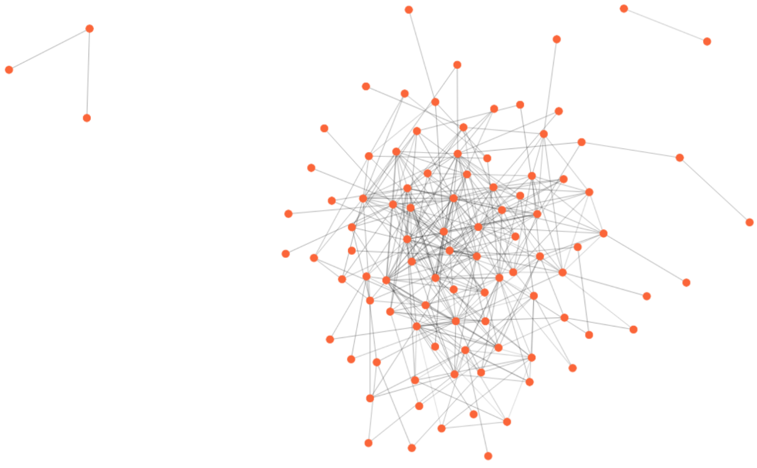

Figure 6: Test visualisation with proper parameters

The sunburst visualisation that we used is useful for selecting and finding out the statistics of the different patients in the TCGA dataset. We use a hierarchical way of the clinical data in order to filter by the demographics of the patients, so at the end we can get their actual proportion.

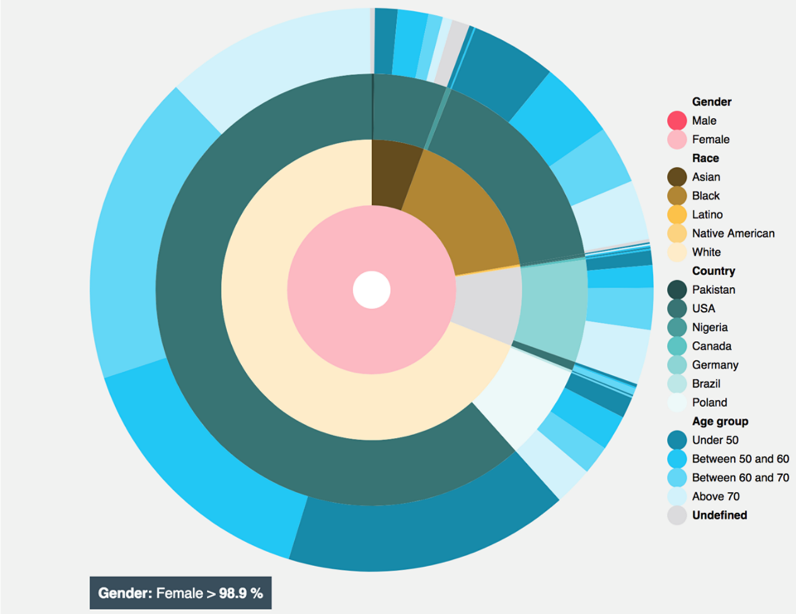

Figure 7: Sunburst

6 What we learnt!

We learnt from our visualization that it is possible to find out where a person stands in the abstract space of patient and gene graphs when it comes to cancer genomics, such that it is possible to tailor the therapies that they receive through the use of such tailored data.

We realised that there is little data to be visualised about people who have cancer from non-Western nations. Adding data from more diverse studies would make this tool useful for a much wider variety of audience.

And of course we learnt a lot about how to use data visualization tools such as javascript.

7 Future Work

We believe (and have been corroborated by Rachel and by our own exploration of other visualisations) that our project has potential to actually be useful in the future, as we envision a new type of clinic in the not too far future, where data science experts, dataviz experts, domain experts in medicine and biology and patients come together to treat the diseases that are stopping humanity from reaching its full potential. We have a number of concrete suggestions, short term and long term that we believe will help us actually deploy this. The team is quite motivated to follow on the work and Rachel also sees potential.

- Find an optimal threshold for displaying the edges and connections in a better way.
- Build a backend that can handle the redrawing of the graph in a dynamic manner so that our original aim of filtering and looking at the graph dynamically is fulfilled.
- We discussed with ourselves and a final outcome that we would like to have is having a platform where methods to calculate novel distance metrics (like the Two Tier Mapper created by Rachel et al.) can be entered, and standard databases like the TCGA can be utilised to check how the distance metric allows things like clustering and prediction.
- We would of course like to extend this to other cancer datasets.
- The gene graph is constant. It could be made into a static 3D graph where distances are the correlations between genes. If it is rendered well, then the filtering of nodes from any initial patient selection would not only give ideas about the relations between patients, but also give ideas about the relationship between related genes. This might even lead to discoveries of new subtypes of cancer and gene coexpressions not previously known.
- The positioning of the patients and genes nodes was something we thought a lot about. If it was not based on some kind of spatial relations, it was a bit meaningless for us. We decided to try and think about the case of say fly genes and position the genes according to the anatomy of the fly, based on which region of the fly body the genes are specifically expressed the most. Also, we may use the same ideas about the breast cancer, though the regions of expression might be the same.