**Breast Cancer Proteomes**

**Correlating breast cancer genotypes to phenotype data using machine learning techniques**

By Robin Paul, Saurabh Jain

**Introduction**

In the “genomics era” where the cost of sequencing entire genomes of individual persons has gone substantially low (~1000$), this has given rise to the idea of “personalized medicine” where treatment of a cancer patient can be given according to the genetic makeup of the patient. According to the central dogma of life, the genome contains all the important information pertaining to life which is encoded in DNA. For specific biological processes, relevant genes are transcribed into RNA which in turn is expressed into proteins for carrying out vital functions of the body. During cancer progression, certain biological processes are carried out by the tumor cell which is different from normal healthy cells. By studying the RNA transcription pattern of genes in individual patients we can correlate genotype (e.g. RNA expression pattern) of patients with patient phenotypes (e.g. histological type, age).

In this project we plan to use a publicly available data for breast cancer patients available in Kaggle. This dataset contains proteome profiling of 77 breast cancer patients. We want to correlate the expression values for 12553 genes for each of these patients with their histological type (ER+, PR+ and TNBC). We also want to correlate other variables such as age etc.

**Datasets**

Reference url: <https://www.kaggle.com/piotrgrabo/breastcancerproteomes/version/3>

Some important features of data:

1. 77\_cancer\_proteomes\_CPTAC\_itraq.csv

Contains the information of various protein for each of the patients

RefSeq\_accession\_number: RefSeq protein ID (each protein has a unique ID in a RefSeq database)

Each row (total 12553) represents a unique protein

Each column (except 2nd and 3rd) represents a patient

1. clinical\_data\_breast\_cancer.csv

column "Complete TCGA ID" is used to match the sample IDs in the main cancer proteomes

Other columns are Gender, Age, ER Status, PR Status, HER2 Final Status etc

Column PAM50 mRNA contains the label for cancer i.e. Basal-like, HER2-enriched, Luminal A, Luminal B. These can be considered as true labels.

**Algorithms :**

In this project we plan to firstly use clustering algorithms to cluster the data into various categories using protein information available for each patient. Some of the scikit learn clustering algorithms that can be considered are K-means, Affinity propagation, Spectral clustering DBSCAN. Using this we intend to judge if we can come up with better clusters which can further be used to classify patients.

Since the number of features is large we also plan to do PCA to reduce features and then perform clustering to assess the overall speed of finding clusters and check the number of clusters formed.

**Metrics for evaluation:**

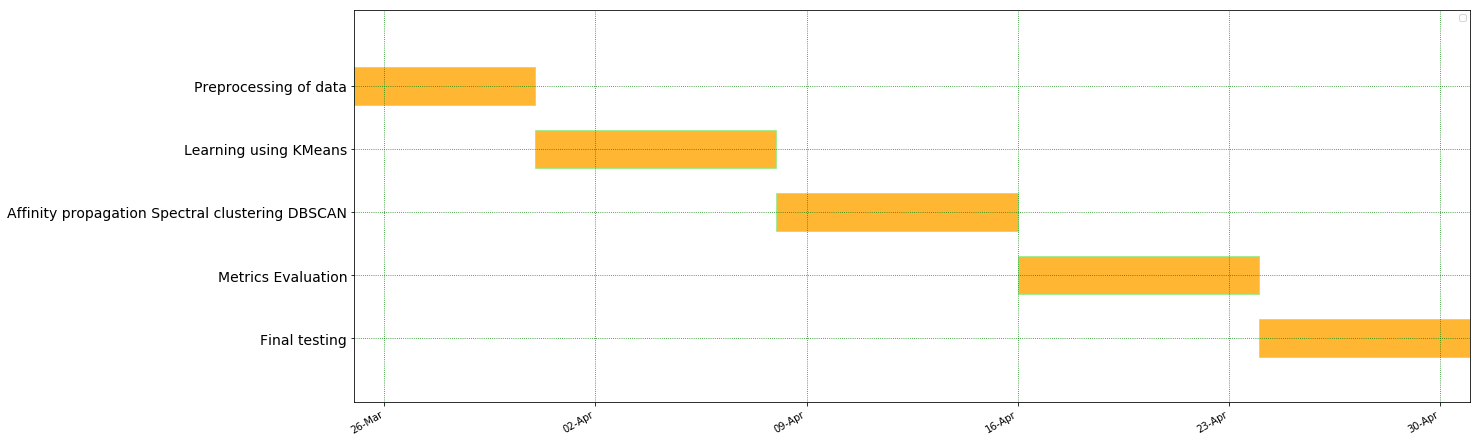
In clustering based the performance is judged by various following ways:

1. Silhouette Coefficient: a higher Silhouette Coefficient score relates to a model with better defined clusters
2. Adjusted Rand index - a function that measures the similarity of the two assignments, ignoring permutations and with chance normalization
3. homogeneity: each cluster contains only members of a single class.
4. completeness: all members of a given class are assigned to the same cluster.

**Key steps:**

1. Reading and loading data
2. Drop unused columns
3. Select specific columns
4. Fill the missing values
5. Set number of clusters – can use multiple values
6. Run various algorithms for each of these cluster size
7. Compute evaluation metrics
8. Plot the resulting clusters

GANTT Chart for timelines:



I plan to use this dataset and run it for various clustering algorithms like Kmeans, Affinity propagation, Spectral clustering DBSCAN and calculate various metrics for evaluation.