Qualitative classification of the breast cancer genome and clustering of the cancer gene network

by

KANIZ FATEMA - 19141026
SHEJUTI SHABNAM - 19141029
AKASH SAHA - 15101085

A thesis submitted to the
Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University

April 2019

Declaration

We, Kaniz Fatema (19141026), Shejuti Shabnam (19141029) and Akash Saha (15101085) from Computer Science and Engineering Department of Brac University hereby declare that

1. The thesis submitted is our own original work while completing the degree.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Signature of Authors:

1.

2.

3.

Approval

A thesis titled "Qualitative classification of the breast cancer genome and clustering of the cancer gene network" submitted by

1. Kaniz Fatema (19141026)
2. Shejuti Shabnam (19141029)
3. Akash Saha (15101085)

Of Spring 2019, has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on April 25, 2019.

Examining Committee:

Signature of Supervisor

Rasif Ajwad

Lecturer, Department of Computer Science and Engineering

BRAC University

Signature of the Chairperson

Dr. Md. Abdul Mottalib

Professor, Department of Computer Science and Engineering

BRAC University

Abstract

The purpose of cancer genome project is to classify the genetic variations that are related to clinical phenotypes. However, some studies showed that some specific cellular pathways are targeted by the cancer mutations genes. But a few of the pathway genes are mutated in each patient. In most approaches, only the existing pathways are considered and the topology of the pathways are ignored. Consequently, new attempts have been targeted on classifying significantly mutated subnetworks and combining them with cancer survival. We had proposed a novel bioinformatics pipeline to identify quantitative classification of the breast cancer genome to verify if the steps will be working or not on real dataset. We have generated a mutation matrix from the collected dataset and calculated pairwise gene similarity. After that, we have also done clustering of the identified cancer gene network, which may help cancer patients by suggesting optimal treatments. We hope our pipeline can also be used for other types of mutation data analysis.

Keywords: Cancer; gene subnetworks; gene similarity; pathways; clustering; bioinformatics.

Acknowledgment

First of all, we would like to thank Almighty Allah to enable us to work on this thesis which has been a great learning experience for us. By the grace of Allah, we could able to put our best efforts and successfully complete it on time. Secondly, we would like to convey our gratitude to our supervisor Mr. Rasif Ajwad for his guidance and handful contribution throughout the whole phase of our thesis work and also to write this report. From the very beginning to the end of the work he has provided us with all kinds of help and inspired us to move forward to our goal. We are also very much thankful to our parents, friends, and well-wishers who have helped us directly or indirectly while conducting our research and continuing our work. We would also like to acknowledge the assistance that we received from a number of resources over the Internet especially from the work of our fellow researches. Finally, we would like to thank BRAC University for giving us the opportunity to conclude the thesis and for giving us the chance to complete our Bachelor degree.

Contents

List of Figures

List of Tables

Acronyms

AMP Adenosine Monophosphate. 26

ANOVA Analysis of variance. 19

BMR Background Mutation Rate. 14

CMDS Correlation Matrix Diagonal Segmentation. 18

CNA Copy Number Alteration. 18

CNV Copy Number Variations. 8

CoMDP Co-occurring Mutated Driver Pathway. 21

CoMET Combinations of Mutually Exclusive Alterations. 22

DiNAMIC Discovering Copy Number Aberrations Manifested In Cancer. 18

DNA Deoxyribonucleic Acid. 4

ER+ Estrogen receptor Positive. 10

ER- Estrogen Receptor Negative. 10

GISTIC Genomic Indentification of Significant Targets In Cancer. 18

GSEA Gene Set Enrichment Analysis. 22

HER2 Human Epidermal Growth Factor Receptor 2. vi, 9

HER2- Human Epidermal Growth Factor Receptor 2 Negative. 10

HETD Hemizygous Deletion. 26

HOMD Homozygous Deletion. 26

HotNet Hybrid Optical Transport Network. 16

HR+ Hormone Receptor-Positive. 10

iMCMC Identifying Mutated Core Modules in Cancer. 21

MCMC Markov Chain Monte Carlo. 19

MDP Mutated Driver Pathway. 19

ME Mutually Exclusive. 20

MEMo Mutual Exclusivity Modules. 18

METABRIC Molecular Taxonomy of Breast Cancer International Consortium. 24

MuSiC Mutational Significance in Cancer. 14

PI3K Phosphatidylinositol-3 Kinase. 16

PPI Protein–Protein Interactions. 5

PR+ Progesterone Receptor Positive. 10

PR- Progesterone Receptor Negative. 10

PSMP Pairwise Search for Mutational Pattern. 20

RNA Messanger Ribonucleic Acid. 7

RTK Receptor Tyrosine Kinases. 16

SNP Single Nucleotide Polymorphism. 7

SNV Single Nucleotide Variants. 6

STAC Significance Testing for Aberrant Copy number. 18

TCGA The Cancer Genome Atlas. 14

TP53 Tumor Protein P53. 15, 16

UC Ulcerative Colitis. 15

Glossary

ANOVA F-test ANOVA can determine whether the means of three or more groups are different. ANOVA uses F-tests to statistically test the equality of means. 20

biomaRt project originated at the European Bioinformatics Institute as a data management solution for the Human Genome Project. 27

ClusterOne graph clustering algorithm. 17

driver genetic mutations mutation within a gene that confers a selective growth advantage. 13

Ensembl provides a genome browser that acts as a single point of access to annotated genomes for mainly vertebrate species. 27

epigenetic relating to or arising from non-genetic influences on gene expression. 5

genome the complete set of genes or genetic material present in a cell or organism. 4

heterogeneity the quality or state of being diverse in character or content. 5

HotNet2 a general algorithm for identifying high weight subnetworks in a vertex-weighted network. 17

hsapiens gene ensembl human gene dataset. 27

malignancy the state or presence of a malignant tumour; cancer. 14

Paired sample t-test a statistical procedure used to determine whether the mean difference between two sets of observations is zero. 19

passenger mutations mutation that do not provide a growth advantage. 13

phenotype the set of observable characteristics of an individual resulting from the interaction of its genotype with the environment. 5

PIK3CA a gene that encodes a lipid kinase involved in multiple signaling pathways. 15

SOMAT a statistical approach or detecting somatic mutations associated with multiple cancer-related traits. 22

somatic mutations the occurrence of a mutation in the somatic tissue of an organism, resulting in a genetically mosaic individual. 4

Student t-test a method of testing hypotheses about the mean of a small sample drawn from a normally distributed population when the population standard deviation is unknown. 19

transcriptomics the study of the transcriptome—the complete set of RNA transcripts that are produced by the genome, under specific circumstances or in a specific cell—using high-throughput methods, such as microarray analysis. 5

Chapter 1

Introduction

Most of the time cancer occurs for somatic mutations that gather in the genome over the lifetime of an individual. It occurs with the assistance from epigenetic and transcriptomics modifications. In past years, DNA sequences has changed the way toward recognizing somatic mutation in cancer genomes. Driver mutations are the hereditary changes that partake in cancer development and other than these transformations are known as passenger mutations. The significance of DNA sequencing helped biologists to quantify single-nucleotide changes with speed and precision [1]. These looks into demonstrated that, the typical cancer genome may have hundreds to thousands of somatic mutations. Cancer cells contain extensive quantities of relatively uncommon mutations, which are not as a rule found in human body. For recognizing driver mutations, genome-wide studies are vital for sequencing various patients. Single Nucleotide Variants (SNV) are enhancement in

a single nucleotide without any imperatives of recurrence and may rise in substantial cells are most basic among the hereditary varieties. These are considered as the 'building blocks' of DNA. The presence of SNV is found in a gene, can coordinate impact in infections. Copy Number Variations (CNV) are two duplicates of every gene found in human body. One duplicate of gene is conveyed from each parent. CNV can likewise be in charge of cancer [2]. Cancer happens when the average rule of cells is meddled. For breast cancer, the genetic changes are gotten in the midst of the lifetime of an individual and impact certain breast cells. These movements are known as somatic mutations. One of the fundamental steps in cancer sequencing is to identify the driver mutations that are responsible for cancer. Single Gene-Level Analysis is a stage for the sequencing. Pathway and Network level Analysis is progressively effective. The advantage of pathway examination is that, results got for different related datasets can be thought easily, as pathway data can guarantee the understanding of the information is done in a typical element space. This is a way to test relationship among transformation and phenotype at the pathway level has been executed in different examinations. There are various methods that can detect mutually exclusive genomic alteration patterns in cancer genomic datasets. De Novo Driver Exclusivity (Dendrix), Mutated Driver Pathway Finder (MDP Finder), PARADIGM, Mutex, Mutually Exclusive (ME) are some types of methods for detecting mutually exclusive genomic alteration patterns in cancer genomic datasets. For our research purpose, we have used two datasets; one is Discovery dataset

and the other is Validation dataset. At first, we retrieved gene names and generated mutation matrix generation. After that, we calculated pairwise weighted gene similarity. Lastly, we have used K-Means clustering algorithm for clustering gene network using weighted values.

Chapter 2

Background

2.1 Cancer and DNA sequencing

Cancer happens for somatic mutations in the Deoxyribonucleic acid (DNA) sequence of an individual that occurs during the lifetime of person. DNA replication is the main cause for the development of the mutations which happens as cells develop and split into two subsidiary cells. Mutations arise because errors in DNA replication process and determine the DNA in the daughter cells from the parental cells. A major goal of cancer genome sequencing is to establish functional links between genetic variations and human diseases. Critical diseases like cancer may often drive by inconstant changes in multiple genes in pathways. Modules or sub networks in biological networks helps to isolate systems with disease related topics. A milestone paper integrates gene expression marks with protein–protein inter-

actions (PPI) to discover predictive modules of cancer result. Most of the time cancer occurs for somatic mutations that gather in the genome over the lifetime of an individual. It happens with the help from epigenetic and transcriptomics alterations. In past years, DNA sequences has transformed the process of identifying somatic mutation in cancer genomes. Driver mutations are the genetic mutations that participate in cancer development and other than these mutations are known as passenger mutations. Additionally, driver mutations are caused by high frequency of passenger mutations which are inappropriate for progression of cancer phenotype. Entire genome sequencing uncovers somatic mutations of different kinds, while sequencing distinguishes coding mutations at a lower cost, however does not permit the analysis of non-coding regions. The importance of DNA sequencing helped biologists to measure single-nucleotide mutations with speed and accuracy [1]. These researches showed that, the typical cancer genome might have thousands of somatic mutations. Cancer cells contain large numbers of comparatively rare mutations, which are not usually seen in human body. For identifying driver mutations, genome-wide studies are necessary for sequencing numerous patients. Though DNA sequencing technology has advanced in recent years, it still faces numbers of complications. Genome sequences reveal all types of somatic mutations. There are three major challenges found in cancer genome sequencing. Firstly, depending on the presence of intra-tumor heterogeneity, somatic mutations from the sequence reads are generated by high throughput technologies. Secondly, recognizing the relatively small number of driver

mutations that are responsible for the growth of cancer from large number of driver mutations which are not responsible for the cancer phenotype. Lastly, the challenge of determining the biological processes that are altered by somatic mutation. The quick developments in high throughput DNA sequencing technologies and their application helps to analyze the resulting data.The recent accessibility of extensive protein systems gives one means to in any event mostly address these difficulties. Utilizing protein–protein connection systems, the yeast two-hybrid framework, or mass spectrometry, various methodologies have been exhibited for extracting pertinent subnetworks dependent on reasonable articulation examples of their genes or on preservation of subnetworks over different species. Each subnetwork is suggestive of a particular practical pathway or intricate, yielding well known and novel pathway speculations in life forms for which adequate protein association information have been estimated. Substantial protein systems have as of late turned out to be accessible for human empowering new open doors for explaining pathways engaged with real illnesses and pathologies [3].

2.2 Single Nucleotide Variants

Single Nucleotide Variants (SNV) are diversification in a single nucleotide with no constraints of frequency and may emerge in somatic cells are most common among the genetic variations. These are considered as the 'building blocks' of DNA. The presence of SNV in the DNA of an individual is

a common consequence. When the existence of SNV is found in a gene, it can direct influence in diseases. SNV can affect gene function. Most of the time SNV does not have important effect in the development of human health. SNV is responsible for the variations among people and the familial characteristics like hair color, eye color and other external characteristics. Sometimes the drug effect differences are also influenced by SNV. SNV may also cause mutations like synonymous and nonsynonymous. Single nucleotide polymorphism (SNP) is the least complex type of DNA variety among people. These basic changes can be of progress or transversion type.They might be in charge of the assorted variety among people, genome advancement, the most well-known familial attributes, for example, curly hair, inter-individual contrasts in medication reaction, and common diseases like diabetes, hypertension, and mental issue. SNPs may change the encoded amino acids or can be quiet or just happen in the non coding areas. They may impact advertiser movement (quality articulation), delegate Messanger Ribonucleic acid (mRNA) adaptation , and subcellular confinement of mRNAs or potentially proteins and henceforth may deliver illness. In this manner, distinguishing proof of various varieties in qualities and examination of their belongings may prompt a superior comprehension of their effect on health of a person [4, 5, 6].

2.3 Copy Number Variations

Copy Number Variations (CNV) are two copies of each gene found in human body. One copy of gene is carried from each parent. There are several cases where copy number of a specific gene varies from person to person. The person can have only one copy or more than two copies of a specific gene. There are also cases like one or two copies of genes are missing. These differences are stated as genetic differences and they are known as CNV. CNV can be responsible for diseases. All mutations are not responsible for the damage of human cells. Copy number variation (CNV) of DNA sequence is practically noteworthy yet still cannot seem to be completely learned. CNVs contain many qualities, sickness loci, useful components and segmental duplication [2].

CNV results three types of mutation sequence in large DNA segments. They are- insertions, deletions and duplications. These variations can work positively sometimes. Extra copy of gene create overflow in the gene. That means extra copy of gene can develop task for adoption. The remaining copies also perform the original task [7].

2.4 Breast Cancer

There are a few genes in our body that controls essential cell functions, for example, cell development, cell growth and cell division. The cell functions should be directed carefully to guarantee that DNA is copied appropriately.

At the point when these genes are done with mutation, it influences the cell functionalities. Therefore, the problem is DNA is not fixed and cell development and division becomes tough to control, which gets the form of tumor. Cancer happens when the typical guideline of cells is interfered. For breast cancer, the hereditary changes are gotten amid in a lifetime of a person and influence certain breast cells. These progressions are known as somatic mutations. Moreover, hereditary transformations present in all cells in the body can also induce the danger of breast cancer. These hereditary changes are known as germline mutations. The main difference between somatic mutation and germline mutation is that germline transformations are acquired from parents [8].

2.5 Various types of Breast Cancer

Depending upon how cancer cells react to various receptors, from an indicative perspective, breast cancer growth has been separated into unique types. This division guarantees that each kind of breast cancer has been treated with various chemotherapy and different types of hormone treatments.

2.5.1 Human epidermal growth factor receptor 2 (HER2)

Sometimes, the cells that contain tumor produces overabundance of HER2 protein. This kind of cancer growth becomes moderately quick in comparison with different sorts.

2.5.2 Hormone receptor-positive (ER+& PR+)

A hormone-receptor-positive HR+ tumor is a tumor which contains cells that express receptors for explicit hormones. The term most ordinarily refers to estrogen receptor positive tumors and likewise incorporate progesterone receptor positive tumors. Estrogen-receptor-positive tumors have receptors that enable them to utilize the hormone estrogen to develop and these rely upon the presence of estrogen for progressing expansion. Moreover, Progesterone-receptor-positive tumors are sensitive to the hormone progesterone, and the cells of these kind of tumor have receptors that enable them to utilize this hormone to develop. Around 80% of all breast cancer growths are supposed to be ER+. Out of these, around 65% develops in response of progesterone hormone, which is known as PR+.

2.5.3 Triple positive and triple negative

Triple positive are usually types of cancer that are HER2+, ER+, PR+. For triple negative, it is HER2-, ER-, PR-.

2.6 Subtypes of Breast Cancer

2.6.1 Luminal A

Luminal A breast cancer is the very first type of breast cancer mentioned in past subsection (ER+ and PR+). In any case, it's HER2-and has low dimen-

sions of the protein Ki-67, which maintains the speed of cell development. Luminal A malignant growths develop gradually and have the best prognosis [9].

2.6.2 Luminal B

Luminal B breast cancer growth is also ER+ and PR+, and either HER2+ or HER2-. The distinction with Luminal A being, in Luminal B there are abnormal amounts of Ki-67.Luminal A for the most part becomes marginally quicker than Luminal A malignant growths and have somewhat more terrible guess than Luminal A [9].

2.6.3 Triple-negative/basal-like

Basal-like breast cancer refers triple negative (ER-, PR-, HER2-).This sort of cancer likewise is progressively seen among more young people and African-American women [10, 11]

2.6.4 HER2-advanced

This subtype of breast malignancy is ER-and PR-yet HER2+. HER2-advanced tumors becomes quicker than luminal cancer growths and can have a negative prognosis [12].

2.6.5 Normal like

This subtype of breast disease resembles Luminal A subtype, ER+ and PR+, HER2-, and has low dimensions of the protein Ki-67. Moreover, it is prognosis marginally bad than Luminal A.

Chapter 3

Mutation Analysis Approaches

There is a lot of information to distinguish driver genetic mutations from the passive passenger mutations. The common steps are to analyze the genomic alterations over a large number of patients where the variations usually correspond to non-random mutations. However, there is another complex approach to identify genes with driver mutations which helps to discover genes with a significantly high mutation frequency over a large number of patients. In the following sections, we will discuss about single gene level, pathway and network level of mutation analysis approaches.

3.1 Single Gene-Level Analysis

As discussed before, one of the fundamental steps in cancer sequencing is to identify the driver mutations that are responsible for cancer. In spite of

the developmental phases which gives us significant advantage generating the output. However, other than mutation frequency, a number of characteristics of mutation rate could affect sequence contexts, mutation types, mutation-specific, gene-specific features scores that evaluate functional impact and so on. Therefore, recently a number of frequency-based methods developed which get a more absolute Background Mutation Rate (BMR) estimation by adopting one or more of these features.For example, both Mutational Significance in Cancer MuSiC [13] and MutSigCV [14] sample-specific rates of mutations and employ the types of mutations. MutSigCV also allows addition of gene-specific features like the replication timing and expression level.Mutational Significance in Cancer (MuSiC) for these extensive data sets. The coordination of analytical activities in the MuSiC structure is generally relevant to a wide arrangement of tumor types and offers the advantages of automation and additionally institutionalization. Thus, we portray the computational structure and factual underpinnings of the MuSiC pipeline and show its execution utilizing 316 ovarian malignancy tests from the TCGA ovarian disease venture. MuSiC effectively affirms many expected outcomes, and recognizes a few conceivably novel roads for disclosure[13].

3.2 Pathway and Network level Analysis

Contrasted with breaking down hereditary transformation information at single quality level, pathway and system examinations can extricate more data

as these strategies manage different qualities in a similar pathway or system, so the likelihood that an atomic occasion will pass the measurable edge is expanded and the quantity of speculations tried are diminished [15]. Another advantage of pathway examination is that outcomes got for various related datasets can be thought about effortlessly, as pathway data can guarantee the understanding of the information is done in a typical element space [16]. This way to deal with test relationship amongst transformation and phenotype at the pathway level has been executed in various investigations [17]. An ongoing report led a pathway-level examination to foresee the general survival of Ulcerative Colitis (UC) patients [18]. They found that 35 out of 103 examples had transformations in no less than one quality in Tumor Protein P53 (TP53) and PIK3CA, which comprises of 16% and 9% of the aggregate number of changes recognized in the examination. The creators additionally found that around 65% of the patients had CNV changes.

Pathway level investigation can expand the measurable capacity to distinguish altogether transformed pathways in particular growths and has better organic translation. Be that as it may, the approach of recognizing pathways being changed in extensive quantities of patients has its impediments as well, on the grounds that exclusive existing pathways are considered, overlooking the topology of the pathways. Besides, the pathways are examined in detachment however they communicate in bigger systems, which may disregard numerous gatherings of interfacing qualities that are not in known pathways but rather have critical relationship with clinical phenotypes [19].

As of late, techniques to distinguish transformed subnetworks among malignancy genomes have been presented. They proposed a system construct approach situated in light of the speculation that cell systems are particular and have between associated proteins that perform particular natural capacities [20]. The creators utilized a bound together atomic association arrange comprising of both protein-protein collaborations and pathways to play out an incorporated system investigation for distinguishing applicant driver changes. Their approach was a mix investigation of grouping changes and duplicate number modifications.

Furthermore, another approach discovered subnetworks by considering that changes in the subnetwork are corresponded with the clinical parameter of survival time of patients [19]. They introduced a calculation called HotNet to distinguish essentially changed subnetworks controlled by the transformation recurrence of individual qualities alongside the associations between them. They considered the change as a wellspring of warmth on the system and extricated the 'hot' hubs. The criticalness of the subnetworks was computed utilizing both the topology of the systems and the recurrence of transformation of the qualities. An alternative proposal was presented which is a changed variant of the beforehand said Hotnet calculation was named Hotnet2 [21]. They utilized this refreshed calculation to examine a Pan-Cancer dataset of 3281 examples from 12 growth writes. The creators distinguished altogether transformed subnetworks with referred to pathways, for example, TP53, RTK, PI3K, and so forth. Nonetheless, the creators noticed that a

few qualities with high individual change scores were missing from the system investigation comes about and expressed this is because of the absence of information and false negatives in the broke down information.

One key restriction of the current subnetwork-based methodologies is that they don't allot the same changed qualities into various subnetworks in spite of the fact that covered subnetworks are conceivable. This propels us to apply two system based grouping calculations to examine bosom malignancy CNV transformation information for recognizing altogether changed subnetworks. The recognized subnetworks can be utilized to test the relationship of transformation status of the qualities in the subnetworks with bosom growth patients' survival. The main approach is called HotNet2 [22] while the other approach is called ClusterOne [23], which has not been connected to examine tumor change information previously, however was produced and connected to recognize covered protein buildings in protein collaboration systems. We embraced this approach since a quality can be doled out to different subnetworks and qualities are generally engaged with numerous buildings and pathways. We grew new change investigation pipeline by considering system topology for estimating quality combines transformation comparability to induce the altogether changed subnetworks [24].

3.3 Comparative Analysis of Different Approaches

Table 3.1: Comparative Analysis

Method	Category	Description	Data Type	Technique
Mutual Exclusivity Modules (MEMo) [25]	Prior Knowledge - Based	determines the applicant driver subnetworks with properties (1) the genes that are part of driver pathway are repeatedly modified among various patients;(2) the genes have a tendency to take part in similar pathway or biological process; and (3) the modified genes inside the driver pathway are mutually exclusive	Somatic Mutation data and Copy Number Alterations (CNAs)	Bayesian Method and Statistical Analysis (STAC, GISTIC, CMDS, DiNAMIC)

Continued from Previous Page				
De Novo Driver Exclusivity (Dendrix) [26, 27]	De Novo Identification	algorithm to identify driver genes with high scopes and high specificity	Mutation Data	Markov Chain Monte Carlo (MCMC) algorithm
Mutated Driver Pathway Finder (MDP Finder) [27]	De Novo Identification	Using stochastic search algorithm and an exact model find mutated driver pathway	Submatrix Problem	Binary Linear Programming (BLP)
Genetic Algorithm (GA) [27]	De Novo Identification	maximize more common and adjustable weight functions	Gene Expression Data	Statistical Test (ANOVA F-test,Student t-test, Paired sample t-test etc.)

Continued from Previous Page				
Mutex [28]	Combination of Prior Pathway Knowledge and Statistics	a method which determines groups of mutually exclusive modified genes having typical succeeding target	Interaction Database	Greedy Algorithm and Permutation Test
Mutually Exclusive (ME) [28]	De Novo Identification	it is used for searching cancer modification data and find faulty gene sets	Gene Expression Data	Statistical Test (ANOVA F-test,Student t-test, Paired sample t-test etc.)
Pairwise Search for Mutational Pattern (PSMP) [29]	Simple Mutual Exclusivity - based	it is the basic for understanding various methods for identifying driver pathway	Somatic Mutation	Bayesian Method

Continued from Previous Page				
Identifying Mutated Core Modules in Cancer (iMCMC) [30]	Network-based for De Novo Identification	to discover various mutated core modules present in initiation of cancer formation pathways	Somatic Mutation, CNAs and gene expressions	Bayesian Method & Statistical Analysis (STAC, GISTIC, CMDS, DiNAMIC)
Co-occurring Mutated Driver Pathway (CoMDP) [31]	Cooperative Pathway	a way to explore combination of different pathway and if they are concurrently mutated in huge group of patients	Simulation Data and Biological Data	Clustering (Cancer Correlation Clustering [32])
PARADIGM [33]	Network- or Pathway-Based Approach	a method for discovering constant pathways in cancer	CNVs and Gene expression	Bayesian Method & Statistical Test(ANOVA F-test,Student t-test, Paired sample t-test etc.)

Continued from Previous Page				
DriverNet [33]	Network- or Pathway-Based Approach	identifies driver mutations evaluating their effect on mRNA expression	Somatic Mutation	Statistical Test (SOMAT)
HotNet [34]	Interaction Network-Based Method	an algorithm to discover notably modified subnetworks present in a huge interaction network	Mutation and Known Pathways	Heat Diffusion process and Statistical Test (Combinations of Mutually Exclusive Alterations (CoMET))
Gene Set Enrichment Analysis (GSEA) [35]	Pathway Analysis and Combinations of Mutations	a technique to sort the mutated gene list and evaluates whether a pre-defined set of genes has more high-ranking genes than expected	Mutation and Known Pathways	Statistical Test (CoMET)

3.4 Motivation

Somatic mutation in DNA sequence is mainly responsible for cancer. Cancer genome sequence is necessary for establishing functional links between genetic variations and human diseases. A type of cancer mutation named driver mutation is responsible for the growth and survival of cancer. On the other hand, passenger mutations do not have any impact on cancer. Since genes (proteins) for the most part interact with different genes to execute their capacities, networks can be measured and separated into subnetworks. Our thesis proposed a new analysis pipeline for identifying mutated subnetworks in breast cancer genome. We utilized one existing clustering algorithm to implement the pipeline. One obvious difference from the existing methods from our method is, our method is a weighted method. The weighted values are used for clustering of the gene networks. Weighted methodology guarantees that the score for every gene are distributed depending on their mutation frequency over every one of the samples. That is, genes that are responsible for cancer are connected with each other. So it is easier to identify the responsible genes in a subnetwork which is more precise than the unweighted approach. The calculated weighted values are used for K-means clustering algorithm.

Chapter 4

Data Analysis and Methodologies

For locating significantly mutated and clinically relevant subnetwork we designed a pipeline as shown in Figure [3.1]. It describes our collection of two sets of data and then pre-processing the datasets for clustering. Finally combining the two result we will compare them.

4.1 Data Analysis

Qualitative classification of the breast cancer genome requires gene interaction networks and patient specific mutations. We have collected CNV datasets for mutation data from a Canada-UK based project known as METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) where

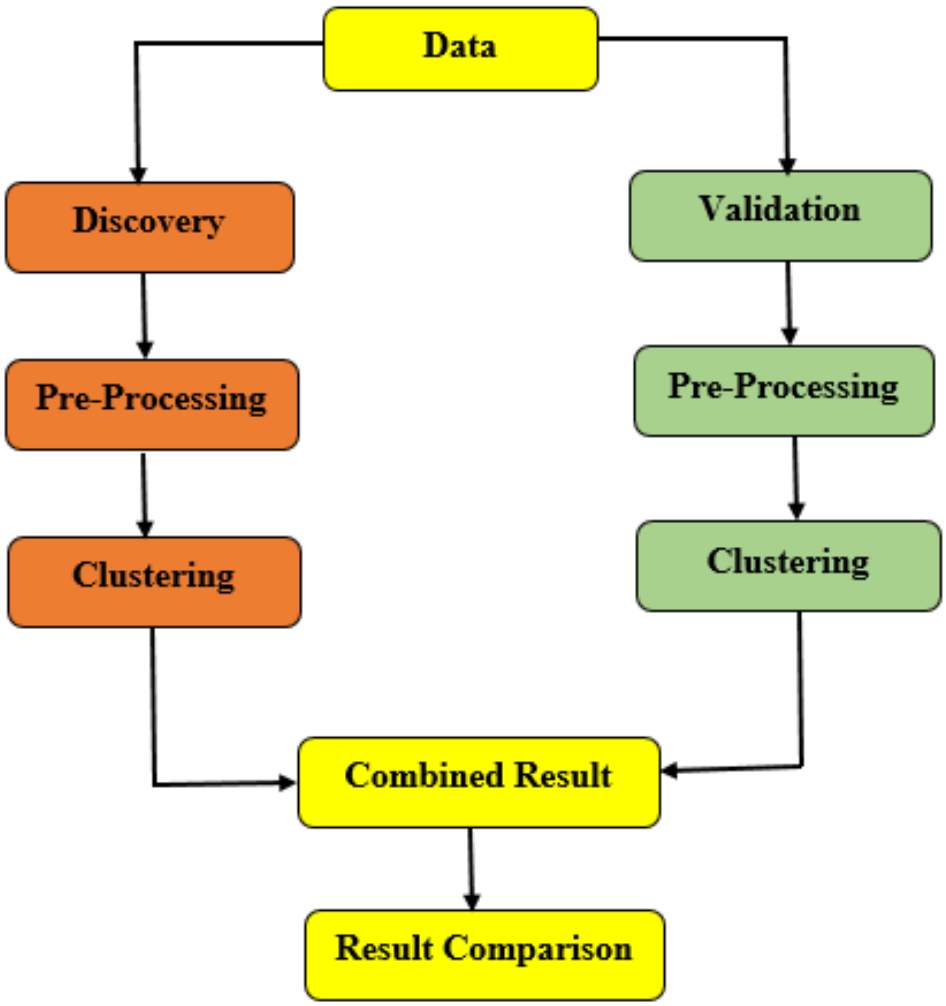

Figure 4.1: Analysis Pipeline

they have over 2000 clinically clarified foremost fresh-frozen breast cancer specimens collected from their tumor banks [36, 37]. These datasets classify the breast tumors into more subcategories. In the first place, location of chromosomes, paired DNA and RNA in a dataset of 994 samples were evaluated. This data set of 994 female patients is represented as 'Discovery' dataset. Another group of 990 data set samples is represented as 'Validation' set. The basis of 'Validation' dataset was to test against the 'Discovery' dataset and to verify whether the accuracy was sufficient. The dataset is composed of chromosomes, their starting location, ending location and the length of the CNV. The chromosomes were given numbers based on their location. Moreover, four types of discrete somatic conditions existed: GAIN, AMP, HETD and HOMD. GAIN and AMP had been converted to gain from these somatic discrete and HETD and HOMD had been converted to loss. Therefore, the CNV calls were represented as 1 and -1 for gain and loss accordingly.

4.2 Methodologies

4.2.1 Mutation Matrix Generation

Our first target was to get the location wise names of the genes as our output from the collected two datasets. For the 'Discovery' set there were a total number of 131956 calls and for the 'Validation' set there were 137896 number of calls in total. Therefore, to get the location wise names of the genes for 'Discovery' set and 'Validation' set, we needed the chromosome number,

their starting location and ending location from the data sets. Thus, for retrieving the CNV specific genes we called 'biomaRt' library on R and read the csv files of the two datasets separately [38]. Moreover, we have also used a dataset for human genes from 'Ensembl' database known as 'hsapiens gene ensembl' [38]. After that we have set 'NULL' for no genes present in any chromosome and 'NA' was set for the genes not located in any chromosome and the rest of them were the name of the genes. We wrote the csv files for 'Discovery' set and 'Validation' set as 'Result_Discovery' set and 'Result_Validation' set respectively. Finally, we found on 'Result_Discovery' set and 'Result_Validation' set that on some locations in the chromosomes there were no genes present and on some there were multiple genes present and some of them were not available. We filtered out the 'Result_Discovery' set and 'Result_Validation' set drawing out the samples which were 'NULL' and 'NA'. In the 'Result_Discovery' set the total number of calls were 165776 and after filtering out the 'NULL' and 'NA' there were 127188 calls. On the other hand, in the 'Result_Validation' set the total number of calls were 173750 and after the omission the total number of calls were 133499 calls. Furthermore, the same genes were found multiple times under one Metabric_ID and under different Metabric_IDs the same types of genes were found. But we needed to find the CNV Gain and CNV Loss for unique genes and for each unique Metabric_IDs. Thereafter, for getting a matrix of rows for unique Metabric_IDs, we turned the same Metabric_IDs found multiple times for same genes into unique Metabric_IDs. Likewise, to get columns in the same

matrix for unique genes, we turned the same genes found multiple times under one Metabric_ID into each unique gene. The CNV types for each gene in each Metabric_IDs were Gain and Loss. If the CNV type was Gain the value was 1, if the CNV type was Loss the value was -1 and for neither Gain nor Loss it was set as 0. After this, we saved these csv files as 'Matrix_Discovery' set and 'Matrix_Validation' set. In the 'Matrix_Discovery' set we found 790 unique genes for 994 Metabric_IDs. Similarly, in the 'Matrix_Validation' set 790 were unique genes and 990 were Metabric_IDs. The total number of Gain and Loss calculated was 111460 and 54316 respectively for 'Matrix_Discovery' set. Similarly, we have also calculated the total number of Gain and Loss for 'Matrix_Validation' set which are 113213 and 60537 accordingly. Besides, we created a line graph for the total gain and total loss in each Metabric_ID for the both sets as shown in Figure[4.1] and developed a bar chart total gain and total loss for both the sets as shown in Figure [4.2].We also generated two curved graphs for Top 200 gain values and Top 200 loss values for both 'Matrix_Discovery' set and 'Matrix_Validation' set as given in both Figure [4.3] and Figure [4.4].

4.2.2 Pairwise weighted gene similarity calculation

Our next step after mutation matrix generation was to calculate the pairwise weighted gene similarity. For similarity calculation, at first, we looked for suitable similarity equation and found cosine similarity equation (3.1) [39]. In this equation m and n represents the genes and t represents their mutation

frequencies. We have two types of mutation frequencies, one is gain *(k=g)* and the other is loss *(k=l)*.The purpose of finding the mutation frequencies is to have a gene-specific mutation score measure so that we can use the score as the weight for each interaction in the network.

$$cos\_sim(m,n) = \frac{\sum_{k \in g,l} t_{mk} t_{nk}}{\sqrt{\sum {t_{(mg)}}^2} \sqrt{\sum {t_{(nl)}}^2}} \tag{4.1}$$

So, to calculate the similarity at first we calculated the total number of gain and total number of loss for each gene for both 'Matrix_Discovery' data set and 'Matrix_Validation' data set and made two csv file named as 'Matrix Discovery Sum Gain_Loss' and 'Matrix Validation Sum Gain_Loss'. After that we called the csv files on R separately and put the above equation inside loops to get the output matrix as csv files named 'Similarity Genes Discovery' and 'Similarity Genes Validation'for 'Matrix Discovery Sum Gain_Loss' and 'Matrix Validation Sum Gain_Loss' respectively. We have set the loop in such a way that each gene will be compared with all other genes in the matrix. The numerator of the equation is the product of the total gain of two different genes and in the denominator, it is the product of the square root of the squared of total gain and total loss of the same gene. Since we have two types of mutation frequencies, we have used this equation in such a way that the similarity of the genes are calculated considering both total number of gain and total number of loss of each gene. The similarity calculation of each gene with every other genes in the datasets is done as shown in Figure [3.2].

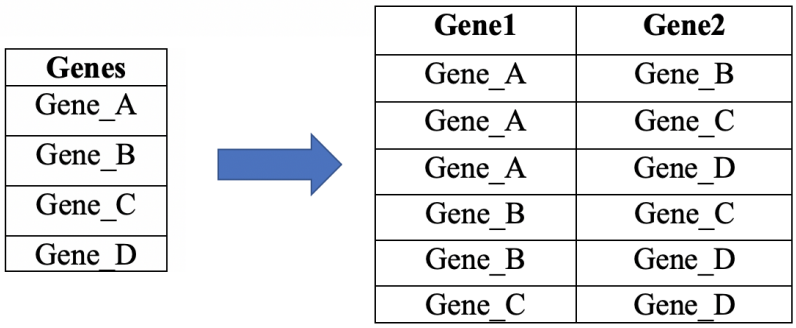

Figure 4.2: Sample Representation of Gene Similarity Calculation

Finally, after calculation we have found out that the total number of data in both the output files are same which is 108264. The similarity values of the paired genes will be used as input for the clustering of gene network.

4.2.3 K-Means Process

We have used a centroid based hard clustering type of algorithm known as K-means for clustering of gene network. It is an iterative clustering algorithm that derives the notion of similarity from how close a data point is to the centroid of the cluster. This method divides the dataset into unique homogenous clusters whose observations are like each other yet not the same as other clusters and the resultant clusters do not overlap with each other [40]. This algorithm consists of two different steps. Firstly, the in advance fixed values of k as cluster centre is chosen randomly. Thereafter, in the second step, each data object is taken to the closest centre after the calculation of the distance between each data object and all k cluster centre

[40, 41]. Generally, the distance between each data object and the center of the clusters are determined by Euclidean distance. At the point when every one of the data objects are incorporated into certain clusters, the initial step is finished and an early grouping is finished. After that, recalculation of the mean of the clusters' center of each clusters which were formed early is done. This iterative procedure proceeds over and over until the criterion function becomes the minimum [40].

An improved k-means algorithm has been proposed by the authors in 2010 from the above explained method [40]. In the improved k-means algorithm the authors' idea was to set two straightforward data structures to hold the names of the clusters and the distance between all the data object and closest cluster during each iteration which can be used in the next iteration. They calculated the distance between the present data object and the new cluster centre. After that, if the calculated distance is smaller than or equivalent to the distance of the old cluster centre, the data object remains in it's previous iteration cluster.[40]

4.2.4 Identification of Mutated Subnetworks Using K-Means

After finding the similarity values of both 'Matrix Discovery Sum Gain_Loss' dataset and 'Matrix Validation Sum Gain_Loss' dataset, we saved it as CSV file named 'SimilarityGenesDiscovery' and 'SimilarityGenesValidation'. As

mentioned above, in the output files of similarity calculation there were two columns for Genes and one column was for the similarity values. These similarity values are weighted values. This weighted methodology guarantees that the score for every gene are distributed depending on their mutation frequency over every one of the samples. The weighted values in the datasets were less than 1 and those values which were equal to 1 are the comparison that occurred with itself. We have seen that both the datasets were too large and we did not have the system for having the whole calculation on process. Thus, we have selected the genes which have got most similarities. From the discovery dataset we considered the genes whose weighted values were greater and equal to 0.9 and less than 1 and finally selected 19462 CNV data. We made a dataset consisting of these 19462 CNV datas and named it 'topdiscovery' and saved it as a csv file. Next, we made the dataset into matrix form and found 7763 genes by 4375 genes and their weighted values and saved it as a csv file named as 'Similarity_disM' We used two libraries on R named 'tidyr' and 'dplyr' to create this matrix format from the column wise values. Since our dataset is too large we added a new row and spread the columns into numeric matrix and saved a new CSV file named 'Similarity_disM'. After that, we have omitted the 'NA' values and replaced it with '0'. Finally, we used K-means algorithm setting the centres of the clusters to '10' and drew a plot as shown in Figure [4.5]. Likewise, from the validation dataset we selected 19010 considering the genes whose weighted values were greater and equal to 0.9 and less than 1 and created a csv file

named ‘topvalidity’. Hereafter, we have followed the same procedure as we did for the discovery dataset using the two libraries named ‘tidyr’ and ‘dyplr’ to create matrix format from the column wise values. We found 6960 genes by 4416 genes with their weighted values and saved the dataset as csv file named as ‘Similarity_valM’. Similarly, we used K-means algorithm setting the centres of the clusters to ‘10’ and drew a plot as shown in Figure [4.6] as we did for discovery dataset.

Chapter 5

Result Analysis

5.1 Mutation Matrix Generation

We found 790 genes in both 'Discovery' dataset and 'Validation' dataset based on the individual CNV positions. We calculated the total number of mutation frequencies in each Metabric_Id individually for both the datasets and created a line graph. We created a 2d line graph as shown in Figure [4.1] where along the x-axis we considered the Metabric_Ids of the patients and along the y-axis we considered the total number of mutation frequencies. We have curved the graph for both the datasets. For identification, gave blue color for 'Discovery' dataset and red color for 'Validation' dataset. From the graph, we have found out that for same gene the mutation frequencies can be different for same patient. To clarify, the CNV type can be both Gain and Loss in one gene for same patient and also for different patient.

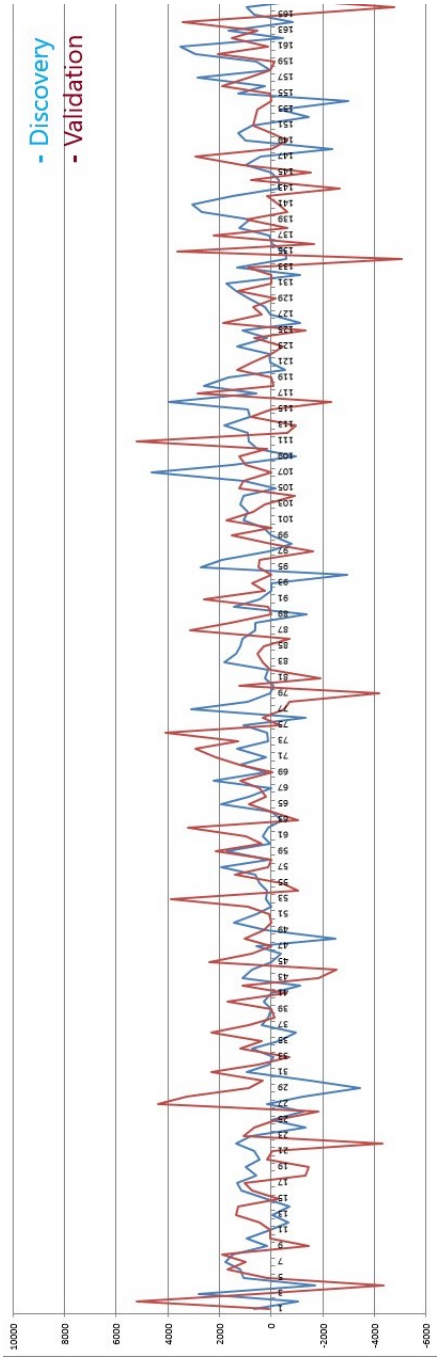

Figure 5.1: CNV Type For Discovery and Validation

Furthermore, we created a bar chart as shown in Figure [4.2] to have a graphical view of the total number of mutation frequencies of all Metabric_Ids. In the bar chart, for identification, we gave blue color to 'Discovery' set and orange color to 'Validation' set. From the bar chart we have seen that the total gain in 'Discovery' dataset is approximately 86% and in 'Validation' dataset the total Gain is around 92%. On the other hand, the total loss in 'Discovery' dataset is around 36% and the total Loss in 'Validation' dataset is approximately 40%.

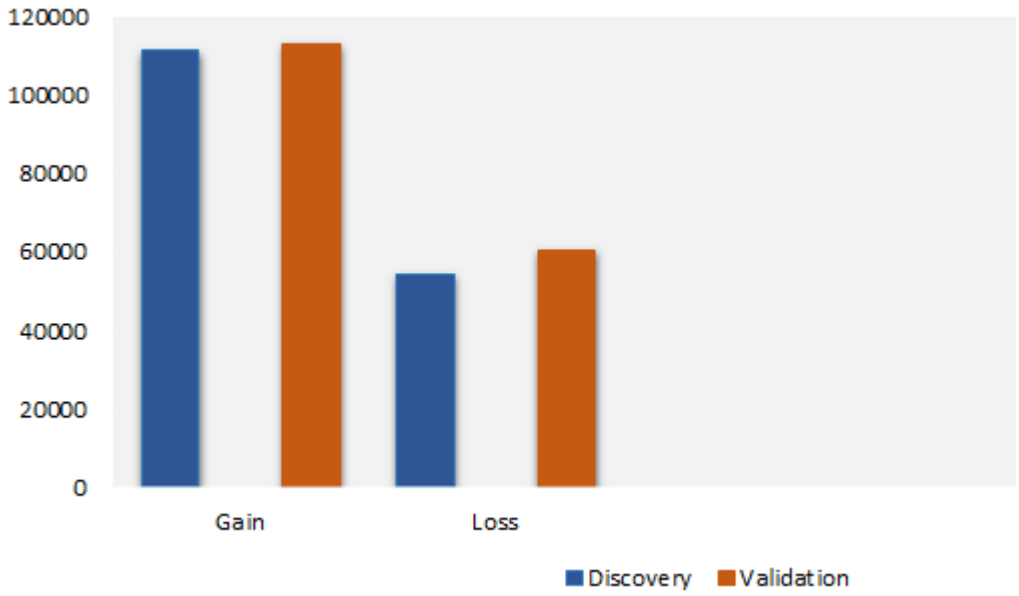

Figure 5.2: Total number of Gain and Loss For Discovery and Validation

Besides, we have also generated two curved graphs for 'Top 200 Gain' as shown in Figure [4.3] and 'Top 200 Loss' as shown in Figure[4.4] for both datasets. From the graph 'Top 200 Gain' we found out that the Gain in 'Validation' dataset is greater than that of 'Discovery' dataset. Likewise, the Loss in 'Discovery' dataset is smaller than that of 'Validation' dataset.

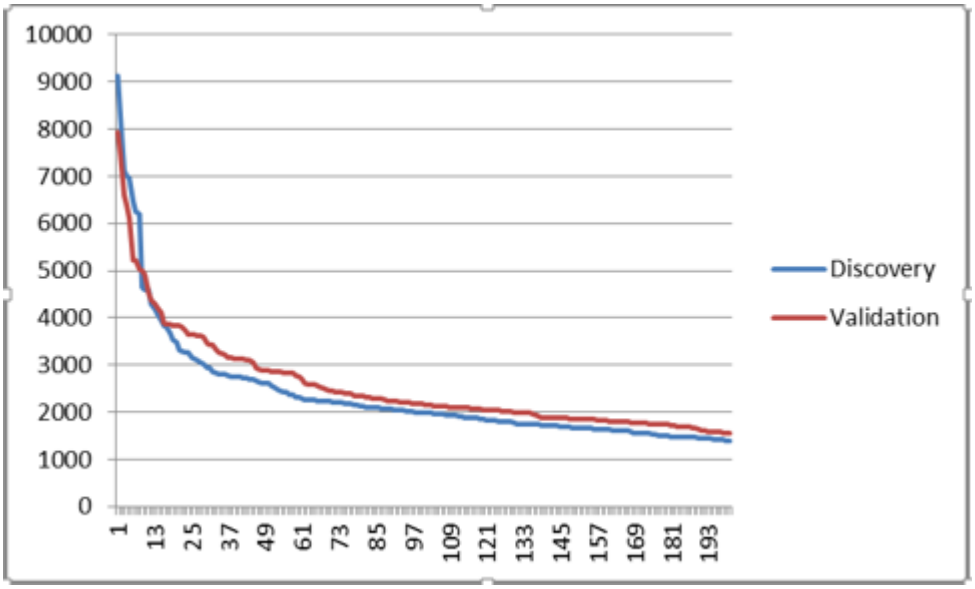

Figure 5.3: Top 200 Gain

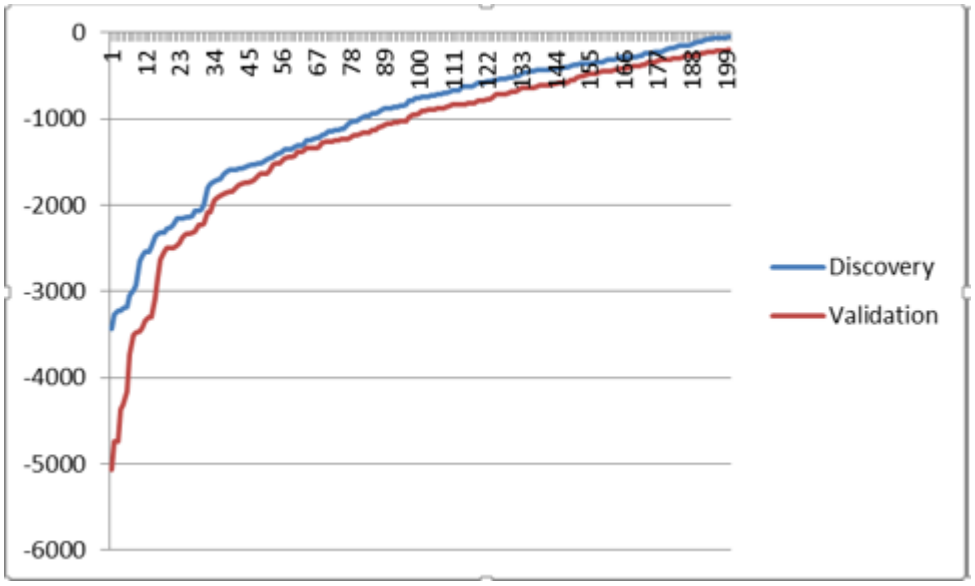

Figure 5.4: Top 200 Loss

5.2 Pairwise weighted gene similarity calculation

We have done the similarity calculation of 27,475 gene interactions for 'Discovery' dataset and for 'Validation' dataset the similarity calculation is of 27,298 gene interaction. We have shown the similarity calculations in the form of tables for both the datasets Table [4.1, 4.2]. From the table we can see that for same type of pairwise gene similarity calculation the values in 'Discovery' dataset and 'Validation' dataset are different. For example, for genes CDH2 and CDH11 the similarity value for 'Discovery' dataset is 0.72868235. However, for these two genes the similarity value for 'Validation' dataset is 0.758765207. These similarity values are weighted vales and thus our method

is weighted. This weighted methodology guarantees that the score for every gene are distributed dependending on their mutation frequency over every one of the samples. We used these weighted values as inputs for clustering of gene network where we used K-means clustering algorithm.

Table 5.1: Gene Similarity Calculation For Discovery Dataset

Gene_1	**Gene_2**	**Similarity**
A1BG	GRB7	0.996778745
GRB7	HADHB	0.818046004
HADHB	HADH	0.976888492
A1BG	SMN1	0.690344383
ADA	ADORA1	0.998731322
CDH2	CDH11	0.72868235
AKT3	CDKN1A	0.953360454
MED6	MED16	0.989825387
NR2E3	CDK9	0.999977251
SIGLEC14	TYROBP	0.893485695

Table 5.2: Gene Similarity Calculation For Validation Dataset

Gene_1	**Gene_2**	**Similarity**
A1BG	GRB7	0.99679647
GRB7	HADHB	0.915119363
HADHB	HADH	0.828442649
A1BG	SMN1	0.640730903
ADA	ADORA1	0.999603489
CDH2	CDH11	0.758765207
AKT3	CDKN1A	0.958537211
MED6	MED16	0.994701142
NR2E3	CDK9	0.997438488
SIGLEC14	TYROBP	0.874449109

5.3 K-Means Clustering Algorithm

As mentioned before, the input datasets that we had was too large to work with because of not having sufficient system to run those inputs we categorized those input values from both the files into six categories and made two new csv files. We have shown our collection of data based on the categorized weighted values in Table 4.3. In the table we have shown the total number of genes that we calculated for the weighted values in both the datasets.

Table 5.3: Total number of genes per weighted value regions for Discovery set and Validation set

Weighted Values	**Number of Genes for Discovery**	**Number of Genes for Validation**
between 1 and 0.9	54733	54113
between 0.9 and 0.8	14151	14478
between 0.8 and 0.7	9606	9950
between 0.7 and 0.6	7552	8142
between 0.6 and 0.5	6280	6527
less than 0.5	16010	15091

We know that K-means is an iterative algorithm which completes in two steps. The first step is 'Cluster assignment' step where the algorithm goes through every one of the data points and relying on the nearest cluster centre, the data points are assigned. The second step is known as 'Move centroid' step where the centroids are moved to the means of the data points in a cluster by K-means. As such, the calculation ascertains the average of the considerable number of data points in a cluster and moves the centroid to that mean area. This procedure is repeated until there is no adjustments in the clusters. Here, we have taken the weighted values which were greater and equal 0.9 and 1 as initial starting point selecting from the huge dataset.

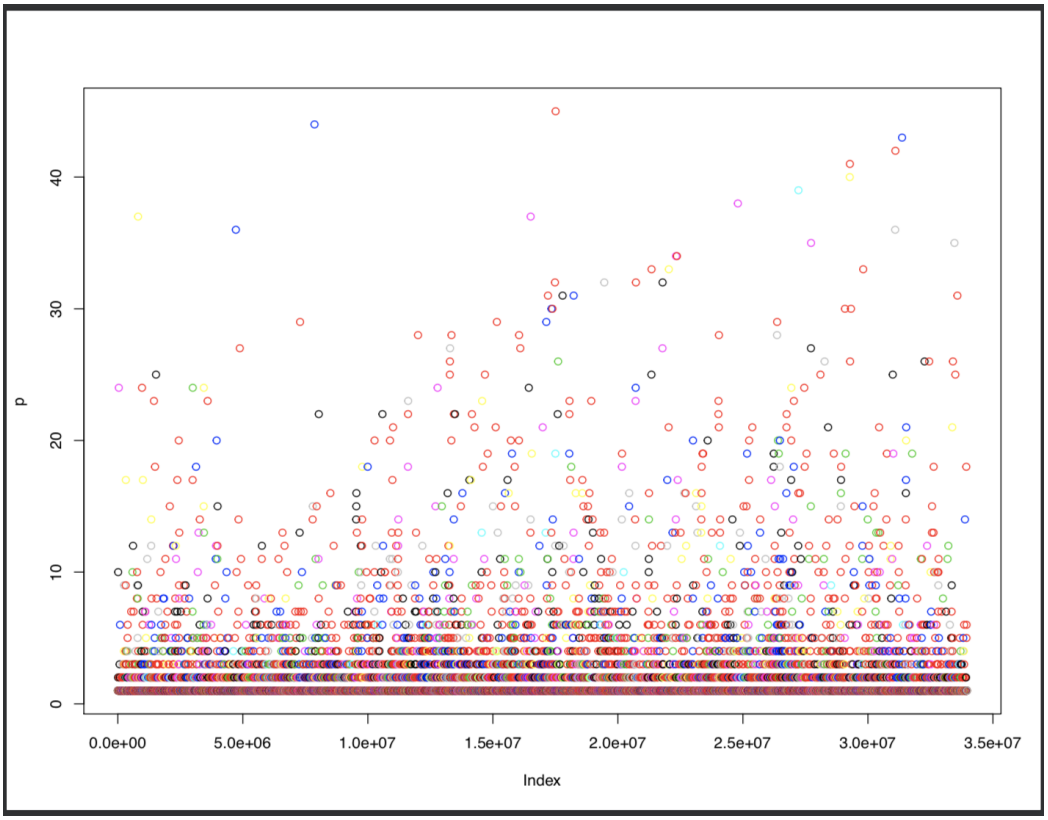

Figure 5.5: Cluster For Discovery Set

We used 'cluster' library on R to get the output clusters. We have shown the clusters of Discovery set and Validation set plotting on graph as shown in Figure [4.5, 4.6]. We can see different colors of dots in the figures. Each color refers to the centre of each cluster. There are 10 colors for 10 different clusters. We have plotted a two dimensional scattered diagrams and it is gene by gene plot. Because of huge dataset and the picture being zoomed out the clusters are found closer to one another.

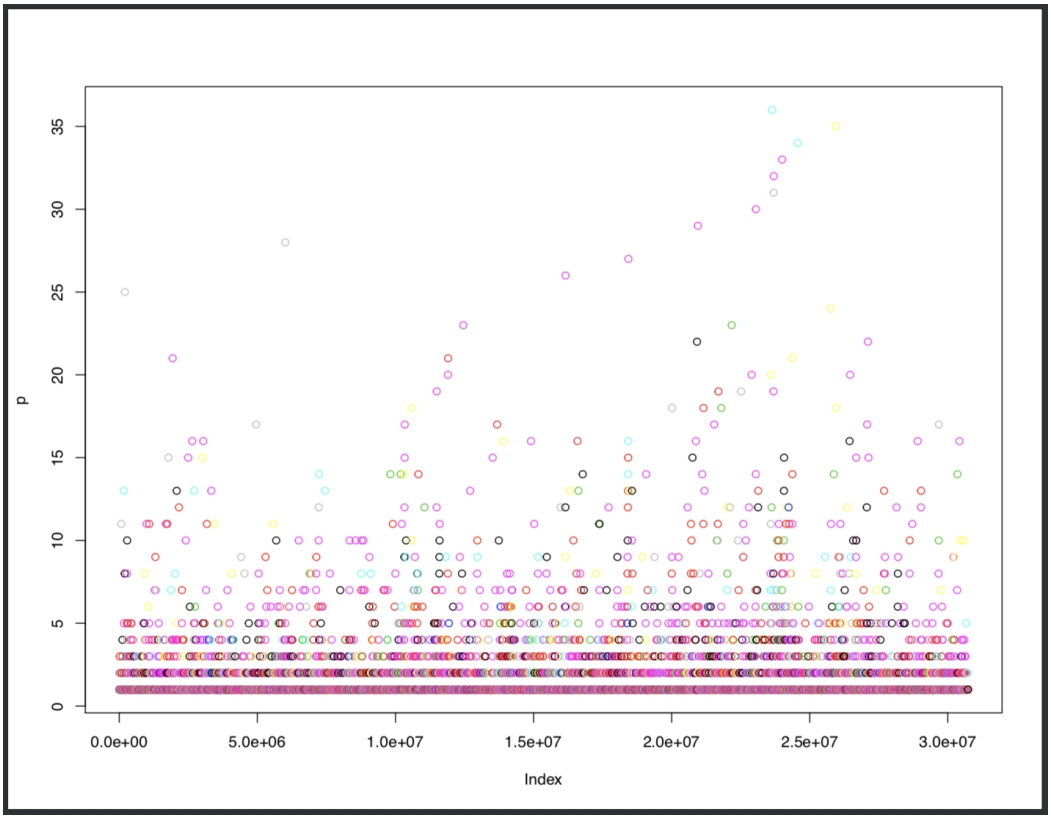

Figure 5.6: Cluster For Validation Set

Chapter 6

Conclusion and Future Plan

6.1 Conclusion

For our work, we have designed a novel bioinformatics analysis pipeline for qualitative classification of the breast cancer genome and clustering of the cancer gene network. We collected datasets which were divided into Discovery dataset and Validation dataset from METABRIC. We retrieved gene names for the cancer patients of the two sets. Next, we found out the total number of CNV mutations in each the genes for all the Metabric_Ids so that we can understand how much the cancer was spreading. After that, we calculated the pairwise similarity of genes for clustering of the cancer gene subnetwork. The advantage of our work is that the values we got after calculation of similarity are weighted values. We used those weighted values as input for identification of mutated subnetworks. Another advantage is that

we have worked with datasets of original cancer patients. The disadvantage of our work is that we used only one clustering algorithm named K-Means. If we could have used more than one clustering algorithm for identification of mutated subnetworks then our work would have been more precise. The analysis pipeline of our paper is a novel bioinformatics pipeline which has never been used to identify qualitative classification of breast cancer genome. The main purpose of our paper was to verify that if the pipeline is working or not on the real dataset that we used. Moreover, we have used weighted analysis which have never been used for clustering of the subnetwork. Weighted approach makes it easier to identify genes that are responsible for cancer. We calculated mutation score for all the genes and used that score to find the weighted values.

6.2 Future Plan

In the future, we will do statistical analysis and survival analysis based on the cluster of the subnetwork and after that we will combine the results of our two data sets and compare between them. Finally, we can say that if our approach is working or not. Moreover, we hope to publish our paper after completion of our research work.

Bibliography

[1] Zhang, J., Chiodini, R., Badr, A., & Zhang, G. (2011). The impact of next-generation sequencing on genomics. Journal of genetics and genomics, 38(3), 95-109.

[2] Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., ... & Cho, E. K. (2006). Global variation in copy number in the human genome. nature, 444(7118), 444.

[3] Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., & Ideker, T. (2007). Network-based classification of breast cancer metastasis. Molecular systems biology, 3(1), 140.

[4] Liu, Z. J. (2008). Aquaculture genome technologies. John Wiley & Sons.

[5] Shastry, B. S. (2009). SNPs: impact on gene function and phenotype. In Single Nucleotide Polymorphisms (pp. 3-22). Humana Press, Totowa, NJ.

[6] Kao, S. L., Chong, S. S., & Lee, C. G. (2000). The role of single nucleotide polymorphisms (SNPs) in understanding complex disorders and pharmacogenomics. Annals of the Academy of Medicine, Singapore, 29(3), 376-382.

[7] Dittmar, K., & Liberles, D. (2011). Evolution after gene duplication. John Wiley & Sons.

[8] Rizzolo, P., Silvestri, V., Falchetti, M., & Ottini, L. (2011). Inherited and acquired alterations in development of breast cancer. The application of clinical genetics, 4, 145.

[9] Braun, L., Mietzsch, F., Seibold, P., Schneeweiss, A., Schirmacher, P., Chang-Claude, J., ... & Aulmann, S. (2013). Intrinsic breast cancer subtypes defined by estrogen receptor signalling—prognostic relevance of progesterone receptor loss. Modern Pathology, 26(9), 1161.

[10] Alluri, P., & Newman, L. A. (2014). Basal-like and triple-negative breast cancers: searching for positives among many negatives. Surgical Oncology Clinics, 23(3), 567-577. Chicago

[11] Dietze, E. C., Sistrunk, C., Miranda-Carboni, G., O'Regan, R., & Seewaldt, V. L. (2015). Triple-negative breast cancer in African-American women: disparities versus biology. Nature Reviews - Cancer, 15(4), 248–254. http://doi.org/10.1038/nrc3896

[12] Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., & Shi, B. (2015). Breast cancer intrinsic subtype classification, clinical use and future trends. American journal of cancer research, 5(10), 2929.

[13] Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., ... & Wilson, R. K. (2012). MuSiC: identifying mutational significance in cancer genomes. Genome research.

[14] Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., ... & Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature, 505(7484), 495.

[15] Chi, M. T., Glaser, R., & Farr, M. J. (2014). The nature of expertise. Psychology Press.

[16] Creixell, P., Schoof, E. M., Simpson, C. D., Longden, J., Miller, C. J., Lou, H. J., ... & Wesolowska-Andersen, A. (2015). Kinome-wide decoding of network-attacking mutations rewiring cancer signaling. Cell, 163(1), 202-217.

[17] Jones, S., Zhang, X., Parsons, D. W., Lin, J. C. H., Leary, R. J., Angenendt, P., ... & Hong, S. M. (2008). Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. science.

[18] Plimack, E. R., Bellmunt, J., Gupta, S., Berger, R., Montgomery, R. B., Heath, K., ... & Perini, R. F. (2015). Pembrolizumab (MK-3475) for

advanced urothelial cancer: Updated results and biomarker analysis from KEYNOTE-012.

[19] Grasso, C. S., Wu, Y. M., Robinson, D. R., Cao, X., Dhanasekaran, S. M., Khan, A. P., ... & Asangani, I. A. (2012). The mutational landscape of lethal castration-resistant prostate cancer. Nature, 487(7406), 239.

[20] Taylor, B. S., Schultz, N., Hieronymus, H., Gopalan, A., Xiao, Y., Carver, B. S., ... & Antipin, Y. (2010). Integrative genomic profiling of human prostate cancer. Cancer cell, 18(1), 11-22.

[21] Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., ... & Zhang, J. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell, 158(4), 929-944.

[22] Zhang, J., & Zhang, S. (2016). The discovery of mutated driver pathways in cancer: Models and algorithms. IEEE/ACM transactions on computational biology and bioinformatics.

[23] Silverstein, R. M., Webster, F. X., Kiemle, D. J., & Bryce, D. L. (2014). Spectrometric identification of organic compounds. John wiley & sons.

[24] Falcon S, Gentleman R (2007) Using GOstats to test gene lists for GO term association. Bioinformatics 23 257–258.

[25] Ciriello, G., Cerami, E., Sander, C., & Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. Genome research, 22(2), 398-406.

[26] F. Vandin, E. Upfal, and B. J. Raphael, "De novo discovery of mutated driver pathways in cancer," Genome Res., vol. 22, pp. 375–385, 2012.

[27] Zhao, J., Zhang, S., Wu, L. Y., & Zhang, X. S. (2012). Efficient methods for identifying mutated driver pathways in cancer. Bioinformatics, 28(22), 2940-2947..

[28] Babur, Ö., Gönen, M., Aksoy, B. A., Schultz, N., Ciriello, G., Sander, C., & Demir, E. (2015). Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. Genome biology, 16(1), 45..

[29] C. H. Yeang, F. McCormick, and A. Levine, "Combinatorial patterns of somatic gene mutations in cancer," FASEB J., vol. 22, pp. 2605–2622, 2008.

[30] Zhang, J., Zhang, S., Wang, Y., & Zhang, X. S. (2013). Identification of mutated core cancer modules by integrating somatic mutation, copy number variation, and gene expression data. BMC systems biology, 7(2), S4.

[31] Zhang, J., Wu, L. Y., Zhang, X. S., & Zhang, S. (2014). Discovery of co-occurring driver pathways in cancer. BMC bioinformatics, 15(1), 27

[32] Hou, J. P., Emad, A., Puleo, G. J., & Milenkovic, O. (2016, August 18). A new correlation clustering method for cancer mutation.

[33] Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., ... & Stuart, J. M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics, 26(12), i237-i245.

[34] Vandin, F., Upfal, E., & Raphael, B. J. (2012). De novo discovery of mutated driver pathways in cancer. Genome research, 22(2), 375-385.

[35] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ... & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102(43), 15545-15550.

[36] Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., ... & Gräf, S. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature, 486(7403), 346.

[37] Margolin, A. A., Bilal, E., Huang, E., Norman, T. C., Ottestad, L., Mecham, B. H., ... & Vollan, H. K. M. (2013). Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. Science translational medicine, 5(181), 181re1-181re1.

[38] Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., ... & Bardou, P. (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic acids research, 43(W1), W589-W598.

[39] Jiang, M., Chen, Y., & Chen, L. (2015). Link prediction in networks with nodes attributes by similarity propagation. arXiv preprint arXiv:1502.04380.

[40] Na, S., Xumin, L., & Yong, G. (2010, April). Research on k-means clustering algorithm: An improved k-means clustering algorithm. In 2010 Third International Symposium on intelligent information technology and security informatics (pp. 63-67). IEEE.

[41] Fahim A M,Salem A M,Torkey F A, "An efficient enhanced k-means clustering algorithm" Journal of Zhejiang University Science A, Vol.10, pp:1626-1633,July 2006.