Analysing Cancer Similarities Among Heterogenous Cancer Datatypes Using Similarity Network Fusion

Submitted by

Abu Jafar Md Asadullah

2013-2-60-016

Supervised by

Md Sarwar Kamal

Senior Lecturer

Department of Computer Science & Engineering (CSE)

A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering

Department of Computer Science & Engineering (CSE)

Faculty of Sciences and Engineering

East West University, Dhaka, Bangladesh

August, 2017

Analysing Cancer Similarities Among Heterogenous Cancer Datatypes Using Similarity Network Fusion

Submitted by

Abu Jafar Md Asadullah

2013-2-60-016

Supervised by

Md Sarwar Kamal

Senior Lecturer

Department of Computer Science & Engineering (CSE)

A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering

Department of Computer Science & Engineering (CSE)

Faculty of Sciences and Engineering

East West University, Dhaka, Bangladesh

August, 2017

DECLARATION

I, hereby, declare that the work presented in this thesis is the outcome of the work performed by me under the supervision of Md Sarwar Kamal, Senior Lecturer, Department of Computer Science and Engineering, East West University, Dhaka, Bangladesh. I also declare that no part of this thesis has been or is being submitted elsewhere for the award of any degree or diploma. This work complies with the regulations of this University and meets the accepted standards with respect to originality and quality. I authorize the University and other individuals to make copies of this work as needed for scholarly research.

Candidate’s Signature

. . . .

Abu Jafar Md Asadullah

ID: 2013-2-60-016

Department of Computer Science and Engineering

East West University

LETTER OF ACCEPTANCE

This Thesis Report entitled **“Analysing Cancer Similarities among Heterogenous Cancer Data Types Using Similarity Network Fusion”** submitted by Abu Jafar Md Asadullah (ID: 2013-2-60-016) to the Department of Computer Science and Engineering, East West University, Dhaka, Bangladesh is accepted by the department in partial fulfillment of requirements for the Award of the Degree of Bachelor of Science and Engineering on August, 2017.

Supervisor

. . . .

Sarwar Kamal
Senior Lecturer, Department of Computer Science and Engineering
East West University
Dhaka, Bangladesh

Chairperson

. . . .

Dr. Md. Mozammel Huq Azad Khan
Chairperson and Professor, Department of Computer Science and Engineering
East West University
Dhaka, Bangladesh

ACKNOWLEDGMENTS

First of all, I would like to thank almighty Allah for giving me the strength, patience and knowledge to complete this Thesis Research.

I would like to express my sincere gratitude to my supervisor Sarwar Kamal for the continuous support, for his patience, for motivating me and most importantly keeping faith on me. His expert guidance helped me through time of research and completion of this thesis. I could not have imagined having a better supervisor and mentor for this Research.

In addition, I would like to express my gratitude to the faculty members of the Department of Computer Science and Engineering (CSE), East West University, Dhaka. who helped me with their feedback.

I would to expresse my deepest gratitude to my friends who supported me when I was about to collapse. Their belief on me gave me strength to keep going.

Last but not least, I would like to thank my family, my parents and to my brother and sister for supporting me.

Abu Jafar Md Asadullah

ID: 2013-2-60-016

ABSTRACT

Patterns from different datatypes can reveal interesting relationship when they are integrated together. In this research a method of aggregating different data types The similarity network fusion was used over five different cancer data. This technique fuses networks from different cancer data according to their similarities to reveal their patterns. A clustering method called spectral clustering was used to find the patterns in the integrated network. Heatmaps were used to visualise patient to patient relationship and the clusters formed, were the subtypes of patient with similar genetic profile. Similar patient subtypes can be studied for further analys and for medical discoveries.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

KNN	K Nearest Neighbour
SNF	Similarity Network Fusion
GBM	Glioblastoma Multiforme
BIC	Breast Invasive Carcinoma
KRCCC	Kidney Renal Clear Cell Carcinoma
SC	Spectral Clustering
LSCC	Lung Squamous Cell Carcinoma
COAD	Colon Adenocarcinoma

CHAPTER 1

1.1 **Introduction**

A tumor is out of control cellular growth that can be resulted in either Benign tumor or Malignant tumor. Benign tumor doesn't invade other tissue and causes cancer but it keeps growing that can result in many complications. Malignant tumor on the other hand is aggressive kind of tumor that invades other tissue and results in cancer. It's said that Cures for a disease relies on the understanding of its biology. Mutations is error in genetic code. Mutations occur when protein makes a mistake while replicating or environment caused mutations or mutation caused by certain diet. If we look further deep into cancer biology we will see certain mutations in gene is responsible for causing cancer.

Gene is the fundamental unit of DNA that encodes for certain protein that carries out all the necessary works in biological process associated with different types of RNA. DNA on the other hand keeps all the records of knowledge preserved as gene. Transcriptome is collection of mRNA that is generated by the process of transcription. Even though all our cells have the exact same DNA but they are all different in nature. That phenomenon is called epigenome where certain gene can be turn off by methylation (adding a methyl group) or turn on by adding acetyl group. The reason why cancer is really hard to cure is each cancer signature is different even between patient of similar cancer, at the same time mutation between two tumors in single patient is different. With the help of recent technology analysis of transcriptome, epigenome helped revealing various information's about cancer. A major breakthrough was discovery of drug called Gleevec that is used to treat a type of blood cancer called chronic myeloid leukemia. As an attempt to contribute to understand cancer biology and discover new relationships among the diverse cancer genome data, I have tried Similarity Fusion

Network to combine data from multiple domain of cancer and analyzed using spectral clustering to yield hidden patterns.

1.2 **Aggregating heterogenous data types**

The main problem of working with Genetic data is the noise in it that often leads to inconsistent decisions[1]. Although there are many methods combining supervised and unsupervised learning, it's easy to miss significant patterns. But if our search doesn't depend on only one type of data (let's say gene expression data), instead we analyze different types of data (let's say gene expression data, the methylation data, miRNA data and siRNA data) and integrate them all together will help us discover new patterns. The same technique I intend to use over multiple types of cancer domain and find out their genetical similarities.

1.3 **Motivation**

Cancer is leading cause of death worldwide. According to WHO there was 8.8 million deaths in 2015[6]. Due to researches and new discovery every day we managed to develop screening process that can easily identify a cancer, but most of the time someone gets diagnosed with cancer its already spread across body. The traditional methods to treat cancer is:

1. Surgery meaning removal of the tumor and nearby lymph nodes

2. Concentrated radiation on targeted location to damage infected cancer cell beyond repair so that it eventually dies.

3. Chemotherapy that targets fast growing cells, as cancer is a fast-growing cell. But it massively damages healthy cells that under goes fast growing process like hair. It damages healthy cells more than it kills cancer cell.

There are many study going on to discover new methods of treatment. I believe my research data will help them in developing new drugs or methods to cure cancer.

1.4 **Research Questions**

I am proposing Similarity Fusion Network can be used find similarities among cancers. Similarity Fusion Network is a great tool to find cancer subtypes I will use to find similarities among different cancers.

1.4.1 **Are there any similarities among cancer?**

Here is some behavioral similarity of cancer. All cancer undergoes these 4 phases.

Cellular proliferation: In life cycle of a cell is 90 percent interphase (cellular growth) and 10 percent mitosis (cell division). A normal cell grows slowly, once they have grown enough it stops growing or it divides itself if it’s necessary. There are receptors that maintains these operations. These receptors can be activated by externally or internally by bonding with a ligand. A ligand is a substance that forms a complex with a biomolecule to serve a biological purpose. Some cancer cell generates these ligands internally to control its own growth. Other cancer cell becomes more sensitive to these ligands due gene amplification. Gene amplification occurs when a specific gene expresses itself a lot or over expression. Due to over expression of those receptors those cells can grow and divide more than a regular cell finally resulting cancer.

Loss of Apoptosis: Apoptosis is a method of controlled cell death that maintains balance between cell growth and cell death. It’s a major key factor in cancer biology. Whenever a cell is damaged it undergoes apoptosis or programmed cell death. This mechanism is also controlled by the receptors that stimulates a certain enzyme called Caspase that breaks down the cell. Over expression of antiapoptotic proteins can hinder this process and leave a damaged cell alive. And increase in cell growth and reduction in cell death destroys the balance and causes it to grow as tumor. Another most important factor in this process is maintained by the p53 gene that is called the guardian of the of genome. Genome is whole DNA sequence. It regulates the cell cycle if any damage to DNA is detected it halts the growth of that cell and initiates repair mechanism. However, if the DNA is beyond repair it initiates apoptosis. But when it becomes corrupted it fails to complete its duty. Since this gene is very unstable it's a most common gene in cancer biology.

Metastasis: Metastasis is invasion of cancer cells into healthy normal cells. This property gives the cancer cell power to escape from its origin and invade other parts of the body. Researchers are trying to understand this property well to stop cancer cell to invade other portions.

Angiogenesis: Tumor cells generates different factors to degrade the Extra cellular matrix and dissolve the membrane that releases angiogenetic signal to grow blood vessels that can provide extra nourishment for the tumor. With extra nourishment cancer cells can grow more and invade other cells.

Now let's look into the physical similarities of cancer. Cancers are divided into different subgroups according which tissue it derives from. Here is the types of cancer and which tissue it derives from.

Carcinoma: Carcinoma is a type of cancer that starts in epithelial cells. It is the most common cancer type. It can be broken down into further subgroups.

1. Basal cell carcinoma: Is a type of skin cancer found in the outer layer of the skin.
2. Squamous cell carcinoma: Is a type of cancer that can be found both in skin and outer layer of organs. Like lung squamous cell carcinoma.
3. Renal cell carcinoma: Is the most common type of kidney cancer.
4. Ductal carcinoma: can be found inside ducts.
5. Adenocarcinoma: It starts in the glandular cells of the body. Like pancreatic ductal adenocarcinoma is the most common type of pancreatic cancer.

Sarcoma: Cancer arising from connective tissue. It has become the new attention of cancer research.

Lymphoma and Leukemia: These two class arise from hematopoietic cell that leaves the marrow and tends to mature in the lymph nodes and blood. Mostly seen in the children.

Germ cell tumor: Cancer deriving from pluripotent cell that presenting in the testicle(seminoma) or dysgerminoma.

After studying the following characteristics both behavioral and physical similarity among cancer I strongly believe relations can be derived from different cancer types.

1.4.2 **Why Similarity Fusion Network?**

Similarity fusion network allows us to integrate patterns of data from different sources. It follows the similarity patterns in different data sets and with every iteration is becomes more similar to each other. Whereas concatenation between data set is not a good idea because it introduces more noise. Again, it may lead to loss of important information's due to incompatible dimensionality. At the same time SNF proved to be very efficient in data integration.

1.5 **Overview**

My main goal of this research is showing the relationships between similar types of cancers. Mainly showing the relationship between Carcinoma if there exist any. My work here has four parts: first one is to collect data and cleanse the data. By data cleansing its meant the selection of the data and removal of data with missing values and normalizing the data. Second step is to integrate those data into a single matrix.

Then the third step is to cluster them using a known clustering technique. And finally analyzing the relationships discovered in the process.

The next chapters will be followed by

Chapter 2 where I will discuss about the background of the method used and the terminologies and the literature review.

Chapter 3 I will discuss the working procedures and methodology I will talk about the data types I am working with their description and the work flow of my procedure.

Chapter 4 will be consisted of experimental results graphs, and tables.

Chapter 5 will include a short summary of the paper and further scopes of future works.

CHAPTER 2: LITERATURE REVIEW

2.1 **Literature Review**

As we all know data integration is an important process in knowledge discovery (KDD) that allows us to integrate data from heterogeneous sources. Thus, comes many challenges. Real life data is not only noisy but may also be incomplete or inconsistent. There may be present Collection bias or differently scaled data. As insight to this Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains and Anna Goldenberg on their paper Similarity Network Fusion for Aggregating data types on a genomic scale proposed a method that can combine heterogeneous data[1] that made the major contribution in this paper. As for other approaches like consensus clustering that requires preselection of the important gene from each data source that leads to biased analysis. It is also mentioned in the paper Critical limitations of consensus clustering in class discovery by George Machilids & Jun Z. Li. They showed how consensus clustering is limited[10]. Consensus clustering performs better in simulated unimodal distribution whereas real life data may or may not follow them. The problem with finding k number of cluster is also very difficult and It also yields ambiguous relations between them. Finally, the gene to gene correlation among most discriminant gene makes it easy to validate any number of k which may deviate us from finding the optimal result. Another popular method is icluster that uses joint latent variable model for integrative clustering but its sensitive to prior gene selection and doesn't cover the full spectrum[1]. There was another method described by Matan Hofree , John P Shen, Hannah Carter, Andrew Gross and Trey Ideker in their paper network-based stratification of tumor mutations[9]. That analyses the somatic tumor genome sequence. Since mutations in two tumors is rarely similar they used Network based stratification to integrate somatic tumor genome into gene network.

Spectral clustering is a method of clustering that uses eigen values corresponding to the similarity matrix to cluster the data. I will be using same clustering method to group our objects. There are other clustering algorithms and comparisons I found in paper “Clustering cancer gene expression data: a comparative study by Marcilio CP de Souto, Ivan G Costa, Daniel SA de Araujo, Teresa B Ludermir and Alexander Schliep”[8]. And there was another paper called “Clustering Algorithms: Their Application to Gene Expression Data by Jelili Oyelade, Itunuoluwa Isewon, Funke Oladipupo, Olufemi Aromolaran, Efosa Uwoghiren, Faridah Ameh, Moses Achas, and Ezekiel Adebiyi,” helped me to understand the unsupervised clustering methods that is used in Bioinformatics. The difference between a hard cluster and soft cluster. When a point is only given to one single cluster is called a hard cluster and when a single point is given to multiple clusters is known as soft cluster. A soft cluster can be converted into a hard cluster by assigning overlapping points to the dominating cluster. A dominating cluster is the one with higher number of point.

2.2 **Background**

Similarity network fusion works by calculating similarity for each pair of samples to construct a sample by sample similarity matrix for each available datatype. I will discuss later the methods that can be used to calculate similarity matrices. Then the similarity matrix is used to calculate the affinity matrix that will represent the data as a graph. The equation used to calculate the similarity is scaled exponential similarity kernel to determine the weights of each edge.

$$W(i,j) = e^{\left(-\frac{\rho^2(xi-xj)}{\mu\varepsilon(i,j)}\right)} \text{ ……….(1)}$$

Where W represents the graph. $\rho^2(xi - xj)$ is the Euclidean distance. It will be changed if other distance measure is used. μ is a hyperparameter which is ranged [0.3-0.8] and can be empirically set. The equation to eliminate the scaling problem is as follows,

$$\varepsilon(i,j) = \frac{mean\left(\rho^2(xi-Ni)\right)+mean\left(\rho^2(xj-Nj)\right)+mean\left(\rho^2(xi-xj)\right)}{3} \ldots\ldots \quad (2)$$

Then the local affinity is calculated by the K nearest neighbor(KNN). Since the affinity graph for each data type is calculated it is combined using nonlinear method based on message passing theory. This nonlinear method makes those networks similar with every iteration until they convergences into one single network.

After the networked are fused together they can be grouped together using clustering techniques I will discuss later.

2.3 **Similarity Measurements**

There are many methods available to calculate similarity between two points. Some of them are discussed below.

Euclidean distance is a method that calculates straight line distance for each pair for points that can be represented in later in matrix. The equation as follows

$$d(p,q) = \sqrt{\sum_1^n (pi - qi)\text{^}2} \ldots\ldots\ldots\ldots(3)$$

This equation calculated distance for n- dimensional data. Since we are working with data of high dimension we will be needing it. By high dimensional data it means the number of genes expressed in each individual. Human genome is consisted of roughly 20000 genes.

Other distance measures like Chebyshev that defines distance as the maximum distance between those two points. And Manhattan distance defines distance as the total distance between each point.

In microarray data other distance measures like Pearson correlation distance, spearman correlation distance or the Kendal tau correlation distance. The other two distances like spearman correlation distance and Kendal tau correlation distance has better statistics in computing distance in microarray data. The equations for each distance measurement is,

$$dspear = 1 - \frac{\sum_{i=1}^{m}(x'i - mean(x'))(y'i - mean(y'))}{\sqrt{\sum_{i=1}^{m}(x'i - mean(x')^{2}\sum_{i=1}^{m}(y'i - mean(y'))^{2}}} \ldots\ldots \quad (4)$$

Here is spearmen's ranking correlation equation. Where x'i is the rank of xi and y'i is the rank of yi.

$$dkendall = \frac{C - D}{C + D} \ \ldots\ldots (5)$$

where C is the concordant pair and D is the number of discordant pair. I will be using this equation to evaluate my fused networks.

The term affinity describes relationship. How each point in a graph is related can be determined by their affinity. This property can be used in mining patterns since the points closed to each other has high affinity rather than the points far apart. This same property is used by the spectral clustering method to divide the clusters according to their affinity.

Unsupervised learning is method to group the similar objects where no label has been given. This technique allows us to decide how the data should be categorized rather than the supervised learning method where we try to learn the relationship of target feature with descriptive feature in categorized data. The reason to use clustering methods in genomic data is our knowledge about it still limited. We use different clustering methods to discover new relationships among genes that can be later identified using the existing know genes

CHAPTER 3: Methodology

3.1 **Working Procedure**

I used R as it is one of the most popular tool in data analytics and as well as for visualizing. I used R for several reasons in machine learning since it was a bit known to me instead of other platforms I used it for preliminary assessment of the data. My working procedure has 4 phases.

3.2 **Data Understanding**

There are two types of data available for analysis. One is the micro array data and another is RNA sequencing data. In my experiment, I am considering the microarray data due to its broad availability. But I will discuss both types of data and their advantages and disadvantages.

Microarray data is formed by isolating mRNA from targeted source and then converting it into more stable form of cDNA. Then the cDNA is placed into a micro array chip where a template sequence is present labeled with fluorescent ink. As the sequences bind to the corresponded template it can be read by a laser. Then the gene expression is calculated by their relation. RNA sequencing data on the other is the actual sequence of the data. It’s a long string of RNA sequence. As for similarity, both of these data suffer from background noise and biases. Both data can be analyzed using similar mathematical methods. but when we are searching for novel gene or iso gene RNA sequence data works better since microarray data can’t discriminate or detect iso genes. but as for my research I am using microarray data due to its robust nature in computation.

The data I have collected are three types. One is the gene expression data. that contains the genes and how they are expressed in certain patient. Second data type I am using is the methylation data. methylation data contains information about how the gene is expressed usually adding a methylation group to the nucleotide represses the gene. And finally, the miRNA data regulates most of the gene expression. Those three data types have patterns in each of them, I am going to uses Similarity fusion network to combine those data types to get the final result. Then I am going to combine them again with cancer data from different domain and observe the relationship between those cancer.

3.3 **Data preprocessing**

Real life data is not perfect there are missing values. The scale of the data may not be compatible with the model used. To handle missing data, one can set a threshold so that certain patients with missing data above threshold cannot exist in the data set. Another way to deal with missing data below threshold is using linear regression model or interpolation to predict the missing value. But here I used K nearest neighbor(KNN) to calculate those missing values. I calculated missing value using 5 nearest neighbors in adjacent cell. As for normalization, I used the difference between mean to readjust each point.

After normalization, I transposed those data frames so that we can calculate distance matrix. The data that was available to me wasn't compatible with the functions. The feature(genes) were in rows whereas the patients(point) was in column. So, I transposed the data to make it compatible with the functions.

Now if we look at the data how they are spread we will know their internal relationship. Since all those data I have is very high dimensional I will only show 15 samples of each data and how they are related. I have collected five cancer data and they are

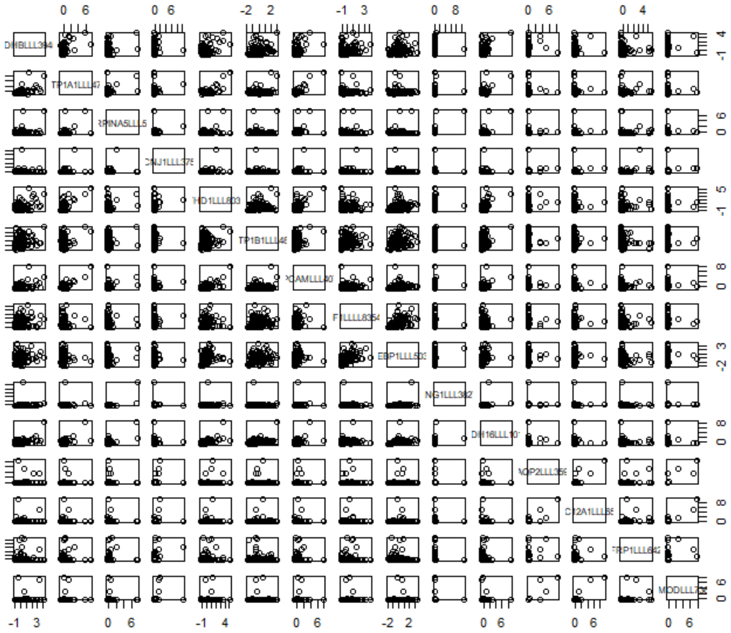

Figure 3.1Gene to gene mapping in gene expression data of

KRCCC

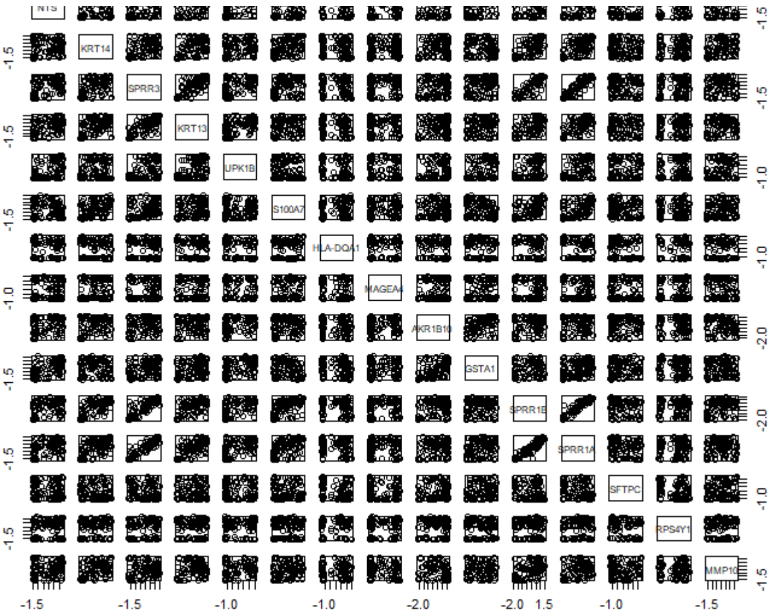

Figure 3.2Gene to gene mapping in gene expression data of

LSCC

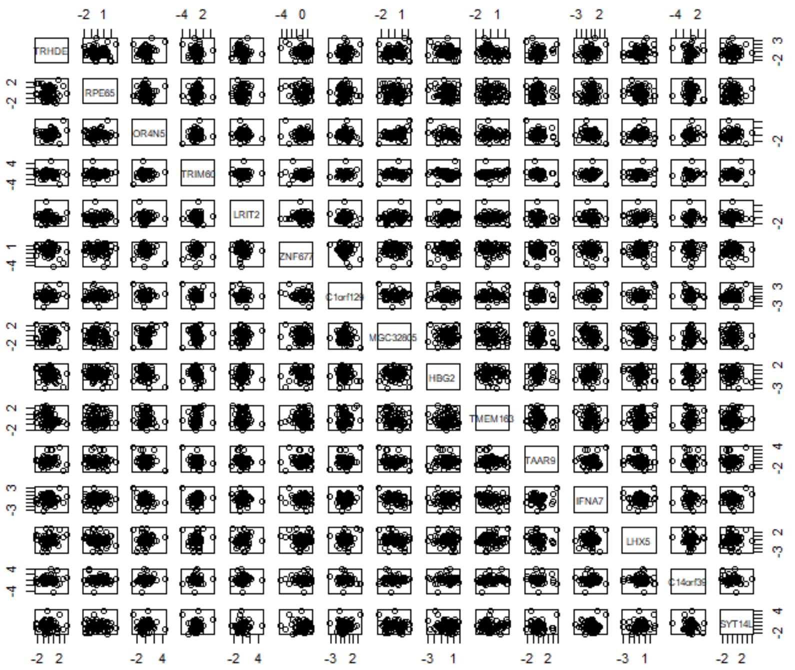

Figure 3.4Gene to gene mapping in gene expression data of

COAD

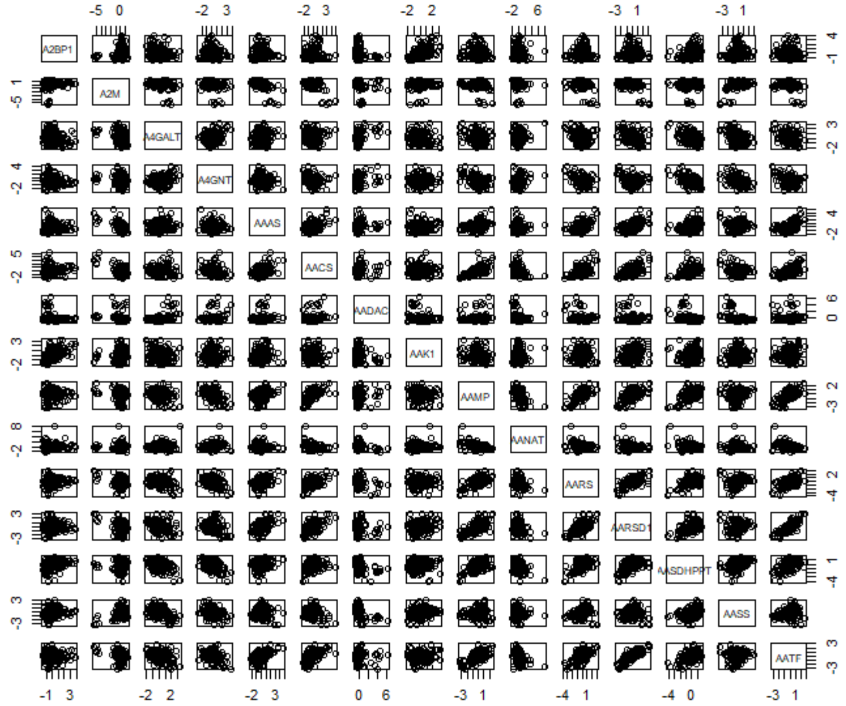

Figure 3.3Gene to gene mapping in gene expression data of

Glioblastoma Multiforme

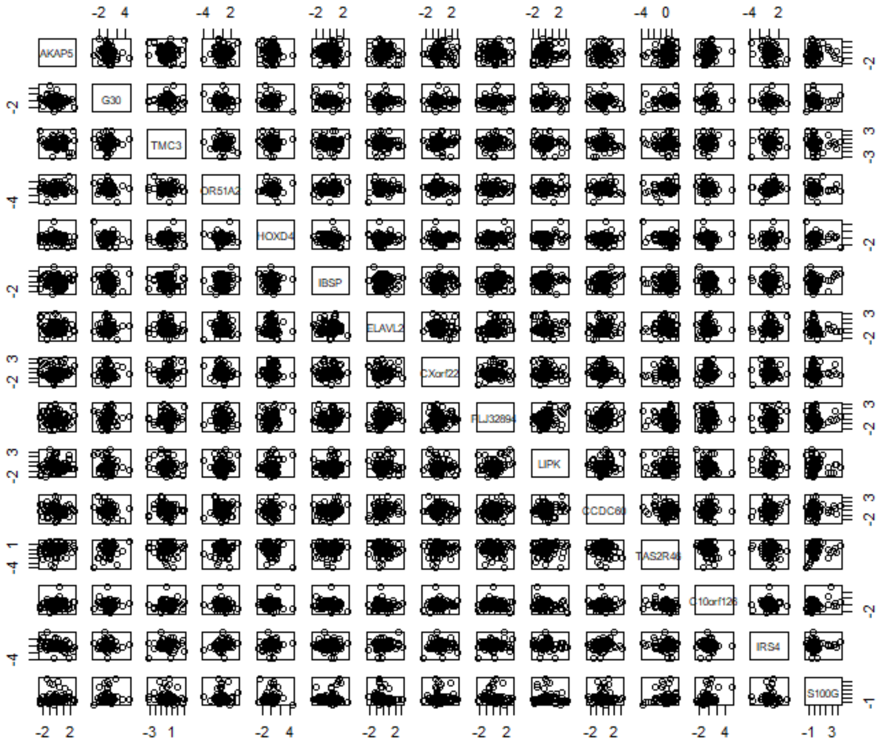

Figure 3.5Gene to gene mapping in gene expression data of BIC

I calculated the Euclidean distance matrix between them. Then I constructed similarity graph by calculating their affinity matrix. Then I fused those matrices from different types of data into one matrix. Then I took different combinations of cancer groups like KRCCC with LSCC, COAD with BIC and clustered them using Spectral Clustering. Here is flowchart of my procedure:

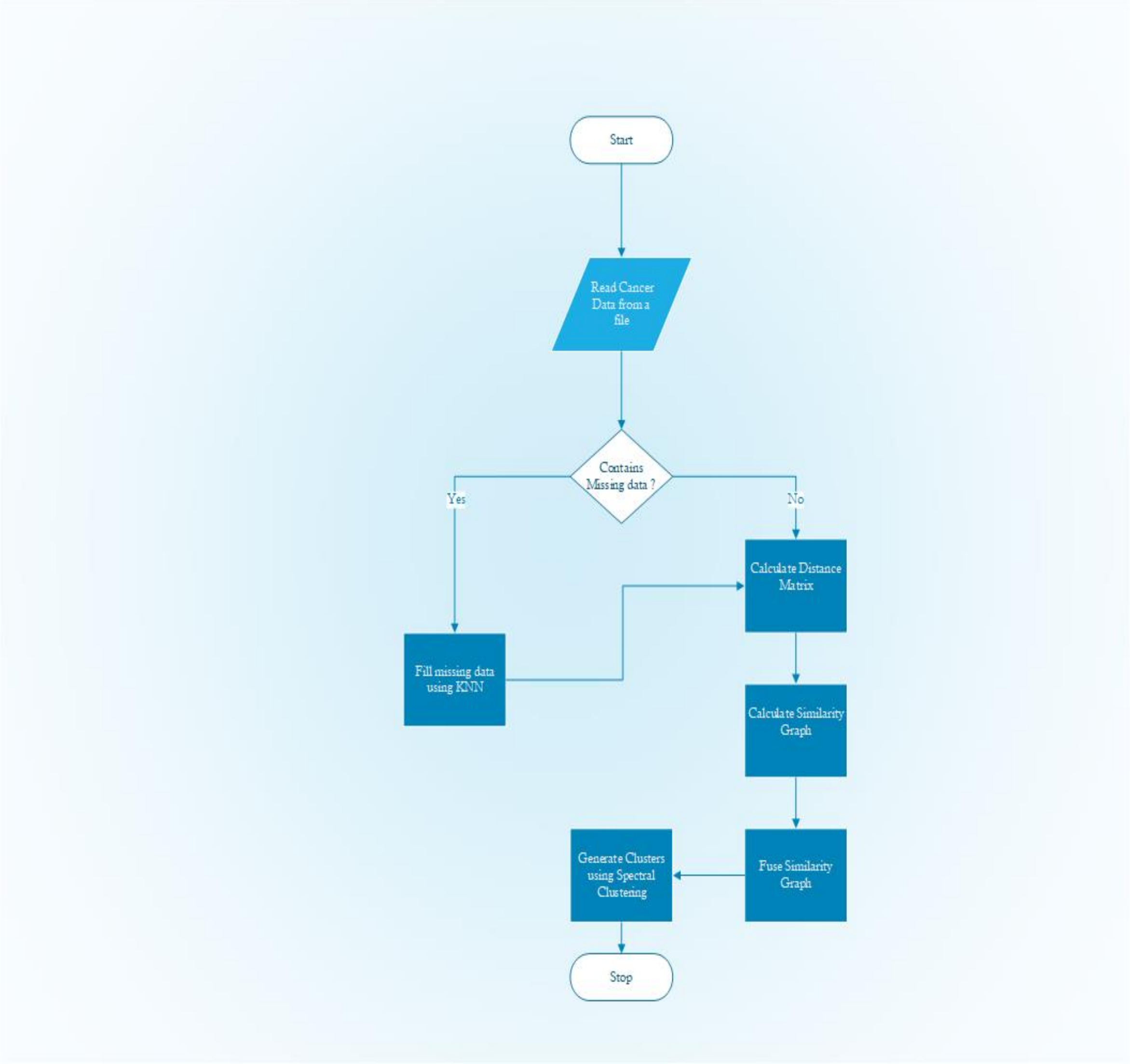
Start
Read Cancer Data from a file
Contains Missing data ?
Yes
No
Calculate Distance Matrix
Fill missing data using KNN
Calculate Similarity Graph
Generate Clusters using Spectral Clustering
Fuse Similarity Graph
Stop

Figure 3.6Flow chart of my program

I allowed to keep those patients data that had less than 20% missing value and disclosed the ones with more that 20% of missing values in data preprocessing step.

Here is a flow chart of data preprocessing step.

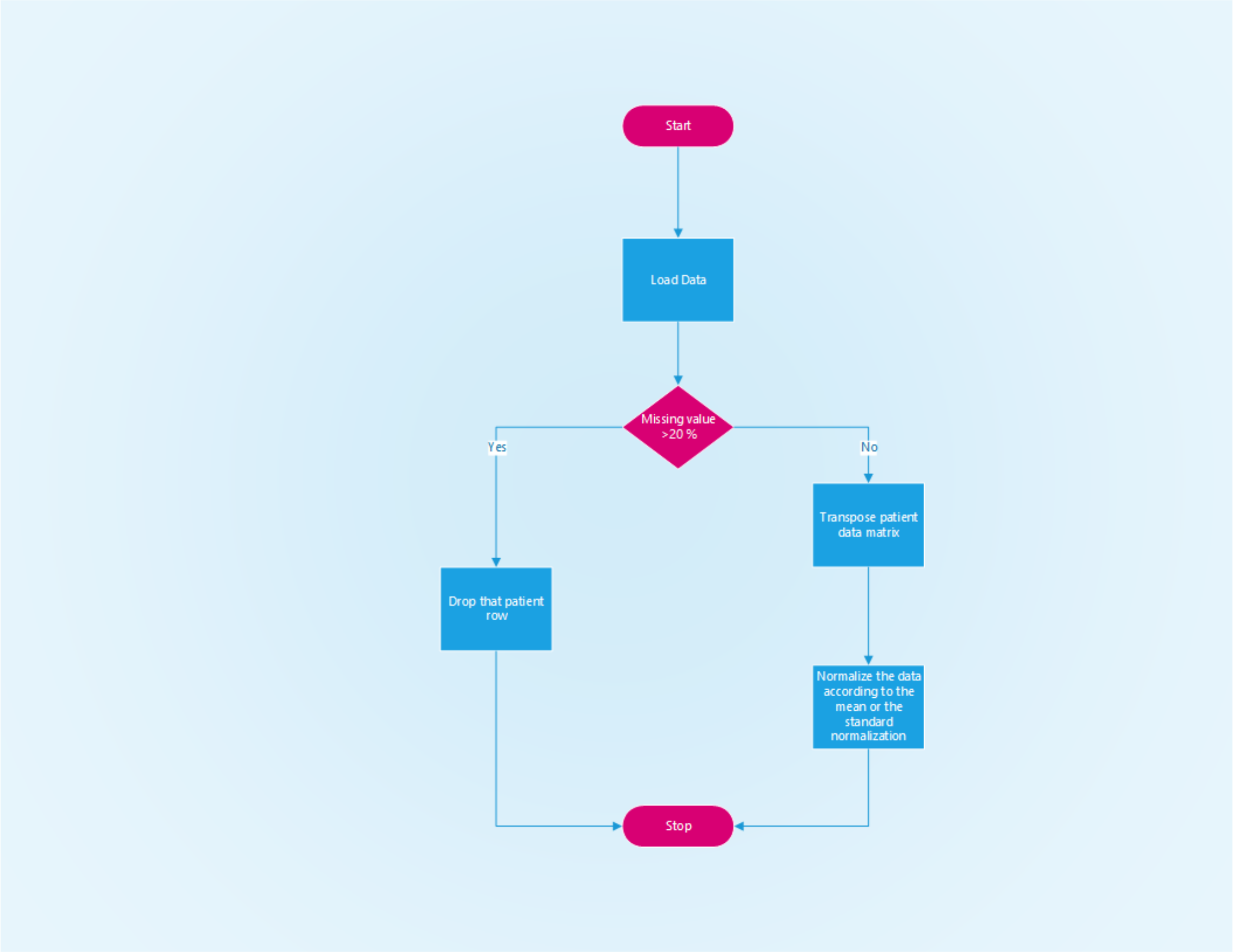
Start
Load Data
Missing value >20 %
Yes
No
Drop that patient row
Transpose patient data matrix
Normalize the data according to the mean or the standard normalization
Stop

Figure 3.7Flow chart of the data preprocessing step.

CHAPTER 4: RESULT AND ANALYSIS

4.1 **Experiment Result**

After conducting the experiment as described in working procedure, I found following relationships from 5 different type of cancers. Before comparing results with real data sets let's see how SNF works in an example data set.

We have two data tables Data1 and Data2. They both have two clusters and they are complementary to each other. Following two figures represents their actual condition, and the third graphs represents how Similarity Network fusion reveals this complementary relationship.

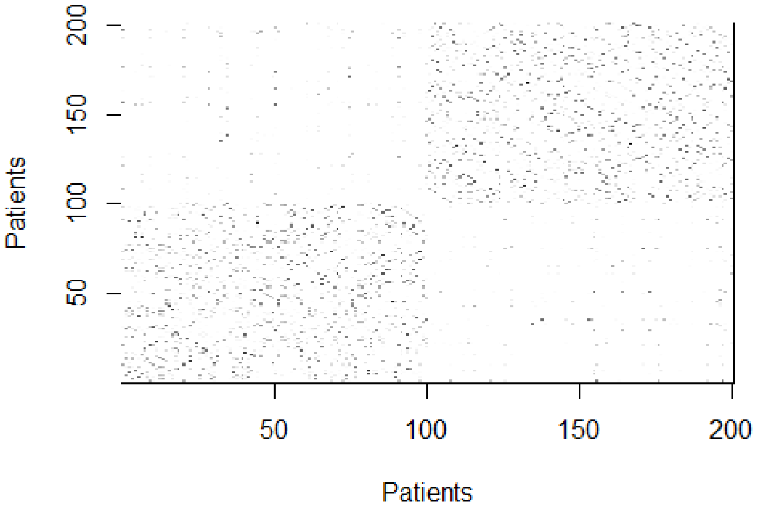

Figure 4.1Two clusters of sample dataset 1

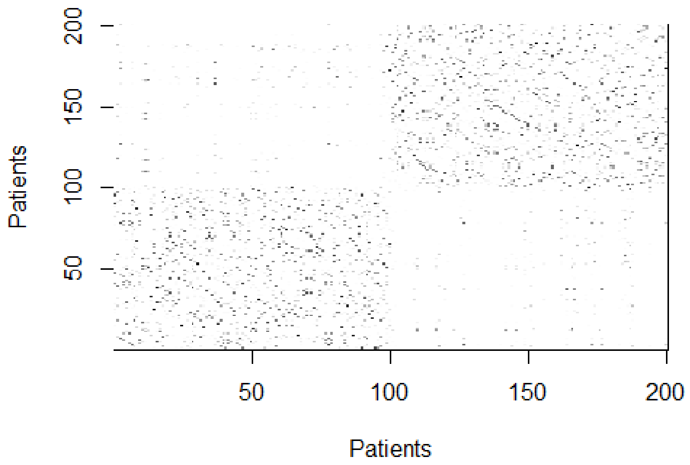

Figure 4.2Two clusters of sample dataset 2

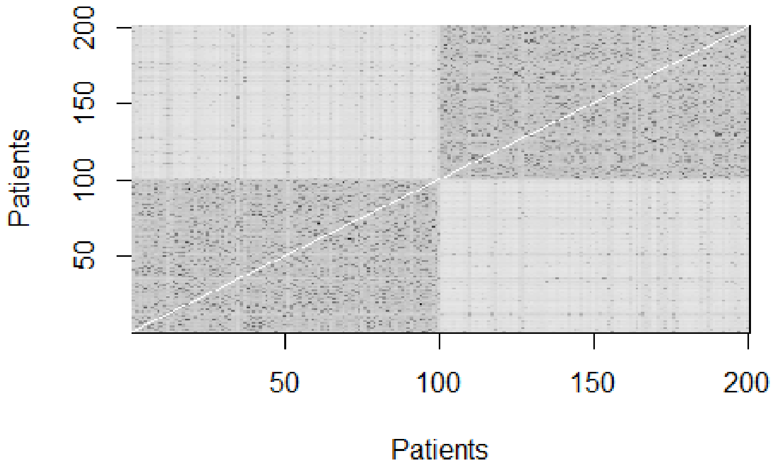

Figure 4.3SNF revealing complementary property of

dataset1 and dataset2 from the graph

Here is a table that shows how the fused data set agrees with other two data set.

Table 4.1Showing the concordance between each pair of Network

	DataSet After Fusion	DataSet1	DataSet2
DataSet After Fusion	1	0.336879878742779	0.127746730109252
DataSet1	0.336879878742779	1	0.0245033133212828
DataSet2	0.127746730109252	0.0245033133212828	1

As we saw from the previous example figures how similarity fusion network reveals the complementary property from both data sets. And the concordance between each network. Now we are going to see how it works on actual cancer data sets. And whether it can reveal the similar properties of different cancer data types. Now, I am going to analyze three different data (Gene Expression Data, Methylation Data, and the miRNA Data) of Breast Invasive Carcinoma(BIC).

Following Three figures represents patterns from each Data types of BIC and the 4th figure represents how the data is clustered after fusing those three data types. Here I am using heatmap to visualize the data.

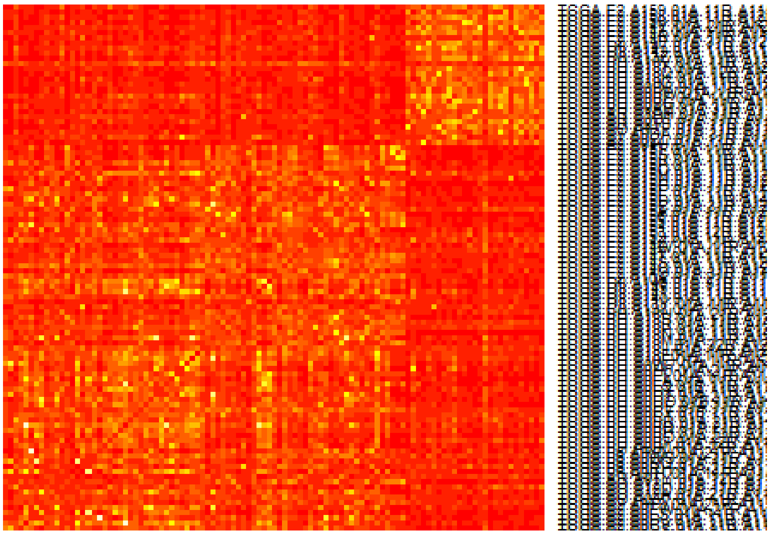

Figure 4.2Heatmap of Patient to patient hierarchy of BIC Gene Expression

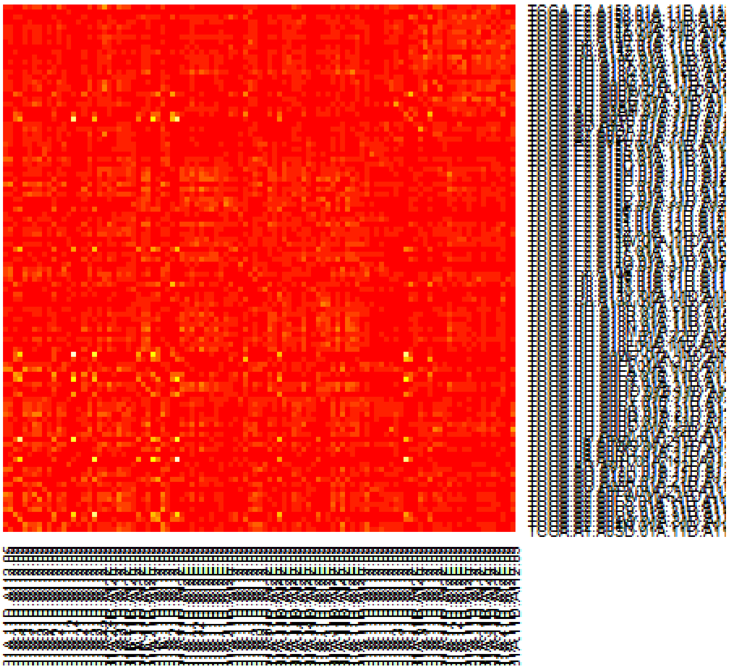

Figure 4.4Heatmap of patient to patient in BIC

Methylation data.

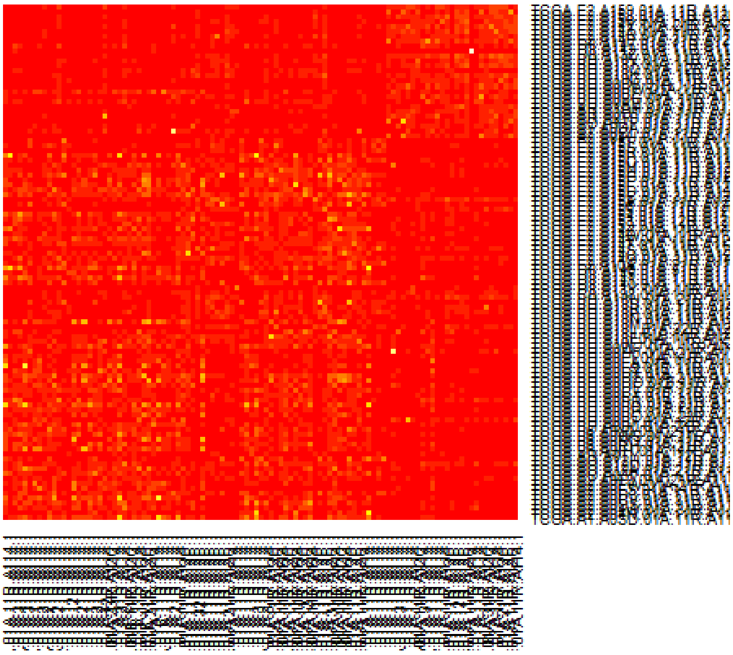

Figure 4.3Heatmap representation of BIC miRNA data

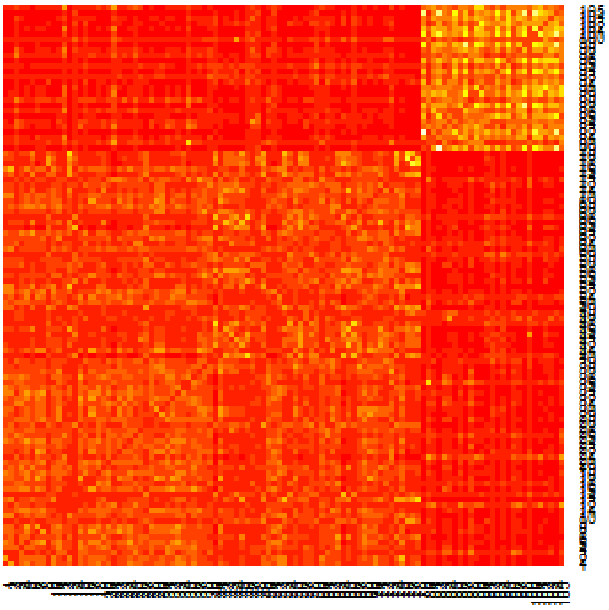

Figure 4.5Heat map representation of BIC after fusing three data types.

As we can see from those comparison how SNF reveals strong patterns from these three data types. For example, in Methylation data it was so noisy that the actual pattern was hardly visible. And in miRNA data the two clusters were somewhat visible. Working with any of those single data types might have led to inconsistent conclusion but after fusion all three similarity networks the final network becomes clearer at the same time it includes all those prior knowledge from those three data types.

Here is the concordance matrix between these four networks.

Table 4.2Showing the concordance between each pair of Network BIC.

	BIC fused network	BIC Gene Expression Network	BIC Methylation Network	BIC miRNA network
BIC fused network	1	0.463560034	0.0039814265	0.3198131181
BIC Gene Expression Network	0.463560034	1	0.0046462509	0.3030847963
BIC Methylation Network	0.003981427	0.004646251	1	0.0009040053
BIC miRNA network	0.319813118	0.303084796	0.0009040053	1

This table shows how the fused network agrees with another network. As we can the concordance between the miRNA and Fused network and the agreement between the gene expression network and the fused network is very higher than the concordance between the methylation network and the fused network. The concordance between methylation network and the fused network shows that no matter how small the value is there is still some components in between these networks are similar. That is indeed the actual purpose of Similarity Fusion Network(SNF).

Now let's look at the results found by fusing networks from 4 different types of cancer data and observe their similarity patterns.

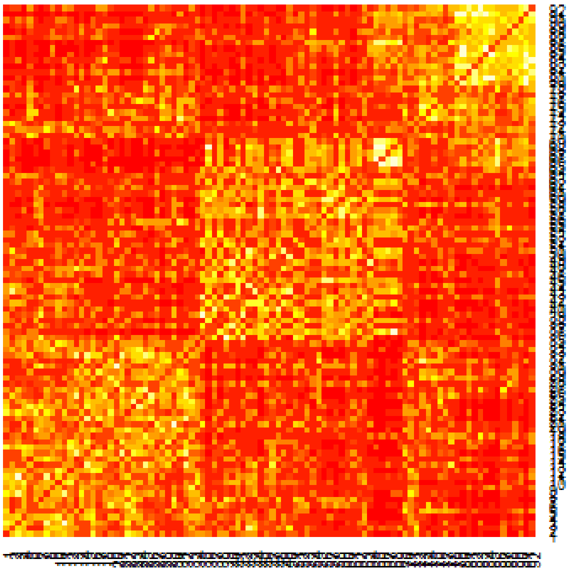

Figure 4.6Clusters between fused network of Colon Adeno Carcinoma and Breast Invasive Carcinoma

Table 4.3Showing Concordance Between Fused network, network of COAD and BIC

	Fused network of COAD &BIC	Network of COAD	Network of BIC
Fused network of COAD &BIC	1	0.26132217	0.2324349
Network of COAD	0.2613222	1	0.01297054
Network of BIC	0.3293585	0.01297054	1

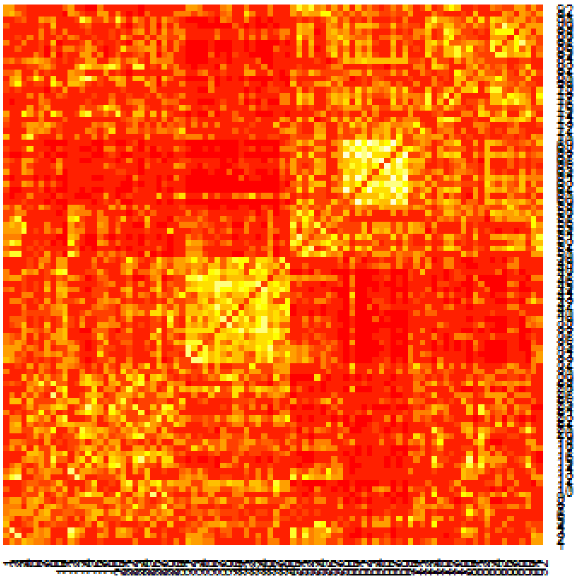

Figure 4.7Clusters between fused network of Lung Squamous Cell Carcinoma and Colon Adenocarcinoma

Table 4.4Showing Concordance Between Fused network, network of LSCC and COAD

	Fused network of LSCC and COAD	Network of LSCC	Network of COAD
Fused network of LSCC and COAD	1	0.08769635	0.32229555
Network of LSCC	0.08769635	1	0.01140381
Network of COAD	0.32229555	0.01140381	1

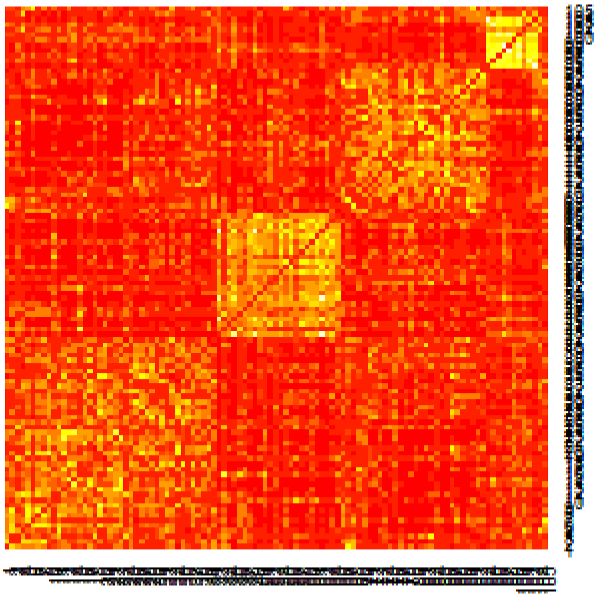

Figure 4.8Clusters between fused network of KRCCC and BIC

Table 4.5Showing Concordance Between Fused network, network of KRCCC and BIC

	Fused network of KRCCC and BIC	Network of KRCCC	Network of BIC
Fused network of KRCCC and BIC	1	0.1525591	0.53661179
Network of KRCCC	0.2840153	1	0.05092961
Network of BIC	0.5366118	0.05092961	1

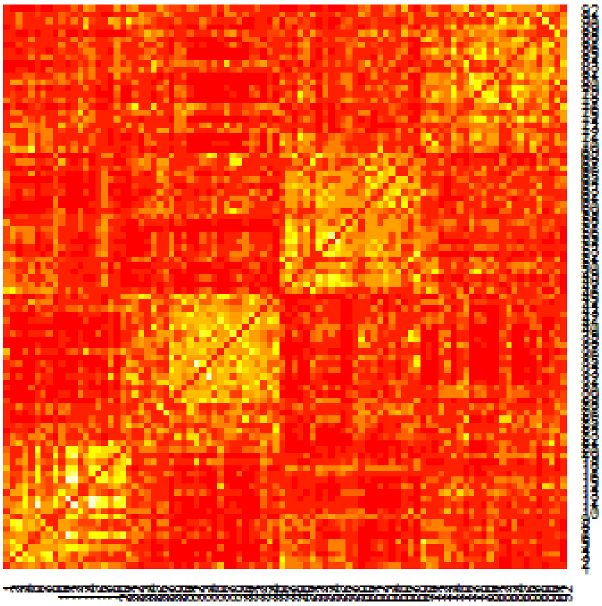

Figure 4.9Clusters between fused network of KRCCC and COAD

Table 4.6Showing Concordance Between Fused network, network of KRCCC and COAD

	Fused network of KRCCC and COAD	Network of KRCCC	Network of COAD
Fused network of KRCCC and COAD	1	0.1476920	0.44813006
Network of KRCCC	0.1476920	1	0.01868906
Network of COAD	0.4481301	0.01868906	1

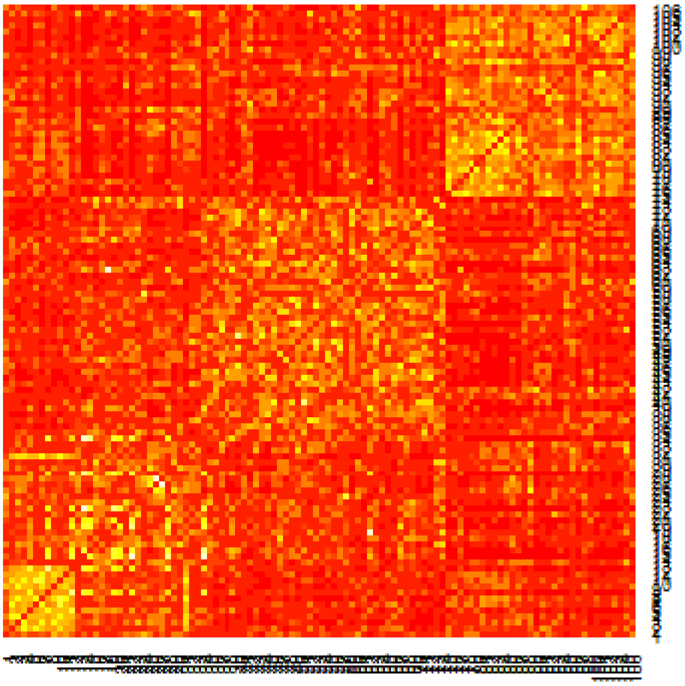

Figure 4.10Clusters between fused network of KRCCC and LSCC

Table 4.7Showing Concordance Between Fused network, network of KRCCC and LSCC

	Fused network of KRCCC and LSCC	Network of KRCCC	Network of LSCC
Fused network of KRCCC and LSCC	1	0.09954951	0.41207231
Network of KRCCC	0.09954951	1	0.0164095
Network of LSCC	0.41207231		1

Those results prove that there are similarities between different types of cancers. Now if we want to view the patient subtypes we will need to analyze these networks in an interactive 3d view to better understand the similarity properties among these cancers.

CHAPTER 5: CONCLUSION

5.1 **Conclusion:**

Similar patient subtypes in different domain of cancer can help in discovery of new drugs. More importantly it can help in discovery of targeted therapies that can target different cancer subtype using the same method. My results show what I thought as an attempt to find subtypes of patients from different cancer domain is possible. However, a further study and different types of analysis with three-dimensional visualization is required to understand how the grouped patient subtypes are related to each other. Analyzing survivability of each subtypes using Kaplan Meier survival curves might also yield important information. Bioinformatics is a huge field with vast possibilities. The results I found from my experiment is just flavor of what can be achieved in larger scale. With technology and people's dedication to contribute in field of bioinformatics and the availability of the data will open new windows of treating cancer soon. As in fact new methods to deal with cancer is being discovered every day like pathological apoptosis or the Entosis cellular cannibalism to limit the growth of cancer cell.

5.2 **Future Work**

The method I used is great method to aggregate different data types. However different clustering methods like Multiple kernel kmeans, Chameleon, or other partitive clustering, hierarchical clustering or hybrid clustering methods could be tried with it to compare the clusters formed. In future, I would like read further deep into mathematics and biology and use those knowledge to help in discover new properties of genetics. I can list my future work as follow:

1. I will collect more data from other different cancers and I will implement it in a very large data set to find relevant information.

2. I would like to study deep into RNA sequencing data and try other methods of analysis to search for similar patterns that can help in discovery of new drugs.
3. I would like to use different algorithms and techniques analyze both microarray and RNA sequencing data.

Our genetical information is continuously evolving I want to design and implement a system that can also evolve through time to keep track of the mutation in very large scale and with my interest in deep learning I sorely intend to implement it real life.

Bibliography

[1] Wang, B., Mezlini, A., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B. and Goldenberg, A. (2017). *Similarity network fusion for aggregating data types on a genomic scale*.

[2] López-Sáez JF, e. (2017). *Cell proliferation and cancer. - PubMed - NCBI*. [online] Ncbi.nlm.nih.gov. Available at: https://www.ncbi.nlm.nih.gov/pubmed/9810511 [Accessed 5 Aug. 2017].

[3] Cancerindex.org. (2017). *Apoptosis and Cancer | CancerIndex*. [online] Available at: http://www.cancerindex.org/Apoptosis [Accessed 5 Aug. 2017].

[4] Cancer.Net. (2017). *What is Metastasis?*. [online] Available at: http://www.cancer.net/navigating-cancer-care/cancer-basics/what-metastasis [Accessed 5 Aug. 2017].

[5] National Cancer Institute. (2017). *Angiogenesis Inhibitors*. [online] Available at: https://www.cancer.gov/about-cancer/treatment/types/immunotherapy/angiogenesis-inhibitors-fact-sheet [Accessed 5 Aug. 2017].

[6] World Health Organization. (2017). *Cancer*. [online] Available at: http://www.who.int/mediacentre/factsheets/fs297/en/ [Accessed 5 Aug. 2017].

[7] Dr Ananya Mandal, M. (2017). *Cancer Classification*. [online] News-Medical.net. Available at: https://www.news-medical.net/health/Cancer-Classification.aspx [Accessed 5 Aug. 2017].

[8] de Souto, M., Costa, I., de Araujo, D., Ludermir, T. and Schliep, A. (2017). *Clustering cancer gene expression data: a comparative study*.

[9] Hofree, M., Shen, J., Carter, H., Gross, A. and Ideker, T. (2017). *Network-based stratification of tumor mutations*.

[10] Şenbabaoğlu, Y., Michailidis, G. and Li, J. (2017). *Critical limitations of consensus clustering in class discovery*.