International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Predicting Cancer by Analyzing Gene Using Data Mining Techniques

Shahida M

M. Tech Student, Department Of Computer Science, Mount Zion College of Engineering Pathanamthitta, India

Abstract: ***cancer is an uncontrollable and abnormal growth of cells in the body. An understanding towards genetics and epigenetics is essential to cope up with the paradigm shift which is underway. Coming days, personalized medicines and gene therapy will a flowing together. This paper highlights the analysis of the gene expression data from cancer. It has become easier to collect extent experimental data in molecular biology. It is important to analyze that extent data and it leads to knowledge discovery that can be declared by experiments. Formerly, the complex genetic disease diagnosis was done based on the non-molecular characteristics such as tumor tissues, pathological characteristics. The microarray data is large and complex datasets. Machine learning Data mining techniques are applied to identifying cancer using gene expression data. In the proposed system the data is more efficient and accurate.***

Keywords: cancer, data mining, gene expression data, classification, clustering, gene therapy, epigenetic,PAM.

1. Introduction

CANCER is a condition of uncontrollable growth of abnormal cell due to damage in the DNA. Cancer cells keep growing and dividing. Cancer is a major cause of all the natural mortalities and morbidities throughout the world. With more than 10 million new cases every year, cancer has become one of the most causes a lot of harm or damage diseases worldwide. .Nearly 13 percent of deaths caused are due to cancer [1]. Some advancement has been reported for its clinical prevention and cure and there has been a noticeable decline in the lives' lost [4], but they are not quite adequate [5]. The growth in a body is observed, when the division and multiplication of cells takes place. Cancer is an abnormal and uncontrollable growth of cells in the body and that turn malignant. So we can say that all cancers are tumors, but all tumors are not cancer. Cancer can form in any organ or tissue, such as the colon, skin, breast, bones, lung, or nerve tissue etc.[2].

Some causes of cancers, including: Benzene and other chemicals, Drinking excess alcohol, Environmental toxins, such as certain poisonous mushrooms and a type of poison that can grow on peanut plants, Excessive sunlight exposure, Genetic problems, Obesity, Radiation, Viruses.

The cause and reason of many cancers remains unknown. The different types of cancer depends some symptoms. For example, Colon cancer causes constipation, blood in the stool, diarrhoea, and dysentery. Lung cancer can cause coughing, heavy breathing, chest pain, etc. May be some cancer have no symptoms. That corresponds to the suitable treatment required for ailing the cancer [2].

In the proposed system, predicting cancer by analyzing gene and converting the gene expression by using data minin g techniques is the concept of the project. Supervised multi attribute clustering algorithm is used to predict accurate diagnosis. Long shelf life is the main advantage.

In the previous study, scientist used to find the affected cells using microarrays to investigate the expression of 1000s of genes at a time. But sometimes may be the errors are occur in the microarray datasets and t is expensive to create, very short shelf life.

2. Epigenetics

In Epigenetics, speaking literally "epi" stands for "on top of" and epigenetics is on top of genetics. It is the study of changes in organisms caused by modification of gene expression rather than alteration of gene code itself. Stable alteration in gene expression pattern. Dynamic process that plays a key role in normal cell growth and differentiation to date the best understood epigenetic mechanism are DNA methylation and histone modifications.

2.1 DNA methylation

In the mammalian genome, the most commonly occurring epigenetics events taking place. This change though heritable, is reversible, making it a therapeutic target. Methylation pattern is determined during embryogenesis and passed over to differentiating tissues and cells. DNA structure is maintained from generation to generation. This structure is modified by base methylation in nearly all cells and organisms. Das and Singal [3] portray DNA methylation as an epigenetic event that highly correlates to the regulation of gene expression. As one facet, DNA methylation exhibits direct interception with the binding sites of particular transcription factors to their promoters. Also, they are involved with the direct binding of specific transcriptors to the methylated DNA. From the cancer perspective, malignant cells as opposed to their normal counterparts show exaggerated disturbances in their DNA [41]. Hypo-methylation is another characteristic of the solid tumor types as cervical and prostate cancers. Role of dna methylation are: plays a role in long term silencing of gene, silencing of repetitive elements, X-chromosome inactivation, in the establishment and maintenance of imprinted genes, suppresses the expression of viral genes and other deleterious elements that have been in corporate in to the genome of the host overtime.

3. Gene Expression

In gene expression, it contains two steps. They are transcription and translation. In transcription, the synthesis of mRNA uses the gene on the template like DNA molecule. This happens in the nucleus of eukaryotes. In translation, the synthesis of polypeptide chain using the genetic code on the mRNA molecule as its guide. RNA (ribonucleic acid) is found all over the cell (nucleus, mitochontria, chloroplasts, ribosome's and the soluble part of the cytoplasm). Some types of RNA are messenger RNA (<5%), Ribosomal RNA (up to 80%), Transfer RNA(15%). In eukaryotes small nuclear ribonucleoproteins. The structural characteristics of RNA molecules are single polynucleotide strand which may be looped or coiled (not a double helix, sugar ribose (not deoxyribose). Adenine, guanine, cytosine, uracil are the base used.

4. Gene Therapy

Gene therapy is the introduction of genes into existing cells to prevent or cure a wide range of diseases. It is a technique for correcting defective genes responsible for disease development. The first approved gene therapy experiment occurred on September 14,1990 in US. Somantic cell gene therapy and germ line gene therapy are the types of gene therapy. In somantic cell gene therapy, therapeutic genes transferred in to the somatic cells. Will not be inherited later generation. In germ line gene therapy, therapeutic genes transferred into the germ cells. It's heritable and passed on to later generations. Gene therapy has the potential to eliminate and prevent hereditary diseases such as cystic fibrosis, ADA-SCID etc. it is possible cure for heart disease, AIDS and cancer. It gives someone born with a genetic disease a chance to life. It can be used to eradicate disease from the future generations. Theoretically, gene therapy is the permanent solution for genetic diseases. At its current stage, it is not accessible to most people due to its huge cost. A breakthrough may come anytime and a day may come when almost every disease will have a gene therapy. Gene therapy have the potential to revolutionize the practice of medicines.

5. Cancerous Gene Identification

Raza and Mishra [7] (2012) attempt to stratify genes within samples (tissues) by recursively filtering genes on the basis of their expression levels and active indulgence in the disease state. The expression level of genes is proportional to various conditions in an organism. It is incumbent to mark reference genes that can be levied as standard for selecting further candidate genes on the basis of a priori criterion. These can be well suited to be potential drug targets and as sites for studying mutations. For such study, gene expression matrix is a good reference source for each gene's expression variance. The algorithm curtails following steps:

1. Ratio and logarithmic conversion of microarray data. (The gene regulation is reflected in the fluorescence intensities that illuminate on superimposition of the multivariate genes. Author has limited the regulation levels by defining up-regulation -> [1, ∞] and down-regulation -> [0, 1]. This categorization brings about rigidity in selecting the acceptable gene expression levels. Through intensity ratio plot, interpretations can be efficiently visualized.)
2. Elimination of gene that fails to provide data in majority of experiments. (Due to several technical issues of improper probing, particular cell orientation of genes, faulty scanner that fails to measure correct expression levels, and due to erroneous manufacturing of microarray chip, gene expression levels can be severely affected. On account of a threshold value of 40 percent, rows holding genes that are not expressed up to the level are abstained from being a part of the experiment in the view of not having significance relevance.)
2) Analysis of significance of data. (Use of t-statistics promulgates the compliance of normal distribution in the data that is responsible for certain pattern generation. This [pattern] can be analyzed and may be interpreted as a result.
3) Replicate handling. (There should be a single entry for each gene.) Elimination of gene having less than two-fold change in expression level. (Irrelevant genes- that do neither show acceptable up-regulation [positive value] nor down-regulation [negative value], are curbed. For convenience of sifting data, genes means of all rows is calculated and genes with –1<mean<1 is selected further.)
4) Conversion of data sets using Log sigmoid function. (The Log sigmoid transformation takes range of input values in between [–∞, ∞] and converges them to the range [0, 1]. The function is given by log sigmoid (x) =1/ (1+$e^{(-x)}$)
5) Elimination of genes that have high variation across the collection of sample. (The genes with sporadically occurring expression levels are omitted. The process accounts for elimination of genes having more than 36 percent variation due to inconsistency.)[6]

6. Scope of the Project

Scientists used to be able to perform genetic analyses of a few genes at once. DNA microarray gives to analyze thousands of genes at one time and it is fast. So we can say that this is a good way to analyze cancer disease. But sometimes may be the errors are occurring in the microarray output. The scope of the project is to eliminate the errors using data mining techniques to predict and analyze accurate data.

7. Module Description

There are mainly four modules in the proposed system. They are microarray, clustering, classification and preprocessing.

7.1 Microarray

Microarray analysis techniques are used in the interpreting the data generated from experimental result on DNA, RNA, Proteins microarray which allow researchers to investigate the expression state of a large number of genes in many cases, an organisms entire genome in a single experiments. First illustrated in antibody. A microarray analysis compares a person's DNA and the control DNA. Microarray can only determine how many copies of each pieces of the chromosomes are present. Microarray cannot tell us how the

chromosomes are arranged. Scientists used to be able to perform genetic analyses of a few genes at once. Usually made commercially. Made of glass, silicon, or nylon. Each plate contains 1000s of spots, and each spot contains a probe for a different gene. The microarray data are images that are transformed into gene expression matrices- tables where genes represented in rows; various samples are represented in columns. So the tissues and experimental result, numbers in each cell characterize the expression level of the particular gene in the particular sample. These matrices are analyzed further. The purpose of microarray is:

1) To measure changes in gene expression levels
2) To observe genomic gains and losses.
3) To observe mutations in DNA.

The latter is called a Gene Chip. The input of the project is the datasets of the microarray.

7.2 Preprocessing

Data in the real world is dirty: Incomplete, noisy, and Inconsistent. A gene expression matrix is obtained, contains gene data. Data pre-processing is indispensable before any cluster analysis can be performed.Methods including fold-change and Significance analysis of microarrays (SAM). Fold-change is a comparatively simpler method; SAM operates on certain statistical assumptions. Fold-change technique works on selecting/eliminating genes with a predetermined threshold level (usually a factor of 2). It compares this level with the mean level of the gene expression and thence chooses/rejects genes on the basis of the calculation. If the gene are not depend the minimum and maximum value we can eliminate or delete such type of genes.

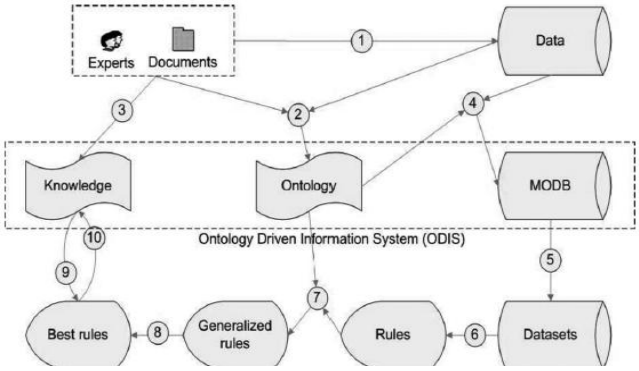

Figure 1: KEOPS Methodology

7.3 Classification

Classification trees recursively partition the space of expressions into subsets that are highly predictive of the phenotype of interest. Create a tree by using distinct values. They are robust, easy-to-use. Prescreening of the genes is not required. Using intuitive graphical representations, the resulting predictive models can be displayed. Here classification tree and predictive analysis of microarray(PAM) is used in classification. In classification, PAM is a straight forward approach and is the nearest centroid classifier. For each class this computes a centroid is given b y the avg. expression levels of the samples. After that assigns new/fresh samples to the class whose centroid is nearest. This is same like k-mean clustering algorithm. Except clusters are now replaced by known classes. PAM software used to implement.

7.4 Clustering

Clustering is a technique for finding similarity groups in data, called clusters. Clustering is often called an unsupervised learning. K-means is a partitional clustering algorithm The *k*-means algorithm partitions the given data into *k* clusters. *k* is specified by the user. Each cluster has a cluster center, called centroid. By using k means algorithm measure the variance for identifies how and where the hierarchical clustering stops or end. In heirarchical clustering, given the input set S, the goal is to produce a hierarchy (dendrogram) in which nodes represent subsets of S. One of the major applications of clustering in bioinformatics is on microarray data to cluster similar genes. So clustering microarray data in a way helps us make hypotheses about:

- potential functions of genes and protein-protein interactions

Features of the tree obtained:

- The root is the whole input set M.
- The leaves are the individual elements of M.
- The union of their children is the internal nodes.

Partitions of the input data into several clusters or groups are represent each level of the tree.

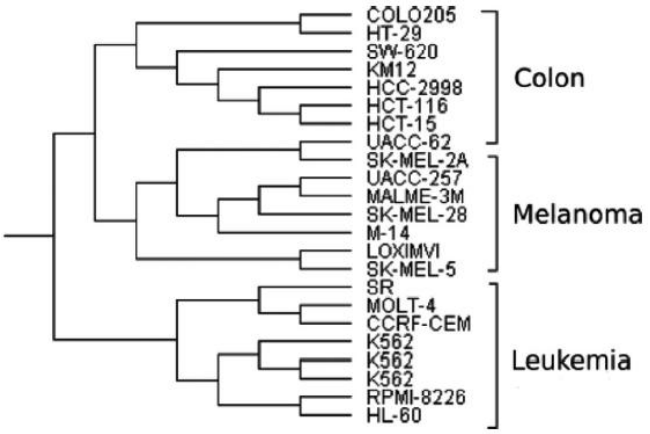

Figure 2: Dendrogram

8. Background study

To find the affected cells using Microarrays to Investigate the "Expression" of Thousands of Genes at a Time. Error occurs in the result dataset during DNA microarray. Expensive to create. Large and Complex data sets. Long time for analysis. Do not have very long shelf life. Size and complexity are the main problem.

9. Result Analysis

This paper implemented in java programming language. The main advantage of this paper is identified; analyze gene and predicting cancer by using data mining techniques. Following screenshots represents the output of the work carried out on the project predicting cancer by analyzing cancer by using data mining techniques.

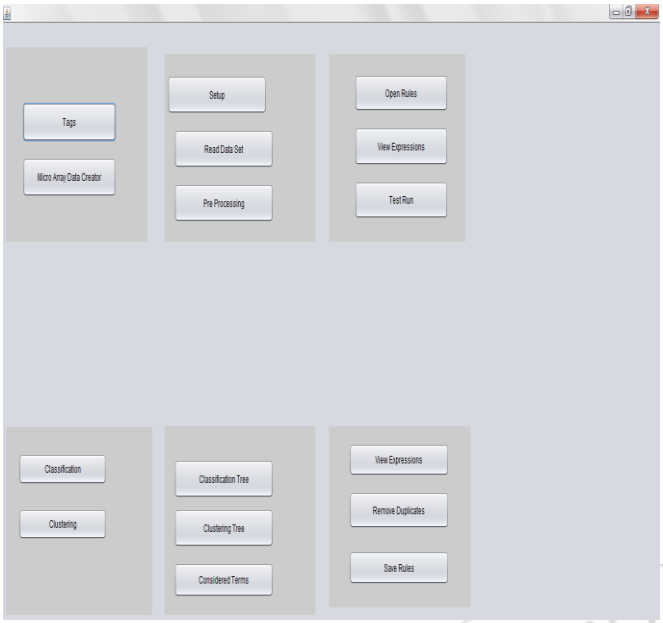

Figure 3: Main Menu

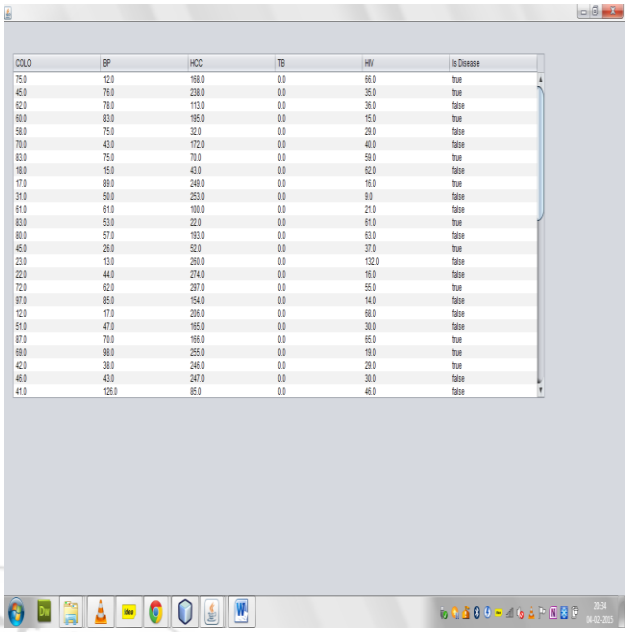

Figure 6: Read datasets

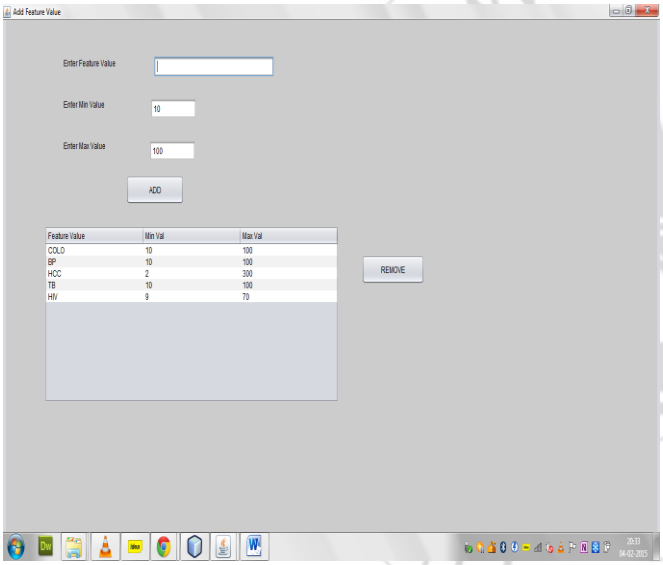

Figure 4: Add factors

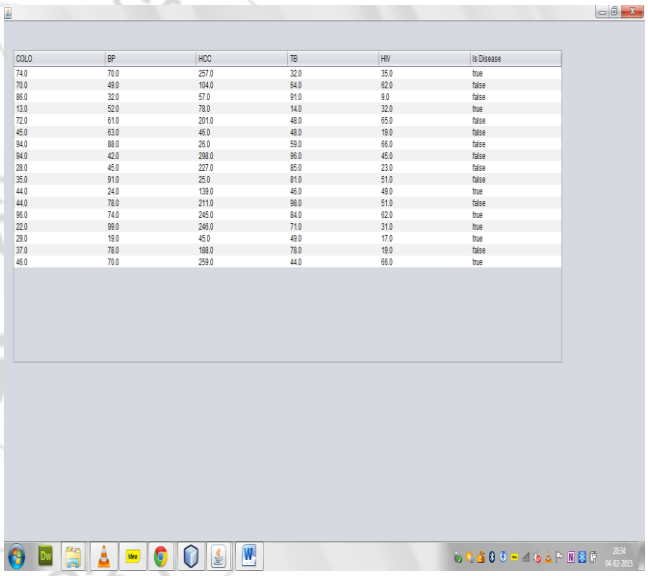

Figure 7: Preprocessing

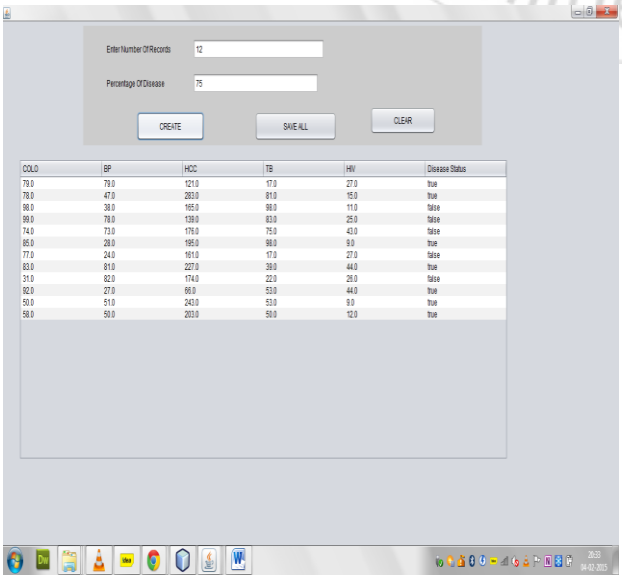

Figure 5: MicroArray creator

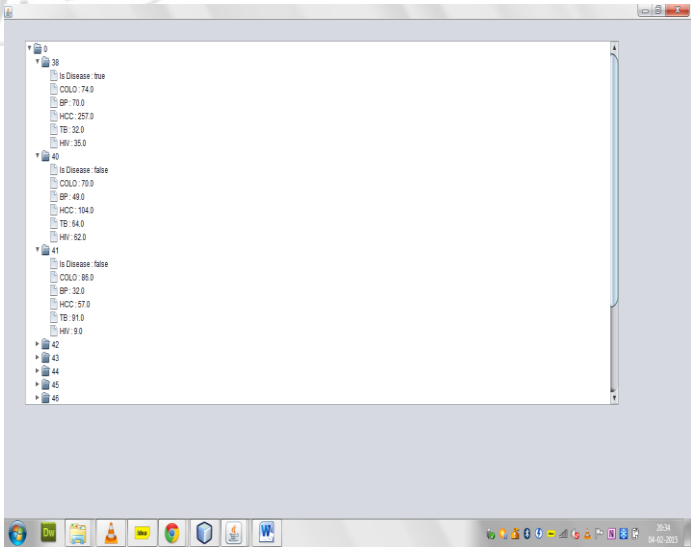

Figure 8: classification

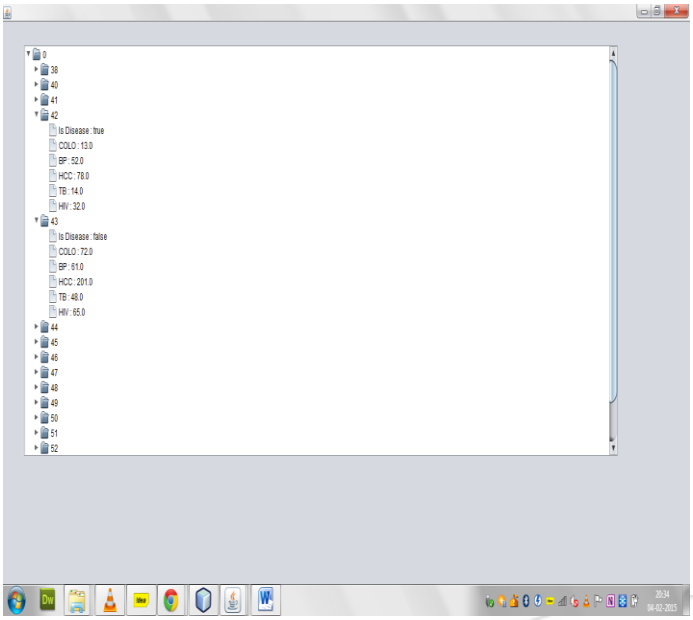

Figure 9: clustering

This proposed method show that identifying and analyzing the gene to predict cancer by using data mining techniques. In this work it has more shelf life and produce accurate data as compared to previous works.

10. Conclusion

In our proposed system, it is observed that a good in quality and accuracy in expression or details classification of tumors is important for successful identification of the nature of an illness and treatment of cancer. Early time, the diagnosis of complex genetic diseases have in accordance with has been based on the non-molecular characteristics like pathological characteristics and clinical phase. By allowing the monitoring of expression levels in cells for thousands of genes done at the same time. Microarray experiments lead to a more full understanding of the molecular variations of tumor cells. A microarray datasets contains many groups of co-expressed genes. Hence to a finer and many informative classification. Several machine learning and data mining techniques are used to analyze or identifying the cancer gene the human body. Comparing the activity of genes in a healthy person and cancerous tissue it may give many clues about gene that are involved in cancer. Moving towards a bygone of personalized medication, gene therapy and Next generation sequencing are producing there mark there. There are more different companies are offering there service and it can be online ordered. Some missing are there that is only few genes are examined or to test in the genetic tests and not the entire genome. Some reasons are occurs that are time and monetary constraints. In the proposed system produce accurate data to find or identify and analyze cancer disease.

References

[1] Data and Statistics. World Health Organization, Geneva, Switzerland, 2006.
[2] PubMedHealth- U.S. Nat. Library Med., (2012). [Online].Available:http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0002267/
[3] P. M. Das and R. Singal, "DNA Methylation and Cancer," J. Clin. Oncol., vol. 22, no. 22, pp. 4632–4642, Nov. 2004.
[4] R. Siegel, J. Ma, Z. Zou, and A. Jemal, "Cancer statistics," CA: A Cancer J. Clinicians, vol. 64, pp. 9–29, 2014.
[5] "International Agency for Research on Cancer (IARC)," WHO, B. W. Stewart and C. P. Wild eds., World Cancer Report, 2014.
[6] Mining Gene Expression Data FocusingvCancer Therapeutics: A Digest ,Shaurya Jauhari and S.A.M. Rizvi 2014
[7] K. Raza and A. Mishra, "A novel anticlustering filtering algorithm for the prediction of genes as a drug target," Amer. J. Biomed. Eng., vol. 2, no. 5, pp. 206–211, 2012.

Author Profile

Shahida M received the Btech degrees in Information Technology engineering from Mount Zion College of Engineering Kadammanitta in 2012. Currently doing Mtech degree in Computer Science Engineering under Mahatma Gandhi University