CANCER CLASSIFICATION OF GENE EXPRESSION DATA USING MACHINE LEARNING MODELS

V. Anu

PG Scholar, Department of Computer Science Engineering, Pandian Saraswathi Yadav Engineering College, Sivagangai, India

Abstract

The World Cancer Report described cancer as a global problem because it affects the whole greater population. There will be a projected increase to 20 million new cases by 2025 [2]. There are several known published literatures on cancer classification techniques with varying models and implementations. This paper presents the existing technology of microarray gene expression and classifies the cancer genes using machine learning algorithms. A logical design was presented using supervised classification and gene selection model. This model can improve the process and method of identifying and classifying cancer disease.

Keywords: *Cancer genomics, microarray, gene expression, cancer classification, supervised classification, machine learning.*

Introduction

Cancer disease has been reported as one of the deadliest genetic maladies of the human genome. It has been the research interest until today by doctors, pathologist, biologist and others life science and health professionals. The World Health Organization (WHO) reported cancer disease having 14million new cases in 2012. This disease is a major cause or morbidity and mortality that accounts of death globally resulting 8.8million deaths in 2015 [1]. The World Cancer Report described cancer as a global problem and projected an increase to 20 million new cases by 2025[2].. There are approaches in technology that reveals the cellular and molecular level of cancer. In a cancer disease sample such a cell biopsy to be processed, thousands of genes at a time can be subjected for analysis in a single chip called microarray. Microarrays are microscopic slides that contain ordered series of samples of DNA (Deoxyribonucleic acids), RNA (Ribonucleic acids), protein, or tissue and others [3]. A single chip microarray can measure the gene expressions of 30,000 gene sample that represents most of the human genome [4]. The challenge of cancer classification using the microarray is the application of model based selection and prediction algorithm that will classify the cancer genes using gene expression data. The computation time, classification accuracy, and its biological relevance in the cancer classification was still in question. The main goal of this study is to explore and analyze the published papers in cancer classification. The scope of this paper is to present the cancer classification using machine learning models.

Related Works

Many of the published research and articles about cancer disease and associating the word “cancer classification and gene expression data” when searched on the Internet. Using Google search engine, it returned 30million search editors (as of August 8, 2018) using the searched keyword. In Google scholar, it’s 1.8 million where most of the cited it ensure cancer genome, proteomics, microarray, machine learning algorithms and others. Similarly, in biomedical and

genomic research, the human genome has been analyzed, sequenced and codified to discover the types of diseases such as cancer and other dreadful diseases. The International Cancer Genome Consortium analyzed more than 25,000 cancer genomes as of 2013 [5]. There was a rapid expansion of the cancer genome data sets also accelerated the genetic analytical tools for genome association studies and analysis through microarray. The result of these analyses was maintained through different online repositories and reported in scientific and research journals such as the Pub Med of the National Center for Biotechnology Information (NCBI) and other life sciences, bioinformatics, and genome science journals [6]. MEDLINE is the journal citation database that has 25 million references. Pub Med has over 28 million citations of biological articles and ever increasing every year. While Pub Med Central is the full-text journal article shave over 3 million articles from Pub Med [7].

Cancer Genome Studies

Cancer in the medical term is abnormal state of a normal cell or a group of cells that mutates and destroys so threat issues in the human body. There are greater than100 different types of cancer diseases [8]. The genome-wide association studies (GWAS) helped in identifying the variants of genetic disease [9]. The cancer research accelerated the reporting of GWAS resulted in the investigation of genetic analysis. In 2007 GWAS publication, there are about 40 unmistakable here ditaryloci have been convincingly distinguished for in excess of two dozen distinct cancers [10].

Microarray and Gene Expression Data

The microarrays contain samples of DNA, RNA, proteins [3]. The sample placed into the slide such as DNA microarray; RNAmicroarrayandotherswillbethetypeofmicroarray.DNA is held in place by chemically reactive aldehydes or primary amine so either synthesizes dbyphotolitho graphic process. The cancer gene expression is made possible from the Internet cancer genomic data[12]. Most of the data available are breast and lung cancer data sets and others have less than 100 sample sizes. Micro array profiling innovation, which has been most generally used to study gene expression in cancer. Cancer Classification methods, evaluation, and accuracy

The weighted voting gene selection works well for classifying binary data [13, 14]. This method works well with some data such as leukemia. The disadvantage of this method was it is not effective in more than 2 classes of data set. In the Fisher's linear discriminate analysis (FLDA) [15] applied to cancer classification tries to find the linear combination of class from sum of its squares.

The similarity based classifiers k-Nearest Neighbor (KNN) [16,17]and Cluster-based classifier (CAST) [18,19] with tuples are not affected by the noise and bias in data. CAST is a cluster based on separable groups containing normal and tumor samples. KNN use less computing time than CAST because of the similarity score evaluation performed on every test and training. These methods are not scalable and not practical for cancer classification because of it use too much computation time. The max-margin classifiers described by [20, 21],Support Vector Machine

(SVM) [22,23,24] used in gene expression data [25,26,27] and used in different cancer classification problems [18,28,29]. SVM has an advantage of selecting a few support vectors of the learning algorithm against the large training set [18,30,31]. However, SVM is limited only for binary class problems. Some extensions of the multiclass SVM methods were presented by [32, 33, and 34]. However, the problem of performance and effectiveness still remain un answered .Boosting improves the classification performance through number of folds of class training. This approach makes improved classification accuracy compared with other algorithms. Boosting was applied to different cancer classification problems by [17, 18]. But the repeated classification of weighted training consumes much time effort. Another approach is the use artificial intelligence techniques such as Bayesian Network (BN) used for gene classification [35, 36] and Neural Network (NN)[37]cancer disease prediction. These can be applied to multiclass classifier.

The disadvantage of the process is a black box and not capable to reveal any biological information in the data. Decision Trees (DT) [38]can be interpret edits meaning and does not require parameter. Trees can be generated right away as the data size increases. DT algorithms are good classifier in terms of scalability. DT implementations were made and improved by proposed [39, 40, 41].

In terms of the classifier accuracy and performance the following experiments made by [18]. While Boosting is better to outperform NN for leukemia, ovarian, colon data set.Similarly, Naïve Bayes (NB) out performs Gene Selection (GS) approach for leukemia and ovarian data sets. The opposite for colon where GS did great compare to NB [35]. Table 1 presented the summary of the cancer classifiers survey result [42].This shows that there is no cancer classifier that is superior to all of the models. This can be a research topic to explore on cancer classifier's accuracy and biological meaning that points to a new classification algorithm or to enhance the capability of the existing algorithm to fit the bio-relevant answer to cancer classification. The limited number of cancer database and data sets varies from each type of cancer genes from each source.

Table 1 Summary of the Cancer Classification [42]

Classification method	Multi class	Strategy Evaluation	Biological meaning	Scalability
Support Vector Machine	No	Max-Margin	No	Good
Boosting	Yes	Max-Margin	Yes	Class dependent
Decision Tree	Yes	Entropy function	Yes	Good
K-Nearest Neighbor	Yes	Similarity	No	Not scalable
Cluster–based Similarity Tuple	Yes	Similarity	No	Not scalable
Gene Selection	No	Weighted voting	Yes	Fair

Fisher Linear Discriminant Analysis	Yes	Discriminant Analysis	No	Fair
Neural Network	Yes	Perceptron	No	Fair
Naïve Bayes	Yes	Distribution modeling	No	Fair

Gene Selection

The feature selection helps to eliminate problems in data set noise and over fitting of the classifier. In addition, these will reveal the bio-relevant information to make use of DT to see the actual view of gene movement and value. The gene selection reduces the large attribute space that helps the classifier to improve the accuracy [13, 14, 17, 19, 43, 44, 45]. The gene feature ranking approach measures the correlation of class labels and attributes values. Using the GS method [13] with the correlation is simple to implement but has a disadvantage for mistaken selection of cancer gene values for normal and tumor types. Comparing the NB and GS selection method [17], NB classifier accuracy is better and has more genes variety than GS. Another approach is the gene subset ranking (GSR). In this method, genes are clustered to obtain the best classifier. Lastly the recursive feature elimination (REF) makes the elimination process retain the best classification power. This is also used in SVM classification as a cost function on the sub set ranking. The REF and GSR work great in cancer classification compared to individual gene ranking (IGR) method.

Methodology

This paper mainly explored and analyzed published papers about the topic. The papers were extracted from Scopus and Pub Med and analyzed using Cited Reference Explorer (CR Explorer) in section A. To illustrate the tasks for the classifier model, the following activities were described in the next sections. Describing the dataset in Section B. The application of machine learning cancer classifier method in Section C.

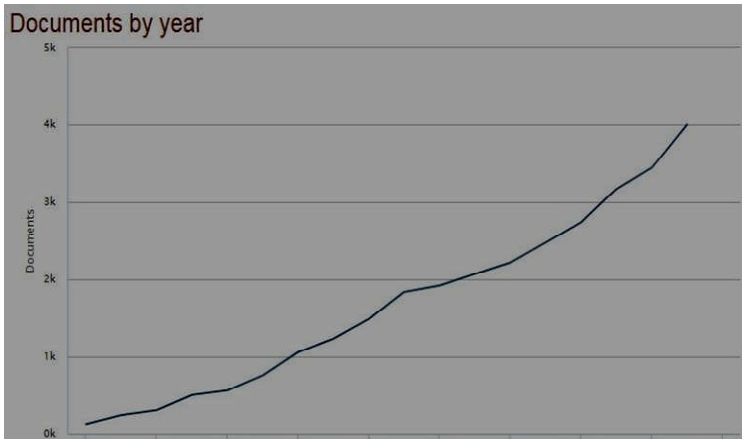

Fig. 1 Documents per year by source

+

Is the frequency of references cited in there searches in terms of the publication years. The RPYS analysis of the topic. There were 55,040 cited references in the 1967-2017 period and 51 different citing publications years with the total of 1448 publications cited The cited publication years considered were from 2000-2017 period with 18 publication year. The researchers on cancer classification using gene expression started its traction on 1995 with its influential work on 1999 of [13, 17, 19, 43, 44, 45].The peak of their search on 2000 with the use of microarray more datasets and the evaluation using classification methods. The work of [45] marked its constant cited reference of the citing years thereafter. The evaluations and classification method were discussed.

Dataset

The data set used is a micro array from the Golub experiment [13]. The dataset is available online from the repository of Stanford Hastie CASI files [46]. The process of obtaining the best model for predicting disease classes from a given raw data set collected by scanning genetic microarrays from 72 patients, each suffering from one of the strains of cancer (leukemia).For each of the patients, the scanner tabulated the values of 7129 genes, each of which was assigned a numerical value. The dataset is 65% ALL and 35% for AML. It contains the gene description and gene accession number. The dataset has training set of 7129 rows (instances) with72columns (features) with 38 samples. The test set contains 6627 unique values of genes from the 7129 values with 34 samples. The data was normalized for this experiment to select the genes that was correlated to the outcome of the combined features. This reduces the genes and to increase the classification accuracy. Using line a method T-values, this reduced the number of genes for the training model. Using the formula below:

$$\text{T-test for mean difference} = \frac{(Av1 - Av2)}{(\sigma 1^{\angle}/N1 + \sigma 2^{\angle}/N2)} \quad (1)$$

Av1 and Av2 average of 1 class from the given gene expression classes. The sigma is the standard deviation of each classes. Then, N1andN2 are samples whose class has T-values. and does not have the T-values. Then running the experiment using selected machine learning tool for cancer classification method.

Machine Learning Tool for Cancer Classification Model

Using the Google Colaboratory (CO) [47] notebook with python programming to run the experiment for the AML and ALL dataset. Using the PyML package, that has classification and regression methods. SVM is a classifier PyML. The classifier used in the experiment was Support Vector Machine (SVM) and Boosting (Extreme Gradient Boosting). SVM is a supervised learning method that analyzed data for classification.

Results and Discussions

The cancer classification logical model used on this paper starts by loading the gene expression data set which is known and normalized from a microarray gene expression platform. The training set has been defined; a technique for the classification is used. Subsequent validation in feature selection of gene is informative in the classification, and classifiers are built based on these dataset. Using the correlation analysis of selecting the features in the dimensionality reduction (gene selection and extraction) this will have the similarity of genes as the featured vector. Then the classifier can predict the new and non-labeled sample genes.

Analyzing Leukemia Dataset

The dataset used in this paper was described in section B. The normalization process used computing for the T-test mean difference values to reduce the number of genes in the training model.

Data Cleaning

Using the Google Colab [47] notebook with python programming to run the experiment for the AML and ALL dataset. Using the PyML package, that has classification and regression methods. SVM is a classifier PyML. View the data set for data cleaning and preparation. Refer to Figure7forthesampleoutputofthedataframeforthecolumn values of test and train data sets. This process will remove the column labels that are not significant in the analysis.

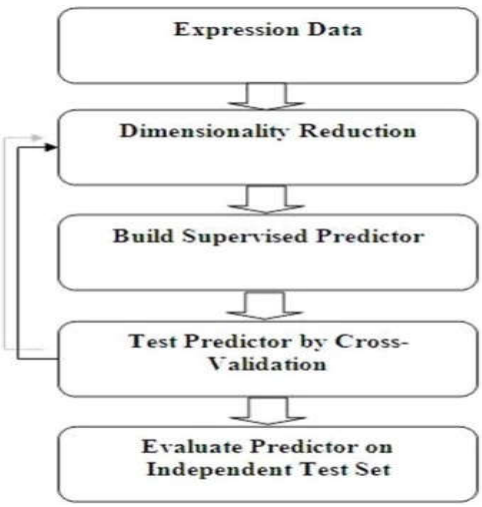

Figure 2 Logical Model for Supervised Learning [53]

Data Standardization and Dimensionality Reduction

This is the preprocessing of the dataset required to standardization form a chine learning estimators. In this process, sci kit-learn [48] was used to make the individual features with zero mean and unit variance. This also computes for the mean and standard deviation. Similarly in

scaling feature values between minimum and maximum. The dimensionality reduction using principal component analysis (PCA) [49] use SVD (sklearn. decomposition.PCA).

Support Vector Machine (SVM*)*

Using the sklearn.svm [50,51] and sklearn.svc method the effective technique for the leukemia dataset. The case of the AML-ALL number of dimensions is greater than that of the samples in the data set. The support vector classification in this experiment described the following outputs:

- Taking the X_pca =pca.fit_transform(X_all);
- model.fit(X_pca[:38,:],labels_train.values.ravel())
- prediction output: Accuracy58.82%
- confusion matrix: 20, 0 | 14,0

The heat maps in Figure 4 and Figure 5 presents the correlation of the gene expression data column values using the Pearson, Kendall, and Spearman correlation and rank coefficients.

Gradient Boosting for Classification

Using the sklearn. Ensemble Gradient Boosting Classifier builds an additive model and optimization of the loss function. Taking the next step in model building of the leukemia AM. The gradient boosting classification in this experiment described the following outputs:

- Xgb = XGB (max_depth=10, loss='exponential', n_estimators=100, learning rate = 0.02, random_state=42)
- xgb.fit(X_pca[:38,:],labels_train.values.ravel())
- pred = xgb. predict(X_pca[38:,:])
- prediction output: Accuracy64.71%
- confusion matrix: 20, 0 | 12,2

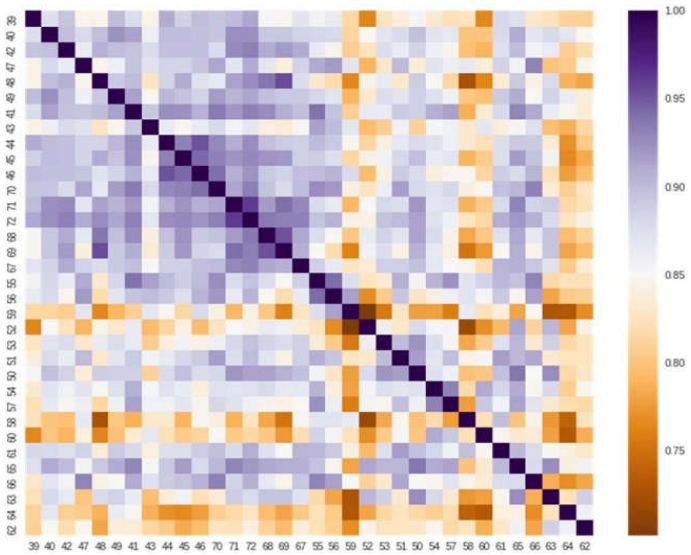

Figure 3 Heat Map of Test Dataset ALL-AML

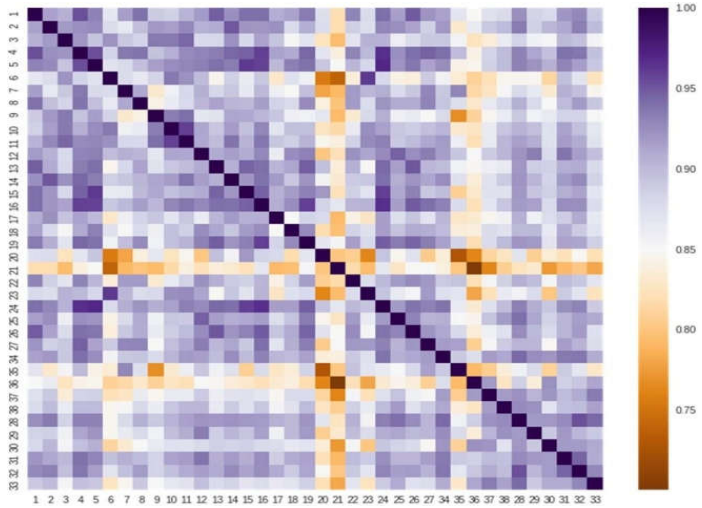

Figure 4 Heat Map of the Training Dataset ALL-AML

Conclusions

This paper presented the gene expression data analysis and classifies the results based on the cancer dataset. The literature analysis presented cancer classification using gene expression was still a topic on interest by most researchers in this field. The survey analysis of the cancer classification model evaluation presented the advantage and disadvantages of each. Gene selection is an important phase in the pre processing and cancer classification.

This paper demonstrated the cancer classification model and applied the SVM and Boosting that provide insights of their application in the gene expression data. The performance and accuracy reported having 58% and 64% can indicate that the experiment can be improved and comparable too there results of published literatures.

The next step is to investigate the cancer classification techniques specific to cancer genome or type of cancer disease using the other machine learning and deep learning techniques. Python packages can be programmed based on the model presented will be the next step. The application of artificial intelligence can be considered as the next step as our research in cancer classification in histology, digital oncology and pathology.

References

1. World Health Organization (2018). Cancer Fact Sheet, Feb.2018 Media Center. Accessed on March 8, 2018 from http://www.who.int/mediacentre/factsheets/fs297/en/
2. Stewart, B. and Wild, C. (2014). World Cancer Report 2014. International Agency for Research on Cancer (IARC), World Health Organization (WHO). WHO Press.
3. Wong, G (2005). Introduction. In Minna Laine. DNA Microarray data analysis (15-24). Helsinki: CSC- Scientific computing Inc.
4. Ramaswamy S., Tamayo, P. & Rifkin, R. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. PNAS Vol. 98(26), 15149-15154.

5. Omics International (n.d.). Cancer Genomics. Journal of Clinical and Medical Genomics. Accessed on June 15, 2018 from https://www.omicsonline.org/scholarly/cancer-genomics-journals- articles-ppts-list.php.
6. National Center for Biotechnology Information US National Library of Medicine (2018). Pub Med Help. Accessed on Une 15, 2018 from https:// www.ncbi.nlm.nih.gov/books/ NBK3827/# pub med help. PubMed _ Quick_Start
7. National Center for Biotechnology Information (n.d.). All Resources, Databases. Accessed on June 15, 2018 from https://www.nlm.nih.gov/bsd/difference.html
8. National Cancer Institute (2018).What is cancer? Accessed on March 20, 2018 from https://www.cancer.gov/about- cancer/understanding/what-is-cancer.
9. Chung CC, Magalhaes WC, Gonzalez BJ, Chanock SJ (2010). Genome wide association studies in cancer – current and future directions. Carcinogenesis, 31:111–120. http://dx.doi.org/10. 1093/carcin/ bgp273PMID
10. Hindorff LA, Gillanders EM, Manolio TA (2011). Genetic architecture of cancer and other complex diseases: lessons learned and future directions. Carcinogenesis, 32:945–954. http://dx.doi.org/10.1093/ carcin /bgr056 PMID:21459759
11. Scopus (2018). Scopus. Elsevier .Accessed on Mar8, 2018 from https://www.elsevier.com/ solutions/ scopus.
12. CR Explorer (2018) Accessed on Mar 20, 2018 from http://andreas-thor.github.io/cre/#
13. Golub,R., Slonim,D., Tamayo,P.(1999).Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, pages531–537.
14. Slonim, D., Tamayo, P., Mesirov, J., Golub, T., and Lander, E. (2000). Class prediction and discovery using gene expression data. In Proc. 4th Int. Conf. on Computational Molecular Biology(RECOMB), pages 263–272.
15. Fisher, R. (1936). The use of multiple measurements in taxonomic problems. Annual of Eugenics, 7:179–188.
16. Fix, E., and Hodges, J. (1951). Discriminatory analysis, nonparametric discrimination: Consistency properties. Technical report, USAF School of Aviation Medicine.
17. Dudoit, S. Fridlyand J., & Speed, T. (2000). Comparison and discrimination methods for the classification of tumors using gene expression data. Technical report no.56. Berkeley. Department of Statistics., Univ. California,43
18. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini (2000). Tissue classification with gene expression profiles. In Proc. of the Fourth Annual Int. Conf. on Computational Molecular Biology.
19. Alon, U., Barkai,N., Gish,K., Levine,AJ., Mack,D., Notterman, DA., Ybarra, S. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Cell Biology, PNAS, Vol. 96,6745-6750.
20. Freund, Y., and Schapire, R. (1998).. Large margin classification using the perceptron algorithm. In Proc. of the 11th Annual Conf. on Comp. LearningTheory.

21. Smola, A., Bartlett, P., and Scholkopf, B. (2000). Advances in Large-Margin Classifiers. MITPress.
22. Boser, B., Guyon, I., and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In Proc. of 5th Annual ACM Workshop on Computational Learning Theory, pages 144–152. ACM
23. Burges,C.(1998).Atutorialonsupportvectormachinesforpattern recognition. Data Mining and Knowledge Discovery, 2:121–167.
24. Vapnik,V.(1998).StatisticalLearningTheory. Wiley,NewYork, NY.
25. Brown, M., et al. (2000). Knowledge based analysis of micorarray gene expression data by using support vector machines. In Proc. of the National Academy of Sciences, volume 97, pages262–267.
26. Fujarewicz, K., Kimmel, M., Rzeszowska-Wolny, J., et al. (2001). Improved classification of gene expression data using support vector machines. Journal of Medical Informatics and Technologies,v.6.
27. Mukherjee, S., Tamayo, P., Slonim, D., Verri, A., Golub, T., Mesirov, J., and Poggio, T. (1999). Support vector machine classification of microarraydata.
28. Furey,T.,Cristianini,N.,Duffy,N.,Bednarski,D.,Schummer, M., and Haussler, D. (2001). Support vector machine classification and validation of cancer tissue samples using microarray expression data.Bioinformatics.
29. Ramaswamy, S., Tamayo, P., Rifkin, R., et al. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. PNAS, 98(26):15149–15154.
30. Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Ares Jr., M., and Haussler, D. (1999). Support vector machine classication of microarray gene expression data. Technical report, Univ. of California at SantaCruz.
31. Shawe-Taylor, J., and Cristianini, N. (1999). Furthur results on the margin distribution. In Proc. 12thAnnual Conf. on Computational Learning Theory.
32. Crammer, K., and Singer, Y. (2000). On the learnability and design of output codes for multiclass problems. Computational Learning Theory, pages35–46.
33. Lee,Y.,Lin,Y.,and Wahba,G.(2001).Multi category support vector machines. Technical report, University of Wisconson- Madison.
34. Hsu, C., and Lin, C. (2001). A comparison on methods for multi- class support vector machines. Technical report, National Taiwan University, Taipei, Taiwan.
35. Keller, A., Schummer, M., Hood, L. and Ruzzo, W. (2000).. Bayesian classification of DNA array expression data. Technical report, University of Washington.
36. Friedman, N., Nachman, M., and Pe'er, D. (2000). Using baysian networks to analyze expression data. In Proc. of the 4th Ann. Int. Conf. on Comp. Molecule Biology.
37. Khan,J.,Wei,J.,Ringner,M.,Saal,L.,et.al.(2001).Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine.