IJRSET March 2016 Volume 3, Issue 3 Pages: 31-37
ISSN 2394-739X
International Journal for Research in Science Engineering and Technology

GENETIC MINING OF LEUKEMIA USING MICROARRAY CLASSIFIED DATA

[1] B. RAJESWARI, [2] Dr. ARUCHAMY RAJINI,
[1] M.Phil Research Scholar, [2] Associate Professor,
[1] PG & Research Department of Computer Science,
[2] PG & Research Department of Computer Application,
[1, 2] Hindusthan College of Arts & Science, Coimbatore.

Abstract:-

Acute Lymphoblastic Leukemia (ALL) is the most common cancer in children and adults. At present, diagnosis, prognosis and treatment decisions are made based upon blood and bone marrow laboratory testing. With advances in microarray technology it is becoming more feasible to perform genetic assessment of individual patients as well. By utilizing the information from a few microarray tests that have produced information about both the SNP paroles of patients and additionally quality expression information. By using Singular Value Decomposition (SVD) on Illumina SNP, Affymetrix and cDNA gene-expression data has performed aggressive attribute selection using random forests to reduce the number of attributes to a manageable size. Then by exploring clustering and prediction of patient-specific properties such as disease sub-classification, and especially clinical outcome. By determining that integrating multiple types of data can provide more meaningful information than individual datasets, if combined properly. This method is able to capture the correlation between the attributes. The most striking result is an apparent connection between genetic background and patient mortality under existing treatment regimes.

Keywords: - proteins, Target identification, characteristic frequency, Computational Identification

1. INTRODUCTION

Cancer, in all of its forms, is the second leading cause of death in worldwide. It is also estimated that approximately 30% of these cancer deaths are preventable [25]. Leukemia is the most common malignancy affecting children under the age of 15, but it also affects many adults. The microarray has become a useful research tool and has allowed researchers to begin looking at problems on a much larger scale. As this technology evolves so to do the applications for microarrays. In cancer research, researchers can now look at the expression levels for many thousands of genes as well as a description of an individual's genome given by a set of Single Nucleotide Polymorphisms (SNPs). The amount of data that is being generated is staggering, and there is a need to develop methods for analyzing this data efficiently. These high-throughput technologies have led to many great discoveries which have had many clinical applications. With 20% of leukemia

cases leading to death, there is an opportunity for this type of technology to have a positive effect in this area of research.

1.1 RESEARCH CONTRIBUTION

The objective of this examination is to investigate the relationship between the hereditary qualities of people who have leukemia and regardless of whether they survive the illness. The speculation is that there is a hereditary relationship between a singular's hereditary qualities and their survivability of this malady. By utilizing the information from a few microarray tests that have produced information about both the SNP paroles of patients and additionally quality expression information. Keeping in mind the end goal to break down this perplexing information building up an information mining process that includes adjusting the information to uproot uninformative information traits and afterward utilizing a framework decay procedure to bunch this information. Information mining is performed in the system on the individual datasets and additionally every conceivable mix of them, to check whether joining information together gives more valuable data to the investigation.

2. DATASET USED

2.1 SNP data

DNA is the blueprint from which all living creatures are created. It is the variations in this DNA that allow for the differences both between species as well as within a species. There are four nucleotides that make up DNA; adenosine (A), cytosine (C), guanine (G), and thymine (T). Because DNA is double stranded, these nucleotides work in pairs; A binds with T and C binds with G. In the lifespan of a living being, this DNA will be replicated numerous times. The process of replication is subject to error, and although there are many error-checking processes, it is still possible for a mistake to happen. This is known as a mutation, and there are many different types of mutations. Some are harmful while some are not. Over the course of time these mutations are passed down from generation to generation, and are subject to their own errors as well. Mutations are the reason why there is so much interspecies individuality [2].

2.2 cDNA data

DNA microarrays are high-throughput devices that allow for the collection of a large amount of information on a small glass slide. The basis of this technology relies on how DNA is transcribed. When a particular gene becomes activated, it is transcribed many times in order to produce the necessary proteins. This process involves creating a complementary strand of mRNA which is then used as a template for building these proteins. Thus, a gene that is highly expressed will have many identical mRNA molecules within a cell [33].

2.3 Asymetrix data

The Asymetrix microarray, known as a Gene Chip, works similarly to the cDNA Microarray. The main deference between these two methods is the way in which they are created. An Asymetrix Gene Chip is created through a process known as Photolithography. This technique uses masks and ultraviolet light to build the DNA Probes directly on the slide. This is deferent from the cDNA method where the DNA probes are spotted into wells on a slide. Before the creation of the Gene Chip, the researchers must decide the composition of the probes so that the masks can be created.

2.4 Clinical data

When it is suspected that an individual may have ALL, many clinical tests are run in order to confirm the diagnosis. These tests include full blood

counts as well as bone marrow biopsies. If the test results show an increased white blood cell count and a decreased platelet count then these are the first signs of leukemia. Other clinical results such as an enlarged spleen or liver, chromosomal abnormalities such as a translocation and cytogenetic counts, subtype of the disease as well as the patient's age and sex are all used to diagnose this disease. The clinical data for each patient in this study was processed at the same facility and therefore can be considered to be comparable. Certain patients were missing various types for clinical data, but the mortality was known for every patient. All of these data is used in this study to represent the view of a patient from a clinical perspective. Since this is the data that is available to clinicians making the diagnosis, using these data with the techniques to see how effective these decisions were.

3. TECHNIQUES DATA MINING

3.1 The random forests algorithm

The random forests algorithm [8] is an ensemble classifier that consists of many binary decision trees. A binary decision tree is a method of using nodes in a tree structure to test the attributes of a dataset. The result of these tests is used to split the training data into subsets which are then passed onto the next layer of the tree. This continues until each subset at a node contains only one class. There are many popular decision-tree algorithms, including ID3, CART and C4.5 [23, 30]. Each of these is a supervised learning method, as they require that the data have class labels. One of the challenges with a decision-tree classier is deciding at each layer of the tree, which attribute will provide the best split of the data. Two popular choices for this task are information gain and the gini index.

3.2 Singular Value De composition

Singular Value Decomposition (SVD) is a matrix decomposition technique The singular values on the diagonal of the S matrix correspond to the importance of the amount of variation captured in each column of U. The V matrix captures the variation along the columns of A, which corresponds to the attributes. One of the many useful properties of SVD is that the results that can be visualized. Since the most variation in the data is captured in the first few columns of the U and V matrices and the variation is captured in an orthogonal manner, it is possible to plot the first 2 or 3 columns of these matrices. The resulting image can show clusters or trends in the data that may have otherwise been difficult to see. It is a useful tool for finding structure in complicated datasets. It is especially useful since it can be used on very large datasets which are difficult to handle [28]. The easiest interpretation of SVD is the geometric interpretation. By plotting the U matrix, the data points correspond to the objects plotted in a new space. Data points that lie close to each other in space are correlated with each other and therefore are more alike. Points that lie opposite of each other are negatively correlated with each other and are less alike. Points that are orthogonal to each other have no association. Also, points that lie at the origin are either correlated with everything or correlated with nothing. Either way these points can usually be discarded as not interesting. The power of SVD as a method of clustering can be seen from this explanation. It is able to find points that are interesting, points that are not interesting, as well as associations between points [28].

3.3 Support Vector Machine

A Support Vector Machine (SVM) is a supervised learning method used for classification. This method uses a decision boundary to separate the classes in space.

However, when picturing two classes of data points that are linearly separable, there are an in finite number of boundary lines that can be drawn and it is impossible to know which of these boundaries the best is. This method uses what is known as the maximum-margin hyper plane. The idea is that a separating line is chosen so as to maximize the distance from the nearest data point on each side. The margin of the linear classier is the width that the boundary can reach before hitting a point on each side. The support vectors are the points that lie on the decision boundary and these are the only points used in determining the best way to separate the classes [28]. One common problem with complex data is that there is no simple linear boundary between classes. The idea behind SVM is to project these data into a higher dimensional space using mathematical functions, known as kernels, to a point where a hyper plane can separate the data. If this is done properly, the necessary number of calculations can be minimized so that only the original attributes are needed, making it an efficient algorithm. It is possible to extend this algorithm to allow misclassifications while incurring a penalty for each. As such, there are several parameters that can be changed and may require testing to determine the best setup for a particular experiment. Although this is primarily a two class separator, it is possible to extend this to multiclass prediction. This is one of the most popular and elective classification algorithms to date [9].

4. EXPERIMENTAL MODEL

In order to gain a better understanding of these datasets, the experimentation process began on a general level. This involved performing SVD on the entire dataset. Once some knowledge was gained about these datasets, attribute selection was then performed using the random forests algorithm, and then further exploration of the data was done using SVD and SVM. To see whether or not data integration would be beneficial, the three datasets were combined together, as well as in pairs.

4.1 SVD analysis of data

Using the geometric interpretation of SVD, the goal of these experiments was to see if there were significant clusters in the data. By labeling these images with different clinical features, e.g. mortality, to be able to understand what the clusters represented. By doing this for each of the datasets listed below, monitoring which datasets held the most structure and if combining them would give more information. The geometric interpretation of SVD is based upon plotting the U*S matrix to see if there are clusters in space. The farther away a point is from the origin, the more interesting that point is. Likewise, the points that lie together in space are more correlated with each other than those points which lay further away in space. For all of the following experiments, the ¯first three columns of the U*S matrix were used to plot the points.

4.2 COMBINATION OF DATASETS

There are two possible methods of combining datasets for performing attribute selection. First, the datasets can be combined and then the random forests algorithm can be run to select the best attributes. Second, attribute selection can be performed on the individual datasets and then the best attributes selected from each and combined together. In order to determine what the most appropriate method was for these experiments all possibilities were created. SVM was then run on all of the combined datasets to determine which approaches were the best.

4.3 ATTRIBUTE SELECTION

Attribute selection is the process of removing attributes which contain less useful information for the task at hand. By

choosing attributes that appear to provide the most useful information, the dimensionality of the problem decreases which helps to improve the quality of the experiments. This is especially true for datasets which are as large as these. For example, in the SNP dataset, not all 13917 SNPs are likely to be relevant to this problem. Having such a large number of attributes not only makes it difficult to perform accurate classification, but the tests themselves become very inefficient to run. There are many ways of performing attribute selection, but because of the size of these datasets and the quality of the process a good choice is to use random forests. At the completion of the random forest algorithm, an output file contains the gini index values for all of the attributes which were used for splitting. One problem that exists when using this algorithm with such a large dataset is that, in order for every attribute to be selected in the algorithm, a large number of trees must be built. However, because of the size of these datasets and the limitations of current hardware, it was not possible to run the algorithm long enough for every attribute to be considered. The solution to this was to run the algorithm several times and combine the results of each trial until all, or almost all, of the attributes have index values. The setup of the algorithm for each dataset is shown in Table 3.1. For each dataset the top 25, 50, 100, 250, 500, 1000, 2500 and 5000 genes were selected for further analysis. These subsets were chosen in order to capture the features of the datasets as they change from very few attributes to a large number of attributes.

Dataset	Patients	Attributes	Trees Built
SNP	137	13917	4x30000
cDNA	68	10027	4x30000
Affy	144	22277	4x30000
SNP & cDNA	49	23944	8x20000
SNP & Affy	118	36194	10x14000
cDNA & Affy	55	32304	10x14000
SNP & cDNA & Affy	49	46221	12x11000

Table 4.1: Attribute selection using Random Forests

CONCLUSION

The goal of this research was to investigate the relationship between an individual's genetics and whether or not they survived their battle with acute lymphoblastic leukemia. The data that was used for this study was produced from microarray analysis of the individual's SNPs and gene expression values. These data are complex and high dimensional which provided many challenges for the analysis. By using data mining techniques to analyze these data and created a process of attribute selection followed by clustering through the use of a Singular Value Decomposition (SVD). Using various clinical labels to understand the results that this technique produced. This study has produced many conclusions about both the data and the techniques that were used. The analysis has shown that a separation can be found between patients who live and who die based on both the SNP values and the gene expression values. This suggests that there may be a genetic explanation for why some patients die within the context of current treatment regimes. This is significant and novel as it is not widely accepted that there is a genetic factor which can distinguish patients who live and die. Rather, the genetic factors that are known are related to individuals developing this disease or not. Not able to pinpoint which attributes are responsible for this, but the attribute selection method creates subsets which contain these informative attributes. This finding was supported through many different analyses. The SNP, cDNA and the combined dataset analysis using the data mining procedure showed a clear separation of the data based on the mortality label. Also, the further analysis of the SNP data using various techniques all showed similar results. The validation technique using also showed that these results were not due to random chance. It would be ideal to obtain new data which could be run through the

model, but at this current time this is not possible. This finding has merit. However, it will take further research and fine tuning of the techniques to discover any biological significance. The process of attribute selection is one which must be done carefully. It is unrealistic to assume that the attribute-selection algorithm, in this case the random forest algorithm, will be able to identify all of the biologically significant attributes with such a large dataset. Showing that by evaluating the attribute selection process through a cross validation of attributes in smaller subsets, there are many attributes which are included in these subsets which may only be informative for that particular dataset and are not globally predictive. It is necessary to be more intelligent about the attribute selection process in order to distinguish between predictive attributes and those that only appear to be predictive. It is also shown that the current process of using clinical data to make decisions about diagnosis, prognosis and treatment is not adequate. Although the survival rate is approximately 80%, it can be seen from the analysis that based on the genetics of these individuals there does not appear to be any meaningful relationship to the risk classification the physicians have assigned as seen in the SVD analysis of the clinical data. The nature of these data being complex, high dimensional, constantly changing and with the datasets being biologically connected, makes it difficult to work with. Data mining provides the necessary tools to attempt to understand and learn from this type of data. The data-mining process is involved and requires the researchers to constantly scrutinize the results and to learn from them in order to develop a more intelligent process. With so much data being produced from these high throughput devices every day, it is necessary to develop intelligent and efficient methods of learning from these data and data mining is necessary to take advantage of the wealth of knowledge hidden in these datasets.

REFERENCES

[1] Affymetrix. Genechip microarrays: Student manual. www.affymetrix.com/ about_affymetrix/outreach/educator/microar ray_curricula.affx, 2004. Accessed on October 20, 2009.

[2] G. Alsbeih, N. Al-Harbi, M. Al-Buhairi, K. Al-Hadyan, and M. Al-Hamed. Association between tp53 codon 72-single nucleotide polymorphism and radiation sensitivity of human ¯broblasts. Radiation Research, pages 535{540, 2007.

[3] Orly Alter. Discovery of principles of nature from mathematical modeling of dna microarray data. PNAS, 103:16063{16064, 2006.

[4] A. Archer and R. Kimes. Empirical characterization of random forest variable importance measures. Computational Statistics and Data Analysis, 52:2249{ 2260, 2007.

[5] E. Asgarian, M.H. Moeinzadeh, S. Sharian, A. Naja, A. Ramezani, J. Habibi, and J. Mohammadzadeh. Solving mec model of haplotype reconstruction using information fusion, single greedy and parallel clustering approaches. Computer Systems and Applications, pages 15{19, 2008.

[6] Deepa Bhojwani, Huining Kang, Renee Menezes, Wenjian Yang, Harland Sather, Naomi Moskowitz, Dong-Joon Min, Je®rey Potter, Richard Harvey, Stephen Hunger, Nita Seibel, Elizabeth Raetz, Rob Pieters, Martin Horstmann, Mary Relling, Monique den Boer, Cheryl Willman, and William Carroll. Gene expression signatures predictive of early response and outcome in high-risk childhood acute lymphoblastic leukemia: a children's oncology group study. Journal of Clinical Oncology, 26(27):4376{4384, 2008.

[7] Sikic Branimir, Robert Tibshirani, and Norman Lacayo. Genoics of childhood

leukemia: the virtue of complexity. Journal of Clinical Oncology, 26(27):4367 ,2008.
[8] L Breiman and A Cutler. Random forests. www.stat.berkeley.edu/ ~ breiman/ RandomForests/cc_manual.htm, 2004. Accessed on October 20, 2009.
[9] Christopher J.C. Burges. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2:121{167, 1998.
[10] Daniel Catchpoole, Andy Lail, Dachuan Guo, Qing-Rong Chen, and Javed Khan. Gene expression pro¯les that segregate patients with childhood acute lymphoblastic leukaemia: an independent validation study identi¯es that endoglin associates with patient outcome. Leukemia Research, 31:1741{1747, 2007.
[11] P. Chopra. Microarray data mining using landmark gene-guided clustering. BMC Bioinformatics, 9(92), 2008.
[12] Nigel Crawford, John Heath, David Ashley, Peter Downie, and Jim Buttery. Survivors of childhood cancer: An australian audit of vaccination status after treatment. Pediatric Blood Cancer, pages 128{133, 2009.
[13] R. Diaz-Uriate and S. Alvarez de Andres. Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7(3), 2006.
[14] Christian Flotho, Elain Coustan-Smith, Deqing Pei, Cheng Cheng, Guangchun Song, Ching-Hon Pui, James Downing, and Dario Campana. A set of genes that regulate cell proliferation predicts treatment outcome in childhood acute lymphoblastic leukaemia. Blood, 110(4):1271{1277, 2007.
[15] Centers for Disease Control and Prevention. Leading causes of death. www.cdc. gov/nchs/FASTATS/lcod.htm, May 2009. Accessed on November 11, 2009.
[16] Leukaemia Foundation. Acute lymphoblastic leukemia. www.leukemia.org/ web/aboutdiseases/leukaemias_all.php, 2004. Accessed on July 25, 2009.
[17] Clare Frobisher, Emma Lancashire, David Winter, Aliki Taylor, Raoul Reulen, and Michael Hawkins. Long-term population based divorce rates among adult survivors of childhood cancer in britain. Pediatric Blood Cancer, pages 116{122, 2009.
[18] Lan Guo, Yan Ma, Rebecca Ward, Vince Castranova, Xianglin Shi, and Yong Qian. Constructing molecular classi¯ers for the accurage prognosis of lung adenocarcinoma. Clinical Cancer Research, 11:3344{3354, 2006.
[19] Katrin Ho®mann, Martin J. Firth, Alex H. Beesley, Joseph R. Freitas, Jette Ford, Saranga Senanayake, Nicholas H. de Klerk, David L. Baker, and Ursula R. Kees. Prediction of relapse in paediatric pre-b acute lymphoblastic leukaemia using a three gene risk index. British Journal of Haematology, 140:656{664, 2008.
[20] Amy Holleman, Meyling Cheok, Monique den Boer, Wenjian Yang, Anjo Veerman, Karin Kazemier, Deqing Pei, Cheng Cheng, Ching-Hon Pui, Mary Relling, Gritta Janka-Schaub, Rob Pieters, and William Evans. Gene-expression patterns in drug-resistant acute lymphoblastic leukemia cells and response to treatment. The New England Journal of Medicine, 351(6):533{542, 2004.