A Conceptual Study of the Genetics of Leukemia: an Advanced Data Mining Approach

Krishna Priya S.[1], Sinhgad IMCA, Narhe, Pune, Maharashtra (INDIA) [1]krishnapriya@simca.ac.in

Shivaji D. Mundhe, Sinhgad IMCA, Narhe, Pune, Maharashtra (INDIA)

RahulMahakal, Sinhgad IMCA, Narhe, Pune, Maharashtra (INDIA)

Poonam Sawant Sinhgad IMCA, Narhe, Pune, Maharashtra (INDIA)

Abstract: **In recent years there has been an explosion in the rate of acquisition of biomedical data. Advances in molecular genetics technologies, such as DNA microarrays allow us for the first time to obtain a "global" view of the cell. It has great potential to provide accurate medical diagnosis, help find the right treatment and cure for many diseases. It is still unknown why one person got the leukemia and others don't. It is possible to analyze from the gene expressions the factors behind it. But these micro-array devices generate bewildering amounts of raw data which is difficult for human to analyze. The present paper proposes a conceptual framework to analyze the reason behind the occurrence of the leukemia by analyzing micro array data using data mining techniques.**

Keyword: Luekemia disease, gene-chip, Blood cancer, Data Mining Tools ,micro array

I.INTRODUCTION

Leukemia is a cancer of the blood or bone marrow and is characterized by an abnormal proliferation of blood cells, usually white cells called 'leukocytes'. Even though many investigations were carried out to find out the root cause of this deadly disease, it is still unknown why one person got the disease while others don't have.

It is essential to know the process of formation of normal blood cells, in order to understand cancerous cell growth. It is strongly felt that information regarding root cause of leukemia can be investigated by extracting hidden and useful information from the genes of the patients. An investigation of this type demands the analysis of large number of genetic datasets of various samples to extract the desirable information, which was an impossible task previously. But now, as data mining has become a synonym for the process of extracting the hidden and useful information from datasets. Data Mining or "the efficient discovery of valuable, non-obvious information from a large collection of data" [1] has a goal to discover knowledge out of data and present it in a form that is easily comprehensible to humans.

II.ORIGIN OF RESEARCH PROBLEM

Leukemia is a cancer of the blood or bone marrow. According to Piatetsky-Shapario and Tamayo, "Microarrays are a revolutionary new technology with great potential to provide accurate medical diagnosis, help find the right treatment and cure for many diseases [2] and provide a detailed genome-wide molecular portrait of cellular states." With advances in "gene chip" technology, gathering micro array data has become faster and more efficient than ever before. Single chip microarrays yielding estimations of the absolute expression values of a particular gene afford insight into what a cell is actually doing and how it reacting and changing given various stimuli[3]. Gene expression information can be therefore systematically harvested, stored, and analyzed at later dates. While traditional statistical methods are helpful in preliminary analysis, data mining techniques can not only draw conclusions accurately but also aid in visualizing patterns within the data set itself.

III. REVIEW OF LITERATURE

A computational procedure for feature extraction and classification of gene expression data was proposed by S. Bicciato *et al*. The Soft Independent Modeling of Class Analogy (SIMCA) approach was implemented in a data mining scheme in order to allow the identification of those genes that are most likely to confer robust and accurate classification of samples from multiple tumor types [4]. K. A. Marx et al carried out the data mining of the NCI Cancer Cell Line Compound GI(50) Values for successfully identifying Quinone Subtypes Effective Against Melanoma and Leukemia Cell Classes [5]

A similar approach related to oral and pharyngeal cancers carried out at National Cancer Institute (NCI) used data mining techniques and a subset (1400) of compounds from the large public compounds data for the studies [6] .This study discussed how data warehousing, data mining, and decision support systems can reduce the national cancer burden or the oral complications of cancer therapies.

Piatetsky-Shapario and Tamayo discussed the challenges facing investigators of microarray data mining, and also provided an overview of microarrays using Affymetrix GeneChip®, and discussion of molecular biology and DNA. Some of the challenges cited are gene selection, classification, clustering and visualization, and low-level analysis [2].

According to Van der Puten (2005), the problem with Leukemia cell data is that different types of leukemia cells look very similar and hence data mining on micro-array data is the solution for the problem. Data mining can solve problem of that given data for a number of patients, data mining of microarrays of Leukemia cell data can help accurately diagnose the disease, predict outcomes for given treatment, and recommend best treatment.

Conger cited the significance of using data mining at the genetic level as comparing to "striking genetic gold" [7]. Dunphy performed 'gene expression profiling' in lymphoma and Leukemia [8]. Very useful result on data mining of leukemia was obtained by Glover et al. [9] and they could relate it to mitochondrial genetic data. Labib and Malek used data mining for cancer management in Egypt especially for for childhood acute lymphoblastic Leukemia[10]. Markiewicz and Osowski [11] performed data mining techniques for feature selection in blood cell recognition, and Marx et al. [12] performed data mining of the NCI Cancer cell compound GI50.

IV.OBJECTIVES OF THE RESEARCH

Considering the social and medical impact of the control of leukemia, it is proposed to carry out a data mining analysis of with the available genetical data on leukemia using advanced data mining tools. Apart from a detailed study about the leukemia disease, it is intended to collect vast the sample of genetic data of the people with and without leukemia.

Advanced data mining tools and algorithms will be developed to analyze the collected data, intending to identify patterns to understand the causes of the leukemia disease. Also it is proposed to suggest remedial measures, if any.

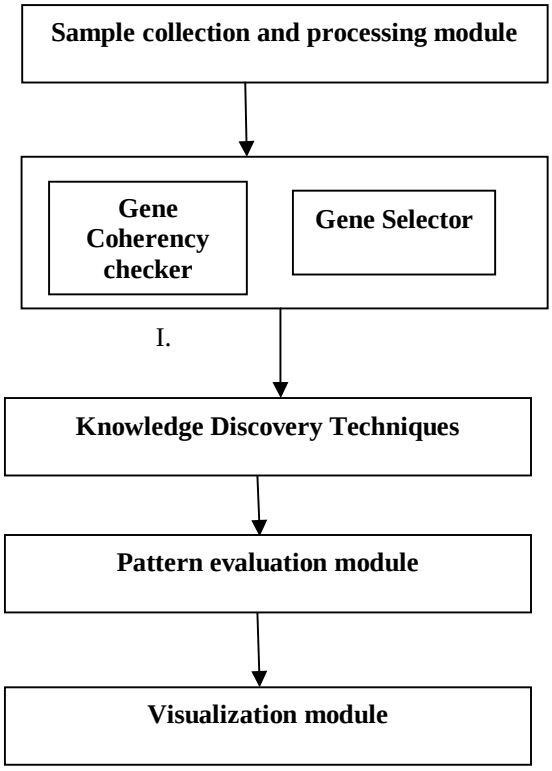
V.SIGNIFICANCE OF THE STUDY

When one hear that he/she have cancer, it's natural to wonder what may have caused the disease. No one knows the exact causes of leukemia. Doctors seldom know why one person gets leukemia and another doesn't. "About 6,000 people in India are diagnosed with leukemia and blood-related cancers out of those 3,800 children each year".

This research is a contribution to the cancer research which may be aims at relieving the pain of lakhs of people including children all over the world who is suffering from the deadly pain of this fatal disease. Microarrays are a revolutionary new technology with great potential to provide accurate medical diagnostics help to find the right treatment and cure for many diseases and provide a detailed genome-wide molecular portrait of cellular states. 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. The amount of data that is being generated is staggering, and there is a need to develop methods for analyzing this data efficiently. In this context it is a candidate for the strength of data mining.

VI. PROPOSED FRAMEWORK



VII. CONCLUSION

Data mining is the process of extracting the hidden and useful information from datasets. It is possible to mine the gene expression data and extract the hidden patterns which will pave the way for new findings about leukemia there by a great contribution to the society. The paper presents a conceptual framework mining the microarray data.

REFERENCES

[1] J. P. Bigus, "Data Mining with Neural Networks", New York: McGraw-Hill, 1996

[2] Piatetsky-Shapario, G. and Tamayo, P , "Microarray Data Mining: Facing the Challenges, SIGKDD Explorations, V. 5, n.2, pp. 1-5,2003.

[3] Segall, R. S. and Pierce, R. M, "Data mining of Leukemia cells using Self-Organized Maps", Proceedings of 2009 ALAR Conference on Applied Research in Information Technology, February 13, 2009.

[4] S. Bicciato, A. Luchini, C. Di-Bello, "Marker identification and classification of cancer types using gene expression data and SIMCA", Germany: Methods-of-information-in-medicine, 2004.

[5] K. A. Marx, P. O'Neil, P. Hoffman, M. L. Ujwal, "Data mining the NCI cancer cell line compound GI(50) values: identifying quinone subtypes effective against melanoma and leukemia cell classes", US Journal-of-chemical-information-and-computer-sciences, 2003.

[6] G. A Forgionne, A. Gagopadhyay, and M. Adya, "Cancer Surveillance Using Data Warehousing, Data Mining, and Decision Support Systems", Topics in Health Information Management, vol. 21(1); Proquest Medical Library, August 2000

[7] Conger, K. (2006), Stanford/Packard, "Scientist's data-mining technique strikes genetic gold," Medical News Today,January11, http://www.medicalnewstoday.com/articles/36009.php

[8] Dunphy, C. H., (2006), "Gene expression profiling data in lymphoma and leukemia: Review of the literature and extrapolation of pertinent clinical applications," Archives of Pathology & Laboratory Medicine, April, v. 130, pp. 483-520.

[9] Glover, C.J., Rabow, A.A, Igsor, Y. G., Shoemaker. R.H., and Couell, D.G. (2007), "Data mining of NCI's anticancer screening database reveals mitochondrial complex I inhibitors cytotoxix to leukemia cell lines, Biochemical Pharmacology, v. 73, n.3, pp. 331-340.

[10] Labib, N.M. and Malek, M.N. (2005), "Data mining for cancer management in Egypt Case Study: Childhood Acute Lymphoblastic Leukemia", Proceedings of World Academy of Science and Engineering and Technology, v. 8, October, pp. 309-314.

[11] Markiewicz, T. and Osowski, S., "Data mining techniques for feature selection in blood cell recognition:, Proceedings of European Symposium on Artificial Neural Networks (ESANN'2006), April 26-28, 2006, pp. 407-412.

[12] Marx, K.A., O'Neil, P., Hoffman, P., Ujwal, M.L. (2003), "Data mining the NCI cancer cell compound GI50 values:

[13] Broad Institute (2007), Cancer Program Data Sets, http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi