International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 3 (2015) pp. 6285-6292

http://www.ripublication.com

Cancer Detection Based Mining System by Importing Intelligence to Genes

P.Suganya[1] and E.Thenmozhi[2]

[1] Student, Department of Computer Science and Engineering.
Sathyabama University, Chennai, India
Email:sugan3991@gmail.com
[2] Assistant Professor, Faculty of Computing,
Sathyabama University, Chennai, India
Email:rajthenu@gmail.com

Abstract

Cancer classification using gene expression data is the process by which information from a gene is used in the synthesis of a functional gene product. Gene expression profiling is a technique used in molecular biology to query the expression of thousands of genes simultaneously. Gene Ontology, is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species. Monitoring the activities of genes by visualizing and normalizing the data sets in order to evaluate the data quality by removing the outliers is very challenging. Then, by introducing several representative approaches to predict the results and further providing various suggestion's for cancer treatment. Finally, this paper suggest the promising trends in the medical field.

Key Terms— Association rules, Gene expression data, Ontological mapping, Semantic Clustering.

I INTRODUCTION

Cancer research is one of the major research area in the medical field. Cancer is a class of diseases characterized by excess cell growth. There are different types of cancer, and each is distributed by the type of cell that gets affected. Cancer is caused in the body when damaged cells divide uncontrollably to form masses of tissue called tumors. Tumors grow in the different kinds of systems, and they can release hormones that alter body function. Tumors that stay in one spot and have limited growth are generally called as benign. When the tumor spreads to other parts of the body and grows, affecting other healthy tissues, it is said to have changeover. This process is

known as metastasis, and in result it is very difficult to treat. The researchers say, cancer mortality is mainly due to metastatic tumors, those that grow from cells that have moved from their original to another part of the body. Only 10% of cancer deaths are caused by the primary tumors. Malignant cells are more agile than non-malignant ones. Malignant cells can penetrate more easily through smaller gaps, also by applying greater force on their environment compared to other cells.

The different sources of cancers, they are:

- Deviations in DNA.
- Carcinogens.
- Genetic Predisposition.
- Other Medical Factors.

However, the cause of many cancers remains unknown. Apart from genetic causes, there are certain environmental and external factors too that participate in cancer formation within an organism, and can be categorized under epigenetics. Epigenetics is the study of changes in gene expression caused by certain base pairs in DNA, or RNA, being "on" or "off" again, through chemical reactions.

Cancer symptoms are quite varied and depend on where the cancer is located, how it has spread, and how the size is. For Example, Skin cancer is often noted by a change in appearance on the skin. Some oral cancers present white patches inside the mouth or white spots on the tongue. Early detection of cancer can greatly improve the odds of successful treatment and survival.

BIOLOGICAL INFORMATION

A gene is the basic physical and functional unit of heredity. Gene expression is the process of transcribing a gene's DNA sequence into RNA. The expression of the genetic information stored in the DNA molecule occurs in two stages: (1) transcription stage where the DNA molecule is transcribed into mRNA, (2) translation stage where mRNA is translated into the amino acid sequences of the proteins that perform various cellular functions.

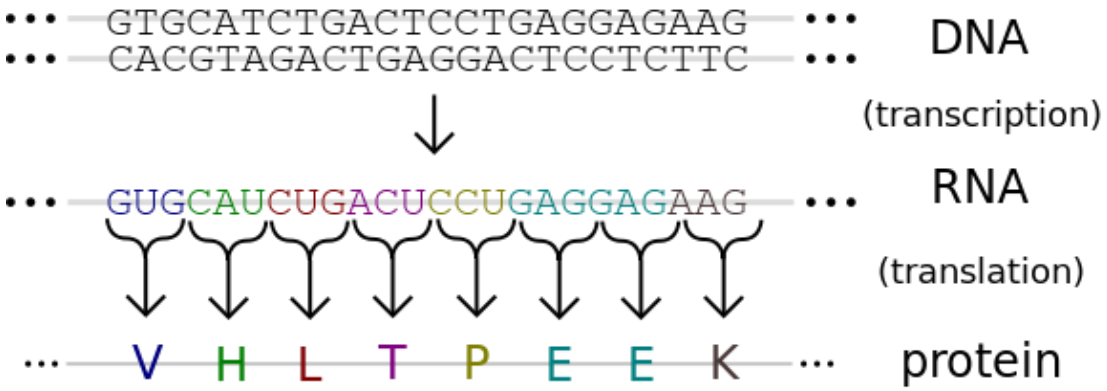

Fig 1: Genetic Code Diagram

Gene Expression matrix analysis

Genome annotation is the process of attaching biological information to sequences. Gene expression matrix analysis can be studied:

Genes expression profiles by comparing rows in the expression matrix;

Samples expression profiles by comparing columns in the expression matrix.

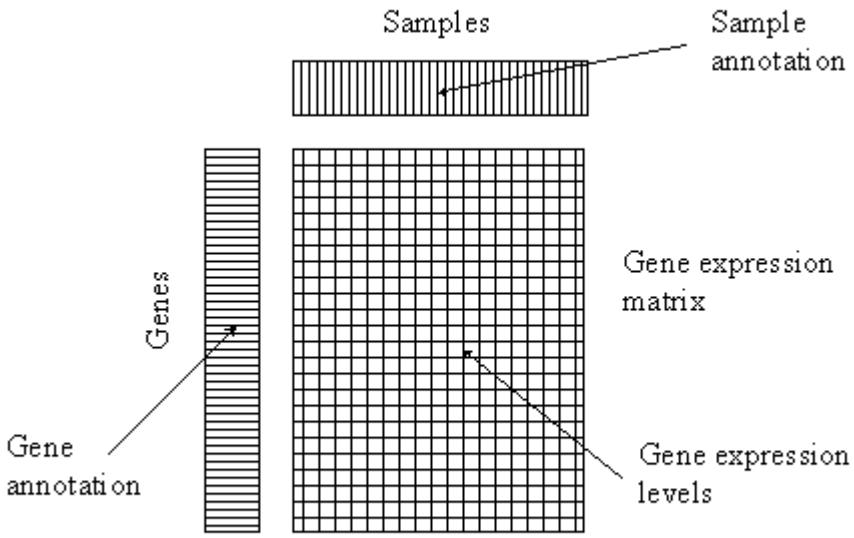

Fig 2: Gene Expression Diagram

By comparing rows and columns, we can look either for similarities or for differences [2]. The method using gene expression profiles is more objective, accurate and reliable [1].

Euclidean Distance is used to measure distance between two objects [9]. We can regard the objects as points in n-dimensional space or as n-dimensional vectors, where n is the number of genes for sample comparison, or number of samples for gene comparison. Clustering analysis is to group together objects (genes or samples) with similar properties, which is the first step in data mining and knowledge discovery [9]. In case of the unsupervised analysis, a clustering algorithm should identify these clusters but in supervised analysis, the task is to find a set of classification rules. Hierarchical clustering works by iteratively partitioning clusters with the complete set. After each joining of two clusters, the distance between all the other clusters and new joined clusters are recalculated. Note that to obtain a particular partitioning into clusters, the distance should be chosen by means. The k-means clustering algorithm typically uses the Euclidean properties of the vector space [9]. The algorithm calculates the center points in each subspace and adjusts the partition so that each vector is assigned to the cluster the center of which is the closest. This is repeated iteratively until the partitioning stabilizes. The gene expression profiles are

meaningful only in case of the experimental method. For this to become a reality, acknowledged ontologies and controlled terminologies for tissues, cell types, and treatments, as well as for array designs, image analysis and various protocols have to be developed.

Gene expression microarrays provide a snapshot of all the transcriptional activity in a biological sample. Multiclass predictions are been applied to the data sets. Support vector machines based classification strategy may not be the optimal method for every type of multiclass problem [4].

II EXISTING SYSTEM

In the existing systems, the system is ranging from old nearest neighbor analysis to support vector machine manipulation for the learning portion of the classification model. Molecular diagnostics offers the promising option of systematic human cancer classification, but these tests are not widely applied because characteristic molecular markers for most solid tumors have not yet identified. Microarrays and Serial analysis of gene expression technology measured thousands of genome-wide expressions values in parallel [2]. Supervised Multiclass Attribute Algorithm is used in order to find the co-regulated clusters of genes. Agglomerative hierarchical Clustering Algorithm uses a bottom-up strategy [10]. It typically starts by letting each object form its own cluster and iteratively merges clusters into larger and larger clusters, until all the objects are in a single cluster or certain termination conditions are satisfied.

Draw backs

- Agglomerative clustering finds the two clusters that are closest to each other and combines the two to form one cluster. Because two clusters are merged per iteration, an agglomerative method requires at most n iterations.
- Existing technique does not measure the expression level of a gene.
- No clear picture of supervised classifier.
- Only few genes are been examined due to lack of time and monetary constraints.
- Prediction of cancer is not accurate.
- Results are vague and imprecise.

RELATED WORKS

In this work, the Gene Expression data set is given as an input in order to analyze it with the Experts documents. Semantics Sequence Structure Algorithm is used for designing gene expression which is used for comparison of gene data's. Based upon the analysis from the input data sets, we will analyze the extracted gene to predict its characteristics and also its features. Gene Knowledge Extraction mainly focuses towards the performance of the individual gene expression data's. Ontological mapping, in which mapping of two different gene expression data's is carried to find the differences in gene characteristics. It is also used as a solution provider in today's

analysis in which it provides insights on the pragmatics of ontological mapping towards gene expression elaboration. Genes work coordinately as gene expression or gene networks in which they are designed to find gene expression patterns by grouping genes. Therefore, the genes whose expressions are modulated by the genetic variants will act as Trans-Regulated gene module in humans to identify the cancer. Association of the genes with multiple traits replicated in Cancer Diogenics (An analysis of cancer prediction in human genes), which is an independent study of Gene Ontology. This Gene Ontology is implemented to enrich the cancer diagnosis of gene analysis. Based on the analysis of the Comparative Knowledge consolidator, the Structural and Semantic rules predicts the cancerous genes. Under the Best Rule Classification constrain, we suggest the final prediction of cancer. Finally, we will evaluate the performance of the genes and also suggest the medicines that are needed to be taken up for the Cancer Diagnosis.

III PROPOSED SYSTEM

The main scope of this work is to face all the challenges in terms of biological relevance. In the proposed system, intelligence is been applied to the genes for faster approach. Swarm intelligence is the collective behavior of distributed, self- organized systems, natural or artificial [11]. Particle Swarm Optimization (PSO) [11] is a global optimization algorithm for dealing with problems in which a best solution can be represented as a point or surface in an n-dimensional space. The main advantage of such an approach over other strategies such as simulated annealing is that the large number of genes that make up the particle swarm make the technique impressively resilient to the problems of local minima. Ant Colony Optimization (ACO) [11] is a probabilistic technique useful in problems that deal with finding better paths through graphs. Genes locate optimal solutions by moving through a parameter space representing all possible solutions. The stimulated genes record their positions and the quality of their solutions, so that in later stimulation iterations more genes locate better solutions.

SYSTEM ARCHITECTURE

In this architecture diagram, the gene expression data sets are been compared with the experts documents to mine the gene expression. In which the gene expression is the most fundamental level at which the genotype develops to the phenotype. The genetic code found in DNA is interpreted by gene expression and their properties develops to phenotype organisms. The Experts documents are used to provide evidence about a suspicious or questionable document using variety of scientific processes and methods. Ontology is a formal representation of knowledge as a set of concepts within a domain and the relationships between those concepts. Gene ontology preserve the integrity of the data and also unify the representation of gene and gene product.

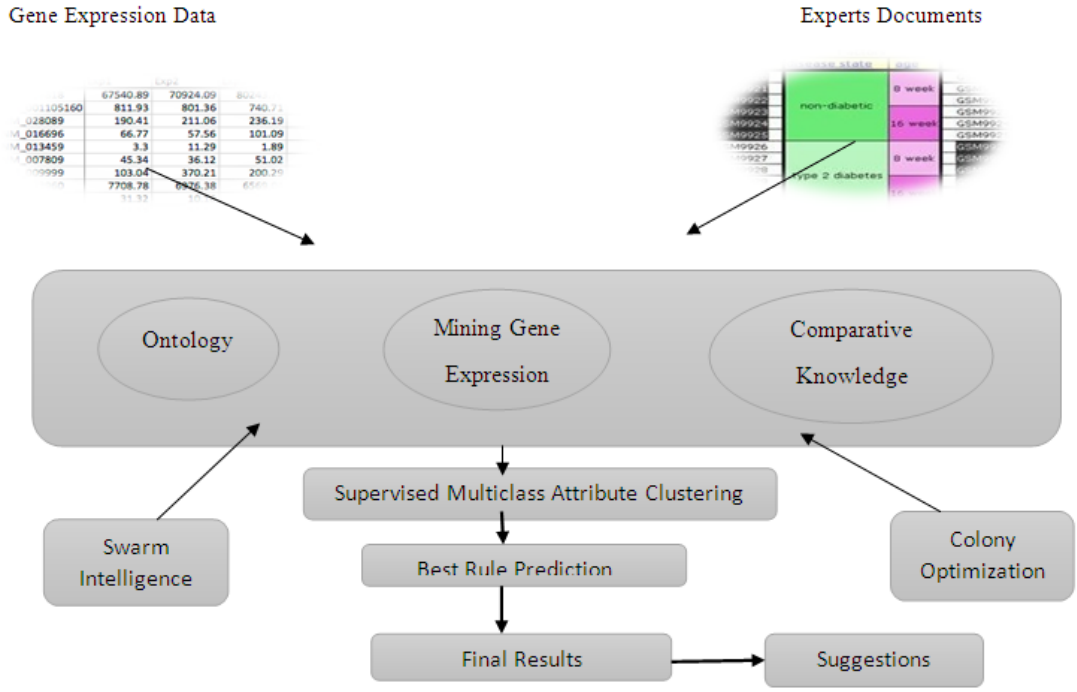

Fig 3: Architecture Diagram

SAMPLE DATA SETS

Cells are the fundamental working units of every living organisms. All the instructions needed to direct their activities are contained within the chemical deoxyribonucleic acid or DNA. A DNA molecule is a double-stranded polymer composed of four basic molecular units known as nucleotides. Each nucleotide consists of a one of the four nitrogen bases and other components. The nitrogen bases such as adenine(A), guanine(G), cytosine(C) and thymine(T) are found. The double helix structures are held together by the hydrogen bonds between the nitrogen base pairs: A with T, C with G. DNA sequence is a particular arrangement of the base pairs in the DNA strand, e.g., CTTGAATCCCG.

Fig 4: Sample Data Sets

IV RELATED ISSUES

In this section, some important issues in cancer classification using gene expression data. They include the issue of biological significance of a cancer classifier. Classifiers are evaluated based on the classification accuracy and efficiency. The major goal of expression data analysis is to provide the biologically meaningful information about the genes and related things.

Classification errors is also an issue in which the genes are grouped into misclassified or non-classified. Gene selection or classification based on genes might lead to classification success but mistake in the process does not provide any biological information about genes related to cancer development. For analysis, the data's are not standardized as the gene expression data are from different laboratories. Because of this, noise and error will be introduced during classification. Sometimes the data's will contain missing values were the efficiency is required to combine all the data's.

V CONCLUSION

Systematic approach to cancer classification is of great importance to cancer treatment. In the past, cancer classification methods are all clinical based and were limited to diagnostic ability. Gene expressions contains the keys to fundamental problems of cancer diagnosis and cancer treatment. Performance evaluation done in three aspects: computation time, classification accuracy and biological relevance gives a precise results.The Framework exploits a number of machine learning and data mining techniques to detect cancer. In future reasearch, cancer have to be

predicted in advance by using genes of previous generation in order to improve the survival of new generation.

Thus,cancer classification using gene expression data has a great future in providing systematic approach for differentiating different tumor types.

REFERENCES

[1] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown, "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray, "*Science*, Vol. 270, pp. 467-470, 1995.

[2] Brazma, Jaak Vilo, "MiniReview-Gene expression data analysis", *FEBS Lett.*, Vol.480, pp. 17-24, 2000.

[3] N. Revathy and R. Amalraj, Accurate cancer classification using expressions of very few genes", *Int. J. Comput. Appl.*, Vol. 14, No. 4, pp. 19-22, Jan. 2011.

[4] Sridhar Ramaswamy and Pablo Tamayo, "Multiclass cancer diagnosis using tumor gene expression signatures", Available: http://www.pnas.org/cgl/doi/10.1073/pnas.211566398.

[5] E.Shay, (2003,Jan), "Microarray cluster analysis and applications" [Online]. Available: http://www.science.co.il/enuka/Essays/Microarray-Review.pdf.

[6] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classificatiom of tumors using gene expression data", *J. Amer. Statist. Assoc.*, Vol.97, No. 457, pp. 77-87, Mar.2002.

[7] Kristopher L. Patton, David J. John, James L. Norris, Daniel R. Lewis, and Gloria K. Muday, "Hierarchical Probabilistic Interaction Modeling for Multiple Gene Expression Replicates", IEEE/ACM *Trans. On Comput. Biology and Bioinformatics.* Vol. 11. No. 2. March/April 2014.

[8] The Gene Ontology Consortium (January 2008). "The Gene Ontology project in 2008". *Nucleic Acids Res.* **36** (Database issue): D440–4. doi: 10.1093/nar/gkm883.PMC 2238979. PMID 17984083.

[9] Daxin Jiang, Chun Tang and Adiong Zhang, "Cluster analysis for gene expression data: A Survey", IEEE *Trans. Knowl. Data Eng.*, Vol. 16, No. 11, pp. 1370-1386, Nov. 2004.

[10] Shaurya Jauhari and S.A.M. Rizvi, "Mining Gene Expression Data Focusing Cancer Therapeutics: A Digest", IEEE/ACM *Trans, On Comput. Biology and Bioinformatics.* Vol. 11. No. 3. May/June 2014.

[11] Available: *http://www.swarm intelligence.com/*