International Journal of Innovative Research in Computer Science and Engineering (IJIRCSE)
www.ioirp.com ISSN: 2394-6364, Volume – 1, Issue – 7. June 2015

Gene Expression Data Analysis Using Ant Based Clustering

[1]R. Rajeswari
Research scholar,
Department of Computer Applications,
St. Peter's University, Chennai. India

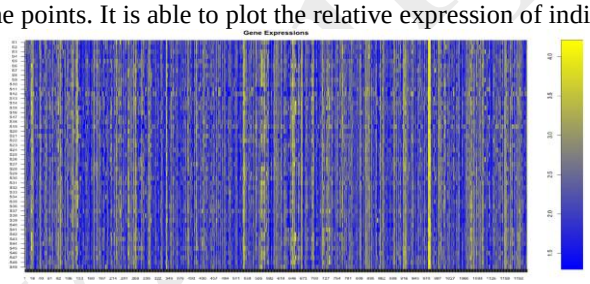
[2]Dr. G. GunaSekaran
Principal
Meenakshi College of Engineering
Chennai. India

***Abstract*— Gene expression data is analyzed to identify the hidden features thereby, predicting the functioning and properties of the gene. In this paper we have presented a new ant-based clustering algorithm for knowledge discovery. Without any pre assumptions regarding the number of clusters or the features of the clusters, thus a reliable clustering algorithm not relying on any pre assumed constants. This paper also provides major steps on collection of gene dataset on basis of different formats, sources, reliability and acquisition of data used for clustering technique, while summarizing their features extraction of gene data.**

Keywords: Data mining, clustering, ant colony algorithm, bio infomatics, Gene expression data

I. INTRODUCTION

Gene is a combination of DNA and RNA chains which controls the levels of proteins that form the structure of any living cell or tissue. This Gene protein controls the functionality, behavior and traits of the life cell while also storing the heredity information. A normal human gene sequence might consist of anywhere around 20,000 to 25,000 strands of gene protein. For the purpose of studying the gene, this sequence of strands is expressed in the form of Gene expression. This Gene expression in other words can also be described as sequence of data that is encoded by the different levels of different protein types that constitute the Gene. This encoded sequence control the functionality of the Gene that includes control functions, behavior, traits, and heredity information from the parent Gene. This gene is stored as a Microarray data in public databases such as the Gene Expression Omnibus (GEO) for further analysis. Microarray was performed a list of gene expression values that identify the gene that are differentially regulated at specific time points. It is able to plot the relative expression of individual gene at a time.



Microarray Gene Expression Data

Microarray data analyze the features of data in many dimensions, there are several techniques for reducing the dimension of the feature vectors extracted from a microarray are,

Feature transformation: It compares nonlinear feature transform schemes with linear feature transform schemes they conclude that the best methods the Locally Linear Embedding feature transform. The low pass filtering the signal is used to increase the dimension of the Orthogonal Linear Discriminate Analysis subspace, and to improve the accuracy, a method for combining the original features to obtain new features is proposed.
Feature selection: The data are projected onto a lower dimensionality space before the selection Sequential Forward Floating Selection (SFFS).

Clustering techniques has been proving itself as a reliable and useful technique over the years to solve this problem. Genes with similar expression patterns can be clustered together to reveal hidden information to help us understand the gene functions, gene regulation and cellular processes as whole.

II. LITERATURE SURVEY

Balaji Venkataraman compares the quality of the generated data with published data and is superior to that obtained using spotted oligonucleotide microarrays [1]. Erliang Zeng produces integrated heterogeneous constrained dataset [2]. Kun Gao in 2014 proposes DNA matures the gene expression data [3]. Chun-Hou Zheng in 2006 the sequential floating forward selection

technique is used to select the independent components of the DNA microarray data for classification [4]. Jesse M. Engreitz in 2010 presents a unique opportunity for intelligent data mining methods to extract information about the transcriptional modules underlying these biological processes [5]. V. Sitras in 2015 share same common traits in their gene expression profile indicating common pathways in their pathophysiology [6]. Soheila Khodakarim in 2014 two new methods, in comparison to the previous ones, is introduced for GSA [7]. Kaiyang Liao, Guizhong Liu and et., produces a result on K-means algorithm is a typical partition-based clustering method [8]. Petri Toronen and et al., produce an algorithm that dealing with noisy data, and is capable of generating an appealing map of data sets in both 2D and 3D space [9] [10]. Md. Bahadur Badsha proved the clustering of gene data using hierarchical clustering algorithm [11]. Zhan C.T., and Janne Nikkila, Petri Toronen provides a graph theoretical approach treats the entire data set as a graph [12]. Adrian E Raftery and Nema Dean clusters the data set as a finite mixture of probability distributions using model-based clustering [14]. Daxin Jiang and etl., clusters the data on the basis of high-dimensional dense cluster [15].

III. EXISTING MODEL

Gene expression data is a standard prefatory technique used for similar gene identification. Some of the data sources are also likely to be of great reinforcement in the analysis of gene expression data. Where the sources include protein interaction data, transcription factor and regulatory elements data, comparative genomics data, protein expression data and much more. [1]

Data Sources:

Data Source is dynamically collects and compiles data from many scientific databases, and thereby attempts to encapsulate the genetics and molecular biology of genes from the genomes. We collected a set of data sources used in bioinformatics experiments obtained from public databases like Bioinformatics laboratory, various hospitals, etc. [2]

Data Format:

Microarray Gene Expression Database Group (MGED), a consortium of academic and commercial organizations with the shared goal of defining standard formats that would allow gene expression data repositories to share and exchange data. [3]

Many new tools are developed and maintained which convert data from sources into a well-defined data format, such as one based on XML, image, chip format, and similar formats.

Data Acquisition:

Data acquisition is the process of sampling the real world signals that can be measured and convert sampling result into digital numeric values that can extracted as an input file for any classification and clustering process. The sample data can be obtained from online open source data or by online payable training data and also by query based data requisition are the methods used to collect the data. [5]

Data Reliability:

Reliability analysis enables filtering of microarray data before estimating the expression ratios. This reliability based filtering can dramatically reduce number of false positives. Assessment of reliability of microarray data and estimation of signal thresholds using mixture modeling. It process on the raw data, not on the expression ratios that may be based on unreliable signals. This accepts single or dual channel data. With several built-in data transformation options, it is adaptable to any data distribution to find the best possible normal mixture model for the data. [4]

Data format:

Gene input values can be represented in 0's and 1's. Where the data's are represented only in digits is called numerical data, and the missing value is represented as "Na".

011 0101 0.98
012 0102 -0.76
013 0103 -0.06

Gene database can also be represented in alphabetical values which is known as text format of gene expression data, [2]

P1 (reverse) CTAACTAATTTTATTGGACTAGGC
P2 (forward) TTCGTAAGCCGAGAGGAATGGGG
P3 (forward) TCTCTGTGCTGTGAATAACT
P4 (reverse) ACAAACAACAATAGCCTTT
Oligo (dG)-adaptor GGCCACGCGTCGACTAGTACG10
Expression primers
P5 CGCGGATCCTTGCCTAGTCCAATAAAATTAG

P6 CCGCTCGAGTTAATCATCCCAATTCAGTG
RT primers
RTF GTCCAGTGCCCATAGTAGTGAT
RTR CGATGCTTCGGGTGTTG
Gene data which represented by a single character, word or numerical number. This is recognized as a categorical factor with several levels.

01	0101	F	Adult	0
02	0203	M	Adult	0
03	0305	F	Adolescent	1

Input quantitative covariates from a plain text file. Each quantitative covariate is recognized as a continuous variable. Where the list of gene values is represented in sequential or continuous values.
#numeric
#AFFX-BioB-5_st
206.0 31.0 252.0 -20.0 -169.0 -66.0 230.0 -23.0 67.0 173.0 -55.0 -20.0 469.0 -201.0 -117.0 -162.0 -5.0 -86.0 350.0 74.0 -215.0 193.0 506.0 183.0 350.0 113.0 -17.0 29.0 247.0 -131.0 358.0 561.0 24.0 524.0 167.0 -56.0 176.0 320.0

Gene dataset is represented in rows and columns in excel spreadsheet. The GRP file format contains a single gene set in a simple newline-delimited text format. The GMT or GMX file formats to create gene sets, rather than using the GRP file format. The GRP format contains a line for each gene, one gene per line. Lines that start with a pound sign (#) are ignored. [6]

Data mining and Clustering has evolved as a collection of gene dataset according to the existing methods of this paper, with a common goal of extracting meaning full predictions from a extensive data sets of gene microarrays. A microarray typically consists of a large number of DNA sequences from a collection of similar or different tissue samples. In this paper, we will focus on the cluster analysis methods used for analyzing gene expression data of DNA sequences to identify the role of a particular gene in the gene expression data. The different clustering techniques applicable to gene data can be categorized into few broad categories like K-Means Algorithms [8], Self-Organizing Map [9], Hierarchical Clustering [10], Graph-Theoretical Approaches [11], Model-Based Clustering and A Density-Based Hierarchical Approach.[12] [14] [15].

IV. PROPOSED WORK

We have observed that most of the algorithms though are fast and robust but leaves a doubt regarding the reliability or correctness of the outcome, since they require us to determine certain initial parameters that determine the reliability of the entire algorithms. Inspired by the natural behavior of ants in finding the shortest path to the best food source even in the absence of any previous knowledge about the sources, we proposes a new clustering algorithm in this paper based on the ant colony to cluster the genes on the basis of the enzymes that can be used to detect a group of genes whose expression level changes in the same pattern. This is a population based Meta heuristic that can be used to find approximate solutions to difficult optimization problems. We expect that the algorithm can effectively uncover the hidden patterns for accurate identification of gene function and predicting its role in gene behavior.

The basic environment of the algorithm consists of randomly placed high-dimensional data objects, having several attributes in a bi-dimensional grid. In this method objects are placed as pick up drop method.
Ant Colony Algorithm:

- Ants (blind) navigate from nest to food source
- Shortest path is discovered via pheromone trails

$$\frac{n-ants*\frac{pheromone}{x-distance}}{\frac{distance_{(longer\ path)}}{time}} < \frac{n-ants*\frac{pheromone}{x-distance}}{\frac{distance_{(shoter\ path)}}{time}}$$

- Each ant moves at random
- Pheromone is deposited on path
- Ants detect lead ant's path, inclined to follow
- More pheromone on path increases probability of path being followed

Algorithm in Pseudo code:

Create construction graph
Initialize pheromone values
while not stop-condition do
Create all ants solutions
Perform local search
Update pheromone values
end while End Do

For the purpose of gene mining we propose to implement this process in three steps

Step: 1 Feature Extraction

In this stage the main features of objects are extracted and the method of comparison.

Step: 2 Similarity Computations

The similarity between the objects taken into consideration in term of these chosen features attributes.

Step: 3 Grouping

The result of similarity or dissimilarity computation is presented in the next step grouping, the form of partitioning these objects into groups.

V. CONCLUSION

In this paper we have presented a new clustering algorithm is ant-based system for knowledge discovery. The ACO introduces new ideas and modifications to improve the convergence providing a reliable method for optimization. The main features of this algorithm are that, it does not require pre-establishing the number of clusters or any other information about the feature of the clusters. Though we are not certain on the time complexity of the algorithms it certainly provides an angle for improvising in the phoneme updating algorithms. As further work, this ant-clustering method is implemented by collecting different types of constrained data, from any hospital database or online gene database sources in text format and to produce the result by using MATLAB, Rapid miner as a supporting tool.

References

[1] Balaji Venkataraman, Madavan Vasudevan, Amita Gupta "A new microarray platform for whole-genome expression profiling of Mycobacterium tuberculosis" in Journal of Microbiological Methods 97 (2014) 34–43.

[2] Erliang Zeng, Chengyong Yang, and etl., "Clustering Genes using Heterogeneous Data Sources" in 3Bioinformatics Research Group (BioRG), School of Computing and Information Sciences, Florida International University, 11200 SW 8th Street, Miami, FL 33199.

[3] Kun Gao, Xiang-yuan Deng, He-ying Qian, Guang-xing Qin, Cheng-xiang Hou, Xi-jie Guo" Cloning and expression analysis of a peptidoglycan recognition protein in silkworm related to virus infection" in journal homepage: www.elsevier.com/locate/gene , Gene 552 (2014) 24–31.

[4] Chun-Hou Zheng, De-Shuang Huanga, Li Shanga, "Feature selection in independent component subspace for microarray data classification" in Neurocomputing 69 (2006) 2407–2410

[5] Jesse M. Engreitz a, Bernie J. Daigle Jr. b, Jonathan J. Marshall a, Russ B. Altman" Independent component analysis: Mining microarray data for fundamental human gene expression modules" in Journal of Biomedical Informatics 43 (2010) 932–944.

[6] V. Sitras , C. Fenton , G. Acharya "Gene expression profile in cardiovascular disease and preeclampsia:A meta-analysis of the transcriptome based on raw data from human studies deposited in Gene Expression Omnibus" in Placenta 36 (2015) 170e178

[7] Soheila Khodakarim, Seyyed Mohammad Tabatabaei, Hamid AlaviMajd," The Multivariate Nonparametric Methods for Identifying Gene Sets with Differential Expression" in Gene 552 (2014) 18–23.

[8] Kaiyang Liao, Guizhong Liu, Li Xiao, and Chaoteng Liu "A sample-based hierarchical adaptive K-means clustering method for large-scale video retrieval", Knowledge-Based Systems, Elsevier,Vol 49, pp 123-133,2013.

[9] Petri Toronen, Mikko Kolehmainen, Garry Wong, and Eero Castren," Analysis of gene expression data using self-organizing maps", FEBS Letters, Federation of European Biochemical Societies,Vol 451,pp 142-146,1999.

[10] Janne Nikkila, Petri Toronen, Samuel Kaski, Jarkko Venna, Eero Castren, and Garry Wong," Analysis and visualization of gene expression data using Self-Organizing Maps", Neural Networks Elsevier,Vol 15, pp 953-966,2002.

[11] Md. Bahadur Badsha, Md. Nurul Haque Mollah, Nusrat Jahan, and Hiroyuki Kurata, "Robust complementary hierarchical clustering for gene expression data analysis by β-divergence", Journal of Bioscience and Bioengineering, Elsevier,Vol 116,No3, pp 397-407,2013.

[12] Zhan C.T.," Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters ", Computers, IEEE Transactions,Vol C-20, issue 1, pp 68-86, Jan 1971.

[13] Wu, Z. and Leahy, R., "An optimal graph theoretic approach to data clustering: theory and its application to image segmentation ", Pattern Analysis and Machine Intelligence, IEEE Transactions on, Vol 15 , Issue 11 ,pp 1101-1113,Aug 2002.

[14] Adrian E Raftery and Nema Dean, "Variable Selection for Model-Based Clustering", Journal of the American Statistical Association, Vol 101, Issue 473, pp 168-178,2006

Daxin Jiang, Jian Pei, and Aidong Zhang," DHC: a density-based hierarchical clustering method for time series gene expression data", In Proceedings of Third IEEE Symposium on Bioinformatics and Bioengineering, pp 393–400, March 2003.