Gene Classification: A Review

Shadi Aljawarneh, Bassam Al-shargabi
Software Engineering Dept, Faculty of Information Technology
Isra University
Amman, Jordan
{Shadi.jawarneh, bassam.shargabi}@ipu.edu.jo

Hasan Rashaideh
Computer Science Dept
Al-Balqa' Applied University, Jordan
rashaideh@gmail.com

***Abstract*—Today, it is possible to monitor a gene expression on a genomic scale using hierarchical clustering, DNA micro-arrays and k-means partitioning which are being the most popular methods. Several tools make use of the GO ontologies or the gene associations provided by consortium members or even individuals. While some progress has been made in addressing the gene classification, current methods are restricted by the limitations of the clustering and visualizations techniques. For example, Avadis, BiNGOb and DAVID tools are based on visualization for gene expression data. In visualization, gene annotations are visualized in as a table view and so the granularity of the GO DAG can be viewed freely by the user or use CLASSIFI (Cluster Assignment for Biological Inference) which is a data-mining tool that can be used to identify significant co-clustering of genes with similar functional properties such as cellular response to DNA damage. Furthermore, Current research is generally more concerned with the clustering and visualizations techniques for gene expression data analysis.**

To enhance the bioinformatics, many researchers and technicians have preferred to match the clustering to the specifications of biomedical applications. In this papered, we have reviewed a number of clustering algorithms for different approaches and data types.

***Index Terms*—Gene expression, Data mining, Semantics, CLASSIFI, DNA.**

I. Introduction

The gene expression is the process by which mRNA and eventually protein is synthesized from the DNA template of each gene. As well as, the portion of each gene which is introduced as mRNA is called as coding sequence of this gene. A gene expression level determines the amount of mRNA produced in a cell during protein synthesis; and is thought to be correlated with the amount of corresponding protein made. Expression levels are impacted by a huge number of environmental factors such as temperature, stress, light, and other signals, that lead to modify in the level of hormones and others [17].

It should be noted that a gene expression analysis gives details about dynamical changes in functionality of living beings. The assumptions that almost human diseases could be accompanied by certain changes in gene expressions has generated much interest among the Bioinformatics community in classification of patient samples based on gene expressions for disease diagnosis and treatment [16].

The latest advances in DNA Microarray introduce the ability to monitor and measure the expression levels of thousands of genes simultaneously in an organism. Several extentisve experiments were conducted which consist of monitoring each gene many times under different situations or evaluating each gene under a single environment with different types of tissues. The first one is beneficial for identification of functionally related genes while the second type of experiment is helpful in classification of various kinds of tissues and identification of those genes whose expression levels are good diagnostic indicators [16]. However, many of these genes are irrelevant to distinction of various samples and have non-positive influence on acquired classification accuracy [16].

The issue of gene identification is a combinatorial optimization issue comprising from two key objectives: i) minimizing as possible the number of selected genes and ii) maximizing the classification accuracy [16]. There is a much different from other functional optimizations that use the values of the functions as fitness, this issue has to something beyond these values. It could be the case that you obtain 100% accuracy on training data but 0% accuracy on test data. The Gene Ontology (GO) is a major bioinformatics initiative to unify the representation of gene and gene product attributes. The objectives of the Gene Ontology project are threefold:

1. Maintaining and further developing its controlled vocabulary of gene and gene product attributes.
2. Annotating genes and gene products, and assimilating and disseminating annotation data.
3. Supporting tools to facilitate access to all aspects of the data provided by the Gene Ontology project.

The GO ontology is a dynamic, and additions, corrections and alterations are suggested by, and solicited from, members of the research and annotation communities. For instance, an annotator could request a specific term to represent a metabolic pathway, or a section of the ontology may be revised with the help of community experts.

The GO ontology files are freely available for downloading from the GO website in a number of formats, or can be accessed online using the GO browser AmiGO. The Gene Ontology project also provides downloadable mappings of its terms to other classification systems.

Example GO Term

id: GO:0000016
name: lactase activity
namespace: molecular_function
def: "Catalysis of the reaction: lactose + H2O = D-glucose + D-galactose." [EC:3.2.1.108] synonym: "lactase-phlorizin hydrolase activity" BROAD [EC:3.2.1.108] synonym: "lactose galactohydrolase activity" EXACT [EC:3.2.1.108] xref: EC:3.2.1.108 xref: MetaCyc:LACTASE-RXN
xref: Reactome:20536 is_a: GO:0004553

Figures 1, 2 and 3 show a case study about Cluster-based network model. 6. Nodes (genes) from clusters 1 and 3 are represented by squares and circles, respectively. In the data and in the intersection networks, positive or negative relationships are represented by continuous or dashed edges, respectively.

As illustrated in Figure 1, 2 and 3, some relationships are supported by both data and the literature, and some are not. Note that, the identification of new relationships that are not mentioned in the current literature has its own merits in generating new biological hypotheses to be further tested in the laboratory. Another note is that we should not expect relationships found in the literature to be necessarily supported by the data model because the following reasons:

- Some of the relationships supported by the literature are related to different biological systems.
- The variation in the expression data can also limit our ability to detect true biological relationships.

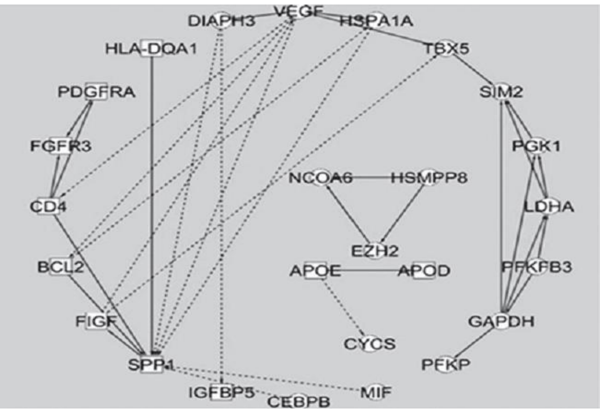

Figure 1: Data Network

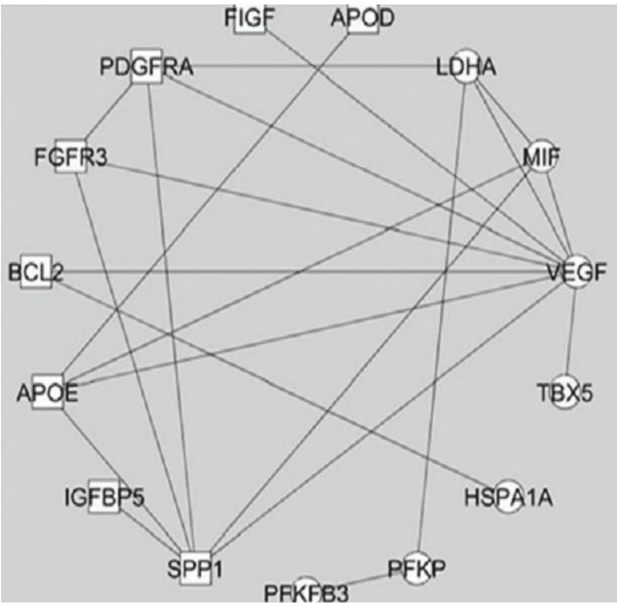

Figure 2: Literature Network

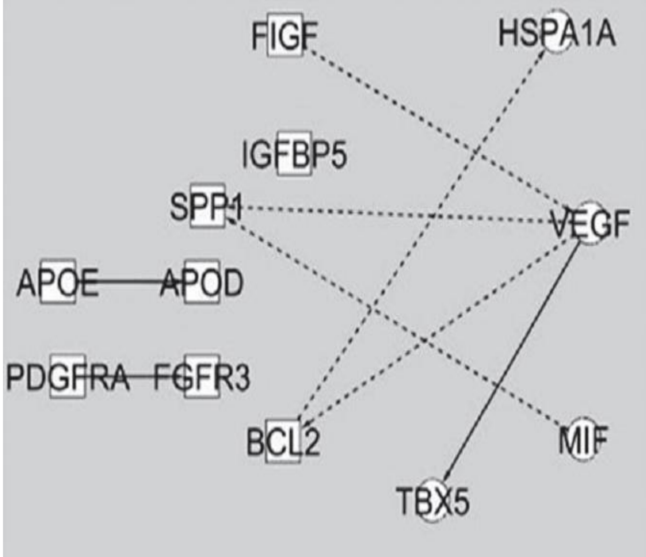

Figure 3: Data/ Literature Intersection Network

The rest of the paper is organized as follows. Section 2 presents the related work in this area. This section also includes a compaersion between the current approaches and methods. Finally, we have concluded the reasons, factors and perspectives that led to this initial research.

II. Rlated Work

In this section we have discussed the recent approaches and methods in this area. Table 1 summarizes the recent approaches and their techniques used, strengths and weaknesses.

Table 1: Compaersion between the current approaches and methods

Methods based on clustering algorithms	Clustering and visualization techniques	Parameter free clustering approach	Bi-clustering and ADP
Clustering methods: ○ Fuzzy C-means ○ Hierarchical clustering ○ K-means algorithm ○ Self-organizing maps ○ Singular value decomposition ○ Support vector machines **Parameters:** required.	**Tools:** ○ Avadis ○ BiNGOb ○ DAVID (Database for Annotation, Visualization and Integrated Discovery) **Parameters:** required.	**Aim:** This approach could lead to appropriate distribution of the given data instances into the most convenient clusters. **Parameters:** Not required.	**Methods:** ○ Bi-clustering ○ Association Pattern Discovery (APD) **Aim:** They adapted to find patterns of co-regulated genes.
Technique: o They have been used to cluster the samples into two or more classes depend on the number of available cancer samples. o k-means involves the number of clusters explicitly specified and approaches like DBSCAN are based on some parameters that implicitly simulate the number of clusters.	**Technique:** o Gene annotations are visualized in as a table view. Avadis has a built-in Gene Ontology browser to view ontology hierarchies. Genes can be clustered based on ontology terms to identify functional signatures in gene expression clusters. o BiNGO is a tool to determine which GO categories are statistically over-represented in a set of genes. o DAVID is a web-based tool that provides integrated solutions for the annotation and analysis of genome-scale datasets derived from high-throughput technologies such as microarray and proteomic platforms.	**Technique:** o Applied multi-objective genetic algorithm is applied to determine some alternative clustering solutions that constitute the pareto front. The result is pool of the clusters reported by all the solutions. Then, The homogeneity of each cluster is determined in the pool to keep the most homogeneous clusters. Finally, as a given data instance may belong to more than one cluster in the solution set this membership is reduced to the cluster in which the instance is closest to the centroid.	**Technique:** o In contrast to the described unsupervised and supervised techniques, these methods are able to discover co-regulated genes not only over the full set but also within and among subsets of conditions (samples). Moreover, each gene and each condition can occur in more than one cluster/pattern. While the idea of bi-clustering comes from the area of traditional clustering, namely to apply a similarity measure to calculate the correlation between cluster members, APD methods are inherited from the area of frequent itemset and association rule mining.
Drawbacks: o One of the main drawbacks of the clustering algorithms is requiring some parameters to guide the clustering process towards a certain solution which may not be necessary the most appropriate to the data	**Drawbacks**: o They can not disambiguate the exact position in pathways for homologous genes. o They do not take into account the reaction rate. o Despite the continuous improvement of visualization algorithms, most visualization	**Drawbacks:** o This approach is not applicable because it has not tested yet.	**Drawbacks**: o This method fails to identify similarly expressed genes whose expressions change between up- and down-regulation from one condition to another.

in hand. ● None of these methods provides formal inferences.	methods reach limitations in terms of user friendliness when thousands of nodes have to be analyzed and visualized.		

It should be noted that Kanehisa et al. [18] developed Kyoto Encyclopedia of Genes and Genomes (KEGG) which is a reference knowledge base for deciphering the genome. An experimental knowledge on systemic functions of the cell and the organism is represented in terms of molecular networks, and a mechanism (KEGG Orthology system) is developed for linking genes in the genome to nodes of the molecular network. In 1996, KEGG had been extended to fit the needs of both large projects and individual labs. In the latest years, the trend is to be focused on capturing and representing knowledge on diseases as perturbed states of the molecular network and drugs to the molecular network [19].

The most widely used method for cancer gene classification is gene expression profiles in terms of accuracy and reliability compared to traditional cancer diagnostic methods based mainly on the morphological appearance of the cancer [20]. A feature or gene selection algorithm using Bayes classification approach was introduced in [21], where the gene feature selection algorithms aim to classify genes that are crucial for accurate cancer classification and also endure biological significance.

The well known Support Vector Machine (SVM) method also is used for gene selection, cancer classification and functional gene classification. SVM classification techniques works well in term of accuracy if the number of training set were small compared to the large number of gene expression data in datasets, which is used as training set for classifying cancer genes selection[22]. A modified SVM were introduced in [23] based on selecting important genes features and building effective cancer classification.

In [24] the authors proposed used SVM for gene selection with a reject option, where gene expression data comprise of expression levels of several thousands of candidate genes. The gene selection procedure is necessary to provide a better understanding of the underlying biological system that generates data and to improve prediction performance [24].

The authors in [25] combined Partial Class Relevance (PCR) and Full Class Relevance (FCR) for cancer gene feature selection and classification, which seeks to reduce the training data while retaining semantic information of the microarray gene expression.

III. Conclusion and Future Work

To improve the bioinformatics, several researchers and practitioners have preferred to match the clustering to the specifications of biomedical applications. In this papered, we have reviewed a number of clustering algorithms for different approaches and data types. As a part of future work, the web semantic technologies will be proposed to assist in the gene classification. The proposed system might include RDFa to annotate for the attributes of genes and classes. In addition, the RDFa extractor and parser components will be used to link among the smellier genes in various classes.

References

[1] TAVAZOIE, S., HUGHES, J. D., CAMPBELL, M. J., CHO, R. J. AND CHURCH, G. M. (1999). Systematic determination of genetic network architecture. Nature Genetics 22, 281–285.

[2] TAMAYO, P., SLONIM, D., MESIROV, J., ZHU, Q., KITAREEWAN, S., DMITROVSKY, E., LANDER, E. S. AND

[3] GOLUB, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proceedings of the National Academy of Sciences of the United States of America 96, 2907–2912.

[4] WALL, M. E., DYCK, P. A. AND BRETTIN, T. S. (2001). Singular value decomposition analysis of microarray data. Bioinformatics 17, 566–568.

[5] BROWN, M. P. S., GRUNDY, W. N., LIN, D., CRISTIANINI, N., SUGNET, C., FUREY, T. S., ARES, M. AND

[6] HAUSSLER, D. (2000). Knowledge-based analysis of microarray gene expression data using support vector machines. Proceedings of the National Academy of Sciences of the United States of America 97, 262–267.

[7] Inoue L. Y., Neira M., Nelson C., Gleave M., and Etzioni R. Cluster-based network model for timecourse gene expression data. Biostatistics, 2007.

[8] Bansal, A.K. (2005): Bioinformatics in microbial biotechnology - A mini review. In: Microbial Cell Factories, 4, S. Article No. 19.

[9] Pavlopoulos, G., Wegener, A.L., Schneider, R.: A survey of visualization tools for biological network analysis. BioData Min 1(1) (Nov 2008) 12

[10] Carmona-Saez P., Chagoyen M., Rodriguez A., Trelles O., Carazo J.M., and Pascual-Montano A. Integrated analysis of gene expression by association rules discovery). BMC Bioinformatics, page 7:54, 2006.

[11] Georgii E., Richter L., Ruckert U., and Kramer S. Analyzing Microarray Data Using Quantitative Association Rules. Bioinformatics, pages ii123–ii129, 2005.

[12] Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhinik Z, Ben-Dork A, et al.: Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 2000, 406:536-540.

[13] Alshalalfa M., Özyer T., Alhajj R. Attractive Clustering Approach for Knowledge Discovery in Gene Expression Data, Proceedings of ICIT 2009, Jordan, 3 Jun 2009.

[14] Gene Ontology - Wikipedia, the free encyclopedia.html

[15] The Gene Ontology Consortium (Jan 2008). "The Gene Ontology project in 2008.". Nucleic acids research 36: D440–4.

[16] Topon KP. Gene expression based cancer classification using evolutionary and non-evolutionary methods. Technical Report No. 041105A1. Japan: Department of Frontier Informatics, The University of Tokyo; 2004.

[17] Schena, M., DNA Microarrays, New York, USA:Oxford University Press, 2000, pp. 3-4.

[18] Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. Nucleic Acids Res. 2012 Jan;40(Database issue):D109-14. Epub 2011 Nov 10.

[19] Kanehisa,M., Goto,S., Furumichi,M., Tanabe,M. and Hirakawa,M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res., 38, D355–D360.

[20] S. Santhosh Baboo, Sasikala S. A Survey on Data Mining Techniques for GeneSelection and Cancer Classification. International Journal of Computer Science and Information Security Vol. 8, No. 1, April 2010.

[21] Sharma, A. and K.K. Paliwal. A gene selection algorithm using Bayesian classification approach. American Journal of Applied Sciences., Vol 9 No 1 pp: 127-131, 2012

[22] J. Zhang, R. lee, Y. J. Wang, Support Vector Machine Classifications for Microarray Expression Data Set, Fifth International Conference on Computational Intelligence and Multimedia Applications, pp 67-71.2003.

[23] Wei Luo, Lipo Wang, Jingjing Sun, “Feature Selection for Cancer Classification Based on Support Vector Machine,” IEEE 978-0-7695-3571-5, 2009

[24] H.k. Choi, D. Yeo, S. Kwon, Y.i Kim, Gene selection and prediction for cancer classification using support vector machines with a reject option, Computational Statistics and Data Analysis, Volume 55, Issue 5, 1 May 2011, pp 1897-1908,

[25] R. Nakkeeran. Hybrid Approach of Data Mining Techniques, PCA, EDM and SVM for Cancer Gene Feature Selection and Classification. European Journal of Scientific Research, Vol.79 No.4 pp641-652.2012