Data Mining in Pathway Analysis for Gene Expression

Amani AlAjlan[1(✉)] and Ghada Badr[1,2]

[1] College of Computer and Information Sciences, King Saud University, Riyadh, Kingdom of Saudi Arabia
aalajlan@ksu.edu.sa, badrghada@hotmail.com
[2] IRI - The City of Scientific Research and Technological Applications, Alex, Egypt

Abstract. Single gene analysis looks to a single gene at a time and its relation to a specific phenotype such as cancer development. However, pathway analysis simplifies the analysis by focusing on group of genes at a time that involve in the same biological process. Pathway analysis has useful applications such as discovering diseases, diseases prevention and drug development. Different data mining approaches can be applied in pathway analysis. In this paper, we overview different pathway analysis techniques in analyzing gene expression and propose a classification for them. Pathway analysis can be classified into: detecting significant pathways and discovering new pathways. In addition, we summarize different data mining techniques that are used in pathway analysis.

Keywords: Pathway analysis · Gene expression · Clustering · Classification · Feature selection

1 Introduction

In 2014, statistics from American Cancer Society [6] stated that cancer is the second cause of death in the United States after the heart diseases. About one every four deaths are caused by cancer [6]. Recently, pathway analysis has lead to better cancer diagnosis and treatment.

Pathway is "a collection of genes that serves a particular function and/or genes that interact with other genes in a known biological process" [29]. In [18] they defined pathway as "a collection of genes that chemically act together in particular cellular or physiologic function". In [1] they defined it as "a series of actions among molecules in a cell that leads to a certain product or a change in a cell". There are three type of pathways: metabolic, gene regulation and signaling pathways. Metabolic pathways are series of chemical reactions in cells [1]. Gene regulation pathways regulate genes to be either active or inhibit [1]. Signaling pathways are series of actions in a cell to move signals from one part of the cell to another. In biological research, they classify genes in pathways to improve gene expression analysis [11] and simplify the analysis by looking to few groups of related genes (pathways) instead of looking to long lists of genes [15].

P. Perner (Ed.): ICDM 2015, LNAI 9165, pp. 69–77, 2015.
DOI: 10.1007/978-3-319-20910-4_6

Pathway analysis concerns with finding out which pathways are responsible for a certain phenotype or which pathways are significant under certain conditions [3]. In addition, pathway analysis is used to explain biological results and as a validation phase in computational research [15]. Khatri et al. [15] pointed out two advantages of using pathway analysis. First, reduce the complexity of analysis from thousands of genes to few hundreds of pathways. Second, identifying significant pathways is more meaningful than a list of different gene expression when comparing two samples such as normal and cancerous.

Pathway analysis has useful applications such as discovering disease occurrences by finding out the disrupted biological pathways. Another application is drugs development that aims to design a drug that target one or two disrupted pathways [1,28]. Moreover, researchers plan to use biological pathway approach to personalized patients treatment and drug development [1,28].

The paper is organized as follows. Section 2 overviews pathway databases. Gen expression, microarray and RNA-seq are presented in Sect. 3. Section 4 overviews pathway analysis techniques. Section 5 describes some miRNA analysis techniques. The conclusion is presented in Sect. 6.

2 Pathway Databases

Pathways are curated manually from biological experiments or automatically using text mining techniques [24]. Manual curation is more accurate and reliable.

There are 547 available biological pathways related resources [2]. For example, KEGG (Kyoto Encyclopedia of Genes and Genomes) database is the most popular pathways resource. It contains manually curated and inferred metabolic, signaling and disease pathways for over 650 organisms [7]. Reactome is another example for manually curated and inferred pathways database. It contains metabolic, signaling and disease pathways for human [7]. Also, BioCarta contains manually curated metabolic and signaling pathways for human and mouse [7].

3 Gene Expression

Genes control cells functions and all cells have the same genetic information. Genes are active or inactive (have different gene expression) according to a cell type and different conditions. Gene expression measures amount of mRNA produced in a cell [13] and gives the degree to which gene is active under different conditions.

3.1 Microarray vs RNA-seq

Sequencing techniques allow scientists to analyze tens of thousands of genes in parallel at any given time [12]. These technologies help us to understand diseases and provide better treatments [5]. Sequencing techniques start with using microarray technologies. Then, next generation sequencing was developed and it has a lot of sequencing that used in gene expression analysis such as RNA-seq.

Microarray is a small glass or plastic or silicon chip in which tens of thousands of DNA molecules (probes) are attached. Microarray is able to detect specific DNA molecules of interest. It works as follow: from two mRNA samples (a test sample and a control sample) cDNAs are obtained and labelled with fluorescent dyes and then hybridized on the surface of the chip. Then, the chips are scanned to read the signal intensity that is omitted from the labelled and hybridized targets [12,23].

RNA-seq is used to rapid profile mRNA expression of whole transcriptome [5,9]. It works as follow: small reads are aligned to an annotated reference mRNA. Then, the number of reads that aligned to one of different cDNAs are counted [4]. RNA-seq outperforms microarray in various aspects. First, the ability of detecting and identifying unknown genes and detecting differential expression levels that have not detected by microarray [5]. Second, it does not require specific probes or predefined transcriptome of interest [5]. Third, it increases specificity and sensitivity for detecting genes [5].

Gene expression usually presents by $i \times j$ matrix as in Fig. 1 where the rows represent expression pattern of genes and the columns represent different conditions such as different samples (normal vs cancer) or different time points [12,13]. In microarray, x_{ij} represents intensity level of hybridization of ith gene in a jth condition. While in RNA-seq, x_{ij} represents the number of reads of gene i observed in condition j.

$$\begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,j} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,j} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i,1} & x_{i,2} & \cdots & x_{i,j} \end{pmatrix}$$

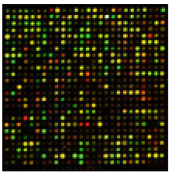
Fig. 1. Gene expression matrix

In microarray, chip is scanned to get hybridization data that are usually represent in a spreadsheet-like format [5] where each cell represents the intensity of hybridization of a specific gene in a specific condition as in Fig. 2. In RNA-seq, sequencing is used to get read counts that represent in spreadsheet-like format. Each cell represents the number of reads that aligned to one of thousands of different cDNAs [5] as in Fig. 3.

Having gene expression available a lot of analysis techniques can be applied. Pathway analysis is one of them that have impact on the development of drugs and disease diagnosis.

4 Classification of Pathway Analysis Techniques

Pathway analysis can be classified into two approaches: detecting significant pathways and discovering new pathways as in Fig. 4. Detecting significant



Gene	Sample1	Sample2	Sample3	Sample4
Gene1	0.6	4.4	1.3	1.0
Gene2	0.7	3.7	2.4	1.9
Gene3	1.5	2.6	5.2	0.8
Gene4	0.3	0.7	0.2	1.3
Gene5	0.5	3.4	3	0.5

Fig. 2. Microarray gene expression matrix

Gene	Sample1	Sample2	Sample3	Sample4
Gene1	7	10	7	9
Gene2	4	3	2	8
Gene3	1	1	2	6
Gene4	2	1	3	8
Gene5	5	4	3	9

Fig. 3. RNA-seq gene expression matrix

pathways approach aims to define and rank significant pathways that related to a specific phenotype either by enrichment score analysis or machine learning techniques [18]. Shin and Kim [23] classified the computational approaches for pathways analysis into three groups: clustering-based methods, gene-based methods and gene set-based methods. Clustering based methods are based on assumption that genes with similar expression would have similar functions or involved in the same biological processes [13,23]. Therefore, genes are clustered and pathways for each cluster are determined. In gene-based methods, differentially expressed genes DEGs between two samples (a test sample and a control sample) are identified, and then significant pathways that DEGs are involved are determined. In gene set-based methods the gene expression and a prior biological resource (i.e. pathway databases) are used to determine the significant pathways (gene sets) [23].

Discovering new pathways can be achieved either by mining the literature through text mining techniques or automatic inferring pathways from network interactions or gene expression data. In this paper, we focus on detecting significant pathways approaches and we categorize the research in the area according to the type of gene expression to be analyzed into two categories: pathway-based microarray analysis, and pathway-based RNA-seq analysis. Next we will review some research related to each category.

4.1 Pathway-Based Microarray Analysis

Most research are focused on analyzing microarray gene expression either to determine significant pathways that contribute to a phenotype of interest or deal with features (genes) selection problem. Next some research related to classification, feature selection and clustering approaches are reviewed.

Classification. It aims to define and rank significant pathways that related to a specific phenotype using machine learning approaches. Zhang et al. [29] used

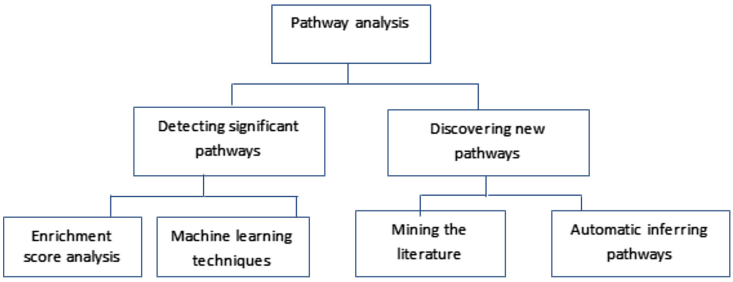


Fig. 4. classification of pathway analysis techniques

machine learning algorithms: nave bayes, support vector machine, decision tree and random forests to rank pathways based on classification error. By using three microarray expression datasets, they proved that machine learning algorithms outperform enrichment score analysis in identifying significant pathways. Pang et al. [20] used random forest classification and regression to analyze and rank pathways. In addition, they pointed out that their method was the first that used continuous measures for ranking pathways.

Features Selection. It aims to select informative genes within pathways before the pathway evaluation process to reduce computational time and improve accuracy [19]. Misman et al. [19] pointed out that when observing a particular biological context such as cancer some genes within pathways are only responsible for a phenotype. Thus, selecting subset of genes is important phase before ranking pathways. Zhang et al. [29] used minimum redundancy maximum relevance mRMR to select representative genes from each pathway. Panteris et al. [22] selected significant genes from each pathway (pathway signature) that describe the pathway at a given experimental condition. Misman et al. [19] used SVM-SCAD to select genes within pathways and have used B-type generalized approximate cross validation (BGACV) to select appropriate tuning parameter for SVM-SCAD. Jungjit et al. [14] proposed a KEGG pathway-based feature selection method for multi-label classification. Their method selects genes based on weighted formula that combines genes predictive accuracy and their occurrence in cancer-related KEGG pathways. Ibrahim et al. [11] selected strongly correlated genes for accurate disease classification by using pathways as prior knowledge. Their method was compared with five feature selection methods using two classifiers: K-nearest neighbour and support vector machine and it preformed the best for three microarray datasets.

Clustering. Detecting pathways in clustering analysis is used as a validation measure or as a partitioning measure. The reason of validation measure is to prove the validity of a clustering algorithm and for partitioning measure to partition datasets into biological meaningful clusters. For example, Shin and

Kim [23] used hierarchical clustering with Euclidean distance to generate gene clusters from gene expressions. Then, pathways are identified in each cluster to check the validity of clustering to identify the subclasses of leukemia. Zhao et al. [30] proposed a pathway-based clustering approach that used pathways to identify clusters. Their aim was to identify subgroups of cancer patients that may respond to the same treatment. Since cancers have similar phenotypes but resulting from different genetic mutations which lead to different responses to the same treatment. Their method is as follow: identify differential gene expression. Then, identify KEGG pathways that enriched with DGEs. Finally, classify the samples according to the expression of genes within the specified pathways. Also, Milone et al. [17] proposed a new method based on self-organizing map SOM clustering that used common metabolic pathways and Euclidean distance as similarity measures to construct clusters. Their objective was to improve the quality of clustering formation by combining pathway information. They used transcripts and metabolites datsets form Solanum lycopersicum and Arabidopsis thaliana species. Their method just improved the biological meaning of clusters compared with classical SOM. Moreover, Kozielski and Gruca [16] proposed a method that combined gene expression and gene ontology to identify clusters. So, the cluster membership should satisfy both gene expression and gene ontology. The proposed method is based on fuzzy clustering algorithm. Pang and Zhao [21] have proposed a method to generate pathways clusters that are related to a phenotype of interest from pathway- based classification [20]. They used class votes from random forest as similarity measure between pathways and tight clustering approach. Table 1 summarizes the research in pathway-based clustering and explains dataset, aim of clustering and aim of pathway analysis either validation or partitioning measure.

4.2 Pathway-Based RNA-seq Analysis

There are limited research focusing on analyzing RNA-seq gene expression to determine significant pathways that contribute to a phenotype of interest. Theses research focusing on statistical approaches. For example, Xiong et al. [27] developed a tool set that have multiple gene-level and gene set-level statistics to determine significant pathways. Fridley et al. [9] proposed using gamma method with soft truncation threshold to determine the gene sets that related to particular phenotype. Then, they applied the method to a smallpox vaccine immunogenetic study to identify gene sets or pathways with differential expression genes between high and low responders to the vaccine. Wang and Cairns [26] proposed combining differential expression with splicing information to detect significant gene sets based on Kolmogorov-Smirnov-like statistic. Xiong et al. [27] pointed out that the Wang and Cairns method is computationally expensive. Also, Hanzelmann et al. [10] developed a method that calculates variation of pathway activity profile over a sample population to analyze gene sets. Their method can be applied to RNA-seq as well as microarray data.

Table 1. Clustering and pathway analysis

Ref	Clustering algorithm	Dataset	Aim of clustering	Aim of pathway analysis
[23]	Hierarchical clustering	Leukemia	To identify subtypes of leukemia	Validation measure
[30]	Hierarchical clustering	NCI60 and DLBCL datasets	To identify subtypes of cancers	Partitioning measure
[17]	Self-organizing map SOM	Solanum lycopersicum and Arabidopsis thaliana	To improve the biological meaning of clusters	Partitioning measure
[21]	Tight clustering approach	Pathways from KEGG, BioCarta and Gen-Mapp	To generate pathways clusters that are related to a phenotype	———

5 Pathway-Based MiRNA Analysis

There are few research focusing on analyzing miRNA to determine significant pathways. Among them, Chen et al. [8] proposed a mathematical model (Bayesian implementation) that used miRNA targets for mapping miRNA to pathways then applied hypothesis test to extract significant pathways. Zhang et al. [28] used sample-matched miRNA and mRNA expression and pathway structure to analyze glioma patient survival. Wang et al. [25] suggested using functional information (gene ontology) to improve miRNA target prediction algorithms since genes that regulated by the same miRNA may share similar functions. Most miRNa target prediction algorithms used physical interaction mechanisms such as free energy, seed match and sequence conservation. Wang et al. [25] built SVM ensemble classifier that combined gene ontology and sequence information to predict miRNA targets.

6 Conclusion

Pathway analysis is reliable in discovering diseases and has various useful applications. In addition, data mining techniques are applied to pathway analysis to discover biological interesting hidden information. Most research in the field are based on analysis of microarray datsets and few are based on RNA-seq. Thus, applying data mining approaches such as classification and clustering to pathway-based RNA-seq analysis leads to more biological results.

References

1. Biological pathways fact sheet (2014). http://www.genome.gov/27530687. Accessed 11 August 2014
2. Pathguide (2015). http://www.pathguide.org/. Accessed 02 January 2015
3. Pathway analysis (2014). http://www.genexplain.com/pathway-analysis. Accessed 08 November 2014
4. Getting started with RNA-seq data analysis (2011). http://www.illumina.com/documents/products/datasheets/datasheet_rnaseq_analysis.pdf
5. Transitioning from microarrays to mRNA-seq, December 2011. http://www.illumina.com/content/dam/illumina-marketing/documents/icommunity/article_2011_12_ea_rna-seq.pdf
6. American cancer society: cancer facts and figures 2014 (2014)
7. Carugo, O., Eisenhaber, F.: Data Mining Techniques for the Life Sciences. Springer, New York (2010)
8. Chen, Y., Chen, H.I., Huang, Y.: Mapping miRNA regulation to functional gene sets. In: International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, IJCBS 2009, pp. 122–125. IEEE (2009)
9. Fridley, B.L., Jenkins, G.D., Grill, D.E., Kennedy, R.B., Poland, G.A., Oberg, A.L.: Soft truncation thresholding for gene set analysis of RNA-seq data: application to a vaccine study. Sci. Rep. **3**, 2898 (2013)
10. Hänzelmann, S., Castelo, R., Guinney, J.: GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinf. **14**(1), 7 (2013)
11. Ibrahim, M.H., Jassim, S., Cawthorne, M., Langlands, K.: Pathway-based gene selection for disease classification. In: 2011 International Conference on Information Society (i-Society), pp. 360–365. IEEE (2011)
12. Jiang, D., Tang, C., Zhang, A.: Cluster analysis for gene expression data: a survey. IEEE Trans. Knowl. Data Eng. **16**(11), 1370–1386 (2004)
13. Jones, N.C., Pevzner, P.: An Introduction to Bioinformatics Algorithms. MIT press, Cambridge (2004)
14. Jungjit, S., Michaelis, M., Freitas, A.A., Cinatl, J.: Extending multi-label feature selection with KEGG pathway information for microarray data analysis. In: 2014 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, pp. 1–8. IEEE (2014)
15. Khatri, P., Sirota, M., Butte, A.J.: Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput. Biol. **8**(2), e1002375 (2012)
16. Kozielski, M., Gruca, A.: Soft approach to identification of cohesive clusters in two gene representations. Procedia Comput. Sci. **35**, 281–289 (2014)
17. Milone, D.H., Stegmayer, G., López, M., Kamenetzky, L., Carrari, F.: Improving clustering with metabolic pathway data. BMC Bioinf. **15**(1), 101 (2014)
18. Misman, M., Deris, S., Hashim, S., Jumali, R., Mohamad, M.: Pathway-based microarray analysis for defining statistical significant phenotype-related pathways: a review of common approaches. In: International Conference on Information Management and Engineering, ICIME 2009, April 2009, pp. 496–500 (2009)
19. Misman, M.F., Mohamad, M.S., Deris, S., Abdullah, A., Hashim, S.Z.M.: An improved hybrid of SVM and SCAD for pathway analysis. Bioinformation **7**(4), 169 (2011)
20. Pang, H., Lin, A., Holford, M., Enerson, B.E., Lu, B., Lawton, M.P., Floyd, E., Zhao, H.: Pathway analysis using random forests classification and regression. Bioinformatics **22**(16), 2028–2036 (2006)

21. Pang, H., Zhao, H.: Building pathway clusters from random forests classification using class votes. BMC Bioinf. **9**(1), 87 (2008)
22. Panteris, E., Swift, S., Payne, A., Liu, X.: Mining pathway signatures from microarray data and relevant biological knowledge. J. Biomed. Inf. **40**(6), 698–706 (2007)
23. Shin, M., Kim, J.: Data mining and knowledge discovery in real life applications. In: Microarray Data Mining for Biological Pathway Analysis, pp. 319–336. I-Tech (2009)
24. Viswanathan, G.A., Seto, J., Patil, S., Nudelman, G., Sealfon, S.C.: Getting started in biological pathway construction and analysis. PLoS Comput. Biol. **4**(2), e16 (2008)
25. Wang, N., Wang, Y., Yang, Y., Shen, Y., Li, A.: miRNA target prediction based on gene ontology. In: 2013 Sixth International Symposium on Computational Intelligence and Design (ISCID), vol. 1, pp. 430–433. IEEE (2013)
26. Wang, X., Cairns, M.J.: Gene set enrichment analysis of RNA-seq data: integrating differential expression and splicing. BMC Bioinf. **14**(Suppl. 5), S16 (2013)
27. Xiong, Q., Mukherjee, S., Furey, T.S.: GSAASeqSP: a toolset for gene set association analysis of RNA-seq data. Sci. Rep. **4**, 6347 (2014)
28. Zhang, C., Li, C., Li, J., Han, J., Shang, D., Zhang, Y., Zhang, W., Yao, Q., Han, L., Xu, Y., Yan, W., Bao, Z., You, G., Jiang, T., Kang, C., Li, X.: Identification of miRNA-mediated core gene module for glioma patient prediction by integrating high-throughput miRNA, mRNA expression and pathway structure. PLoS ONE **9**(5), e96908 (2014)
29. Zhang, W., Emrich, S., Zeng, E.: A two-stage machine learning approach for pathway analysis. In: 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), December 2010, pp. 274–279 (2010)
30. Zhao, X., Zhong, S., Zuo, X., Lin, M., Qin, J., Luan, Y., Zhang, N., Liang, Y., Rao, S.: Pathway-based analysis of the hidden genetic heterogeneities in cancers. Genomics, Proteomics Bioinf. **12**(1), 31–38 (2014)