2009 International Conference on Information Management and Engineering

PATHWAY-BASED MICROARRAY ANALYSIS FOR DEFINING STATISTICAL SIGNIFICANT PHENOTYPE-RELATED PATHWAYS: A REVIEW OF COMMON APPROACHES

M.F. Misman[1] S. Deris[2] S.Z.M. Hashim[3] R. Jumali[4] M.S. Mohamad[5]

Artificial Intelligence and Bioinformatics Lab,
Department of Software Engineering, Faculty of Computer Science and Information System,
Universiti Teknologi Malaysia, Skudai, Johor, Malaysia
[1]faizmisman@gmail.com [2]safaai@utm.my [3]sitizaiton@utm.my [4]wanjumali@yahoo.com
[5]mohd.saberi@gmail.com

Abstract – **In this reiew, we have discussed about approaches in pathway based microarray analysis. Commonly, there are two approaches in pathway based analysis, Enrichment Score and Supervised Machine Learning. These pathway based approaches usually aim to statistically define significant pathways that related to phenotypes of interest. Firstly we discussed an overview of pathway based microarray analysis and its general flow processes in scoring the pathways, the methods applied in both approaches, advantages and limitations based on current researches, and pathways database used in pathway analysis. This review aim to provide better understanding about pathway based microarray analysis and its approaches.**

Keywords-pathway analysis, machine learning, microarray

I. INTRODUCTION

Over the past decade, many researchers focused on the development of techniques that looking into individual gene or one gene at a time, targeting for accurate identification and classification of differentially expressed genes between phenotype and their statistical significance [1, 2]. However, the problem for this single gene analysis lies not in the identification or classification of differentially expressed genes, but in their interpretation of biological meaning [1]. This is because; mostly subtle but coordinated differentially expressed genes cannot be detected as informative genes and usually will be terminated by the strict threshold of cutoff (feature selection) methods in single gene analysis [3, 4]. Moreover, looking at the individual gene at a time cannot provide the complete information about the biological processes such as cancer development because genes chemically act together. Beside that, there are many pathways in cancer development that consists of different genes in each pathway.

Pathway-based microarray analysis was designed to address these limitations of single gene analysis. It uses statistical methods to determine if predefined sets of genes in any pathways are differentially expressed in different phenotypes. Looking at set of genes rather than single gene at a time can bring more advantages in interpreting the biological knowledge from the DNA microarray data.

In this paper, we will be reviewed the current approaches in pathway based microarray analysis for statistically defining significant phenotype-related pathways based on several papers [4, 14, 22]. Aims to give the better understanding on approaches and methods in pathway based analysis, this paper divided into four topics, first topic will be discussed generally about pathway based microarray analysis, framework in scoring the pathways, and issues concern in pathway based analysis. Enrichment Score, one the approaches in pathway based microarray analysis will be discussed. In this part, current methods by several authors with its advantages and limitations classified in certain criteria, pathways database used in research area, and general issues in these approaches will be reviewed. Part three will be reviewed on supervised machine learning approaches and its current methods, classified into single classifier and ensemble classifiers. Part four, the discussion and summary for this paper.

II. OVERVIEW OF PATHWAY BASED MICROARRAY ANALYSIS

Pathways consist of genes that chemically act together in particular cellular or physiologic function [5]. There are consists of two popular types of pathways in genomic studies, metabolic pathways [6, 7] and signaling pathways [8]. Metabolic pathways are biological networks that involve enzymatic catalysis while the signaling pathways are a series of specific actions in a cell in which a signal is passed from one molecule to the next in the series.

Pathway-based microarray analysis is one of the approaches in microarray gene expression analysis. This pathway analysis method integrates the gene expression data with their annotation data such as metabolic

978-0-7695-3595-1/09 $25.00 © 2009 IEEE
DOI 10.1109/ICIME.2009.103

pathways and ontology functional classification rather than on identifying the significant changes of individual gene expression done by single gene analysis [9, 10, 11].

Pathway-based approaches aim to define the biological processes meaning through the finding of significant pathways and the gene member in the pathways using statistical evaluation contrast to single gene analysis that usually used univariate statistical tests that neglect the collaboration between genes [12, 13]. Moreover, pathway-based analysis approaches can detect subtle and coordinated changes in expression level of a group of genes in a pathway or with related function that usually single gene analysis cannot detect [4, 14].

Generally in pathway analysis, each pathway will be ranked based on the score obtained from statistical methods (figure 1). The highest score will be given to the pathway which had most relevant genes to related phenotypes. There are two methods in scoring pathways, the enrichment analysis and machine learning approaches [14]. Although these two methods have different processes but usually came out with the same output, the differentially expressed pathways and phenotype-relevant pathways [14].

There are several issues concerned by several authors in pathway-based analysis such as the quality of the pathways, since the pathways data are usually taken from the literature or other resources, non-relevant genes maybe included, or relevant genes maybe excluded from the pathways [14]. Several researches attempted to minimize these misspecification by defining signature genes to represent pathway behavior [15], refining pathways to adapt to specific conditions by removing unaltered genes from the dataset [16, 17, 18], and improving the functional interpretation of gene groups by including additional information associated with the group [19].

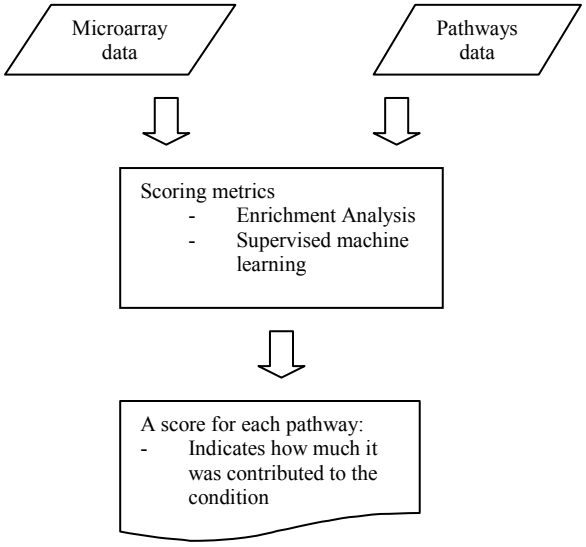

Figure 1: General process in scoring the pathways

III. ENRICHMENT ANALYSIS APPROACHES

Enrichment analysis (EA) is the common approaches in pathway based analysis. Generally, the gene expression profiles are viewed as functional pathways, and significant pathways are the pathways with the large number of differentially expressed genes [14]. Mootha et al. [12] firstly used Gene Set Enrichment Analysis (GSEA) discovered the alteration of genes of oxidative phosphorylation pathways and other metabolic pathways in human diabetes disease. Even before GSEA by Mootha et al. [12], a few researches already applied the concept of integrating the microarray data with their annotation such by Virtaneva et al. [20] calculated scores for the pathways curated from SWISS-PROT [21] database and applied sample randomization to evaluate the significance of each category.

Thereafter, various EA methods have been developed (Table 1) based on different null hypothesis and statistical methods and only GSEA are in both classes. EA are classified into two kinds of null hypothesis, competitive null hypothesis and self-contained null hypothesis [4, 22]. Detailed interpretation about null hypothesis can be obtained from [22].

Table 1: methods in Enrichment Analysis classified by null hypothesis

Null hypothesis	Methods	Author(s)	Statistical test
Self-contained	GSEA	[12]	Kolmogorov-Smirnov, sample randomization
		[28]	
	Globaltest	[23]	Sample randomization
	SAM-GS	[24]	Sample randomization
Competitive	Catmap	[25]	Gene randomization
	JProGO	[26]	Fisher's exact test, hypergeometric test
	GeneTrail	. [27]	Fisher's exact test, hypergeometric test
	GSEA	[12]	Kolmogorov-Smirnov, sample randomization
		[28]	

Nam and Kim [4] also have studied the comparison between this two null hypotheses including GSEA mixed null hypothesis. The experiment consists of 2000 genes expression data divided into two classes where each class has 20 samples. The expression values were sampled from a standard normal distribution in both groups. For 600 randomly selected genes, a random value between 0.5 and 1 are added to the second group to generate differentially expressed genes. After that, the genes were

divided into 100 gene sets where each gene set contained 20 genes. The average *t*-statistic score function was used in order to differentiate these three hypotheses. In this experiment, it is expected that no genes were enriched with differentially expressed genes.

As a result, the competitive null recognized no differentially expressed gene sets, so the *p*-values were distributed uniformly. It is differ to self-contained method, where this method detected about 83% gene sets as differentially expresses with a *p*-value cutoff of 0.05. The mixed approach, showed an intermediate performance. Nam and Kim. [4] have conformed that if the purpose of the research is to find gene sets relatively enriched with differentially expressed gene, a competitive method should be used. While if the purpose is to find gene sets clearly separated between the two sample groups, a self-contained method can be use. Nam and Kim. [4] prefer to use the mixed method, in order to avoid the drawbacks of each method.

GeneTrail by [27] provided two statistical methods, Over-Representation Analysis (ORA) that comparing reference set of genes to a test set, and GSEA that scoring sorted lists of genes. Besides that, GeneTrail improved the calculation of *p*-value for each gene from GSEA by using dynamic-programming algorithm. Using pathways data from KEGG [29], TRANSPATH [30], TRANSFAC transcription factors [31], protein-protein interaction data from DIP [32], MINT [33], HPRD [34], and InAct [35] databases as an annotation data. Catmap used same methodologies as GSEA but it used Wilcoxon rank sum to improve the calculation of the *p*-value from GSEA and using Gene Ontology [36] database as an annotation data. Using three well established statistical methods of the threshold value-based Fisher's exact test and threshold value-independent Kolmogorov_smirnov and *t*-test, JProGo was implemented for the functional intrepetation of high-throughput gene expression data based on the identification of Gene Ontology (GO) nodes.

Basically, almost all of the current methods in EA are enhanced from GSEA. This is because; classic GSEA has certain limitations such as classic GSEA cannot handle more than two classes [1, 37]. Furthermore, Dragichi et al. [16] has proved that current GSEA considering that all the genes in the pathway equally important is inaccurate. It is showed in their research where including information on the topology or position of the differentially expressed genes in the pathway helped to identify pathways that may be relevant to lung cancer but were otherwise missed [14]. Dinu et al. [24] has showed in their research that GSEA can give statistical significance to gene sets that have no genes with moderately or strongly associated with the phenotype. Moreover, EA evaluates one pathway at a time and this leads to the neglecting of pathway interdependences that contribute to changes in the phenotypes [14, 38].

IV. SUPERVISED MACHINE LEARNING APPROACHES

Contrast to EA, supervised machine learning can evaluate multiple pathways simultaneously, and thus it could account pathway interdependence [13, 38]. In supervised machine learning approaches, significance pathways are the pathways that can improve the prediction of the phenotype [14]. Although machine learning approaches are not a popular approach compared to EA, but there are several authors such as Tomfohr et al. [41], Tai and Pan. [42, 43], Wei and Li. [38], Luan and Li. [44], and Pang et al. [13] that used machine learning approach in pathway based analysis to define significant pathways that related to phenotypes. These supervised machine learning approaches used pathways as input variables, classified into two models, single classifier model and ensemble classifier model (Table 2).

In single classifier model, certain preprocessing methods have been applied in order to determine pathway activities by evaluating the gene expressions in the pathways [14]. Such as singular value decomposition-based method by Tomfohr et al. [51] used first principle component to represent pathway, centroids of the gene expressions in the pathways as preprocessing in shrunken centroid-based methods by Tai and Pan. [41, 42]. Utilization of preprocessing methods in single classifier model can led to lost of some informative genes are something that need to be concern.

Table 1: Methods in supervised machine learning approaches classified into single and ensemble classifier

Model	Methods	Author(s)
Single classifier	Singular value decomposition method	[51]
	Discriminant analysis	[41]
	Partial least square regression	[42]
Ensemble classifier	Non-parametric machine learning	[38] [43]
	Random Forest	[13]

The goal of ensemble learning methods is to construct a collection (an ensemble) of individual classifiers that are diverse and yet accurate, and if this can be achieved, then highly accurate classification decisions can be obtained by voting the decisions of the individual classifiers in the ensemble [44]. Many authors have demonstrated significant performance improvements through ensemble methods [45-48]. Three of the currently most popular techniques for constructing ensembles are bootstrap aggregation [49], the Adaboost family of algorithms [50], and Random Forest the ensemble of classification trees [39].

In pathway analysis, this ensemble model can surmount the drawback from the single classifier models in term of losing informative genes by the preprocessing methods. A nonparametric pathway-based regression based [38, 43] use expression of the genes in pathways to characterize the activity of the pathways, and the activity level are regressed to the phenotype to form a predictive model [14]. Pang et al. [13] use Random Forest classification and regression methods to define the significant genes and pathways that related to the phenotypes.

V. DISCUSSION AND CONCLUSION

Pathway based microarray analyses are one of the method that are becoming popular nowadays. Ranking pathways are relevant to a particular phenotype; it can help researchers focus on a few sets of genes. They are particularly useful for generating further biological hypotheses of interest. In addition, pathway analysis proved that it can identify more subtle changes in expression than the gene lists that result from univariate statistical analysis. A review paper describing the advantages of performing pathway-based tests has been published [40]. Furthermore, as pathways are functional subunits of the cellular systems, looking at them may improve the ability to tease out biologically meaningful information from microarray data.

These considerations have motivated various research groups to look at gene sets or pathways rather than single genes. In addition, pathway analysis proved that it can identify more subtle changes in expression than the gene lists that result from univariate statistical analysis.

This review has discussed the approaches in pathway based microarray analysis and methods applied. In Enrichment Score topics, we have also covered the topic about null hypothesis based on research done by [4]. Classified the methods into this null hypothesis can help to facilitate the understanding of this approaches. For the supervised machine learning approach, we also classified the recent methods into two models, single classifier and ensemble classifiers.

ACKNOWLEDGEMENT

This work was supported by Ministry of Science, Technology and Innovation through Research Management Center, UTM. By research grant vot 79228. Also special thanks to our project leader, Prof. Dr. Safaai Deris for his full support.

REFERENCES

[1] Tian, L., Greenberg, SA., Kong, SW., Altschuler, J., Kohane, IS., Park, PJ. (2005). Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America*. 102(3): 13544-13549.

[2] Speed , T. Statistical Analysis of Gene Expression Microarray Data. London: Chapman and Hall. 2003.

[3] Shi, J. and Walker, MG. (2007). Gene Set Enrichment Analysis (GSEA) for Interpreting Gene Expression Profiles. *Current Bioinformatics.* 2:000-000.

[4] Nam, D, and Kim, S.Y. (2007). Gene-set Approach for Expression Pattern Analysis. *Briefings in Bioinformatics.* 9(3): 450.

[5] Kurhekar. M.P., Adak, S., Jhunjhunwala, S., Ragupathy, K. Genome-Wide pathway analysis and visualization using gene expression data. *Pacific Symposium on Biocomputing.* January 3-7. Lihue, Hawaii. 462-473.

[6] Greenberg, D.M. Metabolic pathways. 3rd. ed. New York.: Academic Press; 1975.

[7] Horton, H.R., Moran, L.A., Ochs, R.S., Rawn, J.D., Scrimgeour, K.G. Principles of Biochemistry. 3rd ed. Englewood Cliff, N.J: Prentice-Hall. 1996.

[8] Krauss, G. Biochemistry of signal transduction and regulation. 3rd. ed. Weinheim: Wiley-VCH.: 2008.

[9] Dinu, L., Zhao, H., Miller, PL. (2007). Integrating domain knowledge with statistical and data mining methods for high-density genomic SNP disease association analysis. *Journal of biomedical informatics.* 40(6): 750-760.

[10] Manoli, T., et al. (2006). Group testing for pathway analysis improves comparability of different microarray datasets. *Bioinformatics*. 22, 2500–2506.

[11] Mansmann,U. and Meister, R. (2005). Testing differential gene expression in functional groups: Goemano´ s global test versus an ANCOVA approach. *Methods of Inf. Med*. 44, 449–453.

[12] Mootha,V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D., Groop, L.C. (2003). PGC-1 Alpha-Responsive Genes Involved in Oxidative phosphorylation are Coordinately Downregulated in Human Diabetes. *Nature Genetic*. 34(3): 267–273.

[13] Pang, H, and Zhao, H. (2008). Building Pathway Clusters from Random Forest Classification Using Class Votes. *BMC Bioinformatics.* 9(87).

[14] Wang, X., Dalkic, E., Wu, M., Chan, C. (2008). Gene Module Level Analysis: Identification to Networks and Dynamics. *Current Opinion in Biotechnology.* 19(5): 482-491.

[15] Panteris, E., Swift, S., Payne, A., Liu, X. (2007). Mining Pathway Signatures from Microarray Data and Relevant Biological Knowledge. *Journal of Biomedical Informatics.* 40 (6): 698-706.

[16] Dragichi, S., Khatri, P., Tarca, A.L., Amin, K., Done, A., Voichita, C., Georgescu, C., Romero, R. (2007). A Systen Biology Approach for Pathway Level Analysis. *Genome Research.* 17(10): 1537-1545.

[17] Novak, B.A., Jain, A.N. (2006). Pathway Recognition and Augmentation by Computational Analysis of Microarray Expression Data. *Bioinformatics.* 22(2): 233-241.

[18] Gat-Viks, I., Shamir, R. (2007). Refinement and Expansion of Signaling Pathways: The Osmotic Response Network in Yeast. *Genome Research.* 17(3): 358-367.

[19] Hummel, M., Meister, R., Mansmann, U. (2008). GlobalANCOVA: Exploration and Assessment of Gene Group Effects. *Bioinformatics.* 24(1): 78-85.

[20] Virtaneva, K., Wright, F.A., Tanner, S.M., Yuan, B., Lemon, W.J., Caligiuri, M.A., Bloomfield, C.D., De La Chapelle, A., Krahe, R. (2001). Expression Profiling Reveals Fundamental Biological Differences in Acute Myeloid Leukemia With Isolated Trisomy 8 and Normal Cytogenetics. *Proceedings of the National Academy of Sciences of the United States of America.* 98(3): 1124-1129.

[21] Bairoch, A., Boeckmann, B. (1991). The SWISS-PROT Protein Sequnece Data Bank. *Nucleic Acids Research.* 19(SUPPL.): 2247-2249.

[22] Goeman, JJ. and Buhlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*. 23(8): 980-987.

[23] Goeman, J.J., Van de Geer, S., De Kort, F., Van Houwellingen, H.C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*. 20(1): 93-99.

[24] Dinu, I., Potter, J.D., Mueller, T., Liu, Q., Adewale, A., Jhangri, G.S., Einecke, G., Famulski, K.S., Halloran, P., Yasui, Y. (2007). Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*. 8(242).

[25] Breslin, T,. et al. (2004). Comparing functional annotation analyses with catmap. *BMC Bioinformatics*, 5, 193.

[26] Scheer, M., Klawonn, F., Munch, R., Grote, A., Hiller, K., Choi, C., Koch, I., Schobert, M., Hartlig, E., Klages, U., Jahn, D. (2006). JProGO: A Novel Tool for the Functional Interpretation of Prokaryotic Microarray Data Using Gene Ontology Information. *Nucleic Acids Research* 34 (WEB. SERV. ISS.).: W510-W515.

[27] Backes, C., Keller, A., Kuentzer, J., Kneissl, B., Comtesse, N., Elnakady, Y.A., Muller, R., Meese, E., Lenhof, H.P. (2007). GeneTrail-Advanced Gene Set Enrichment Analysis. *Nucleic acids research.* 35 (Web Server issue): W186-192.

[28]Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P. (2005). Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. *Proceedings of the National Academy of Sciences of the United States of America.* 102(43): 15545-15550.

[29] Kaneisha, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M. (2006). From Genomics to Chemical Genomics: New Developments in KEGG. *Nucleic acids research.* 34 (Database issue): D354-357.

[30] Krull, M., Pistor, S., Voss, N., Kel, A., Reuter, I., Kronenberg, D., Michael, H., Schwarzer, K., Potapov, A., Choi, C., Kel-Margoulis, O., Wingender, E. (2006). TRANSPATH: An Information Resource for Storing and Visualizing Signaling Pathways and Their Pathological Aberrations. *Nucleic acids research.* 34 (Database issue): D546-551.

[31] Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A.E., Wingender, E. (2006). TRANSFAC and its Module TRANSCompel: Transcriptional Gene Regulation in Eukaryotes. *Nucleic acids research.* 34 (Database issue): D108-110.

[32] Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 Update. *Nucleic Acids Research.* 32 (DATABASE ISS.).: D449-D451.

[33] Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., Cesareni, G. (2002). MINT: A Molecular INTeraction Database. *FEBS Letters.* 513 (1): 135-140.

[34] Peri, S.,Navarro, J.D., Amanchy, R., Kristiansen, T.Z., Jonnalagadda, C.K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T.K.B., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H.N., Rashmi, B.P., Ramya, M.A., Zhao, Z., Chandrika, K.N., Padma, N., Harsha, H.C., Yatish, A.J., Kavitha, M.P., Menezes, M., Choudhury, D.R., Suresh, S., Ghosh, N., Saravana, R., Chandran, S., Krishna, S., Joy, M., Anand, S.K., Madavan, V., Joseph, A., Wong, G.W., Schiemann, W.P., Constantinescu, S.N., Huang, L., Khosravi-Far, R., Steen, H., Tewari, M., Ghaffari, S., Blobe, G.C., Dang, C.V., Garcia, J.G.N., Pevsner, J., Jensen, O.N., Roepstorff, P., Deshpande, K.S., Chinnaiyan, A.M., Hamosh, A., Chakravarti, A., Pandey, A. (2003). Development of Human Protein Reference Database as an Initial Platform for Approaching Systems Biology in Humans. *Genome Research.* 13 (10).: 2363-2371.

[35] Kerrien, S.Y., Alam-Faruque, B., Aranda, I., Bancarz, A., Bridge, C., Derow, E., Dimmer, M., Feuermann, A., Friedrichsen, R., Huntley, C., Kohler, J., Khadake, C., Leroy, A., Liban, C., Lieftink, L., Montecchi-Palazzi, S., Orchard, J., Risse, K., Robbe, B., Roechert, D., Thorneycroft, Y., Zhang, R., Apweiler., Hermjakob, H. (2006). IntAct - Open Source Resource for Molecular Interaction Data. *nucleic Acid Research*. 32(Database issue): D561-D565.

[36] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G. (2000). Gene Ontology: Tool for the Unification of Biology. *Nature Genetics.* 25 (1).: 25-29.

[37] Edelman, E., Porello, A., Guinny, J., Balakumaran, B., Bild, A., Febbo, P.G., Mukherjee, S. (2006). Analysis of Sample Set Enrichment Scores: Assaying the Enrichment of Sets of Genes for Individual Samples in Genome-Wide Expression Profiles. *Bioinformatics.* 22: e108-e116.

[38] Wei, Z and Li, H. (2006). Nonparametric Pathway-Based Regression Models for Analysis of Genomic Data. *biostatistics.* 8(2): 265-284.

[39] Breiman, L. (2006). Random Forests. *Machine Learning.* 45(1): 5-32.

[40] Curtis, RK., et al. (2005). Pathways to the analysis of microarray data. *Trends Biotechnol*. 23, 429–435. [44] Tomfohr, J., Lu, J., Kepler, T.B. (2005). Pathway Level Analysis of Gene Expression using Singular Value Decomposition. *BMC Bioinformatics.* 6: 225.

[41] Tai, F. and Pan, W. (2007). Incorporating Prior Knowledge of Gene Functional Groups into Regularized Discriminant Analysis of Microarray Data. *Bioinformatics.* 23(23).: 3170-3177.

[42] Tai, F. and Pan, W. (2007). Incorporating Prior Knowledge of Predictors into Penalized Classifiers with Multiple Penalty Terms. *Bioinformatics.* 23(14).: 1775-1782.

[43] Luan, Y., Li, H. (2008). Group Additive Regression Models for Genomic Data Analysis. *Biostatistics.* 9(1): 100-113.

[44] Dietterich, T.G. (2000). An Experiments Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning.* 40(2): 139-157.

[45] Breiman, L. (1996b). Bias, Variance, and Arcing Classifiers. Technical Report 460, Department of Statistics, University of California, Berkeley, CA.

[46] Kohavi, R, and Kunz, C. Option Decision Trees with Majority Votes. *Proceedings of the Fourteenth International Conference on Machine Learning.* July 8-12. San Francisco, California: Morgan Kaufman. 2007. 161-169.

[47] Bauer, E, and Kohavi, R. (1999). An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting and Variants. *Machine Learning.* 36(1/2): 105–139.

[48] Maclin, R, and Opitz, D. An Empirical Evaluation of Bagging and Boosting. *Proceeding of the Fourteenth International Conference on Machine Learning.* July 8-12. ambridge, MA: AAAI Press/MIT Press. 1997. 546-551.

[49] Breiman, L. (1996a). Bagging Predictors. *Machine Learning.* 24(2): 123–140.

[50] Freund, Y and Schapire, R.E. Experiments with a New Boosting Algorithm. *Proceeding of 13th International Conference on Machine Learning.* July 3-6. Morgan Kaufman. 1996. 148-146.