2009 International Conference on Information Management and Engineering

A Study of Network-based Approach for Cancer Classification

R. Jumali[1] S. Deris[2] S.Z.M. Hashim[3] M.F. Misman[4] M.S. Mohamad[5]
Department of Software Engineering, Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia
Skudai, Johor, Malaysia
[1]wanjumali@yahoo.com [2] safaai@utm.my [3]sitizaiton@utm.my [4]faizmisman@gmail.com [5]mohd.saberi@gmail.com

Abstract—**The advent of high-throughput techniques such as microarray data enabled researchers to elucidate process in a cell that fruitfully useful for pathological and medical. For such opportunities, microarray gene expression data have been explored and applied for various types of studies e.g. gene association, gene classification and construction of gene network. Unfortunately, since gene expression data naturally have a few of samples and thousands of genes, this leads to a biological and technical problems. Thus, the availability of artificial intelligence techniques couples with statistical methods can give promising results for addressing the problems. These approaches derive two well known methods: supervised and unsupervised. Whenever possible, these two superior methods can work well in classification and clustering in term of class discovery and class prediction. Significantly, in this paper we will review the benefit of network-based in term of interaction data for classification in identification of class cancer.**

Keywords-classification; DNA microarray data; interaction gene

I. INTRODUCTION

Cancer that begins from normal cell will become tumor and spreading rapidly in human body can extremely threat human life without early diagnosis. Although there are many treatments in our advanced medical, they are very costly and affect another side effect. Thus, research in bioinformatics is beneficial to address this problem. Actually, bioinformatics is a combination of biology, computer science, mathematics and statistics. Cancer occurs once the cell regulation goes wrong [1]. This occurs from replication of certain genes in the change of the expression values that mutate the DNA.

With emergence of microarray data, bioinformatics field has been grown up and focused on many parts of research such as classification [2], clustering and others whereby each one produced promising results for cancer identification. The nature of this biological data (microarray) leads to many problems that need researchers to undergoing many more research for problems such as high dimensionality, noise and irrelevant gene.

Statistical techniques coupled with artificial intelligence methods especially machine learning has been adapted to solve problems in bioinformatics. It derives two main approaches; supervised learning and unsupervised learning meanwhile the other approaches include semi-supervised learning, reinforcement learning, transduction and learning to learn.

Unsupervised learning is an agent which models a set of inputs without learner. For the purpose of cancer identification, clustering method which is one of unsupervised learning approaches has been used to discover the type class cancer in cluster form.

Instead, classification was used for the same task as well. Differ to clustering in unsupervised approach, supervised learning classification require learner to build classifier and regarded as training and test set. Many approaches has been introduced and used for classification and each has their own strengths and weaknesses. For the next section, details regarding classification include problems, methods used, trends and arisen issues will be discussed in term of microarray gene expression data.

II. DNA MICROARRAY DATA

Microarray technology such as DNA microarrays enabled us to study the behavior of the cell globally [3] rather than using microscope [4]. With this high throughput technique, biologists are able to measure the expression levels of thousands of genes in a single experiment [5-11] and used as a diagnostic tool also stimulating the discovery of new target for the treatment of diseases. This ability is also known as microarray analysis or gene expression profiling (molecular signature) and this study also referred as genomic.

Basically, DNA microarray data also known as gene expression data derived from samples tissues or blood to cDNA (spotted array) by hybridization of mRNA as well as from hybridization of oligonucleotide of DNA (Affymetrix chips). Gene expression is the information within a cell that represents the transcription process of a gene (DNA Sequence) into mRNA and ultimately into a functional gene product (protein or RNA) through translation process. Actually, microarray gene expression data play an important role for gene association studies, predictive modeling and reconstruction of gene networks [12] and deal with three major approaches: network-based approaches, expression-based approaches and prior pathways-based approaches [13].

978-0-7695-3595-1/09 $25.00 © 2009 IEEE
DOI 10.1109/ICIME.2009.104

According to [14], to find the functionality of genes related to cancer, most researchers focusing on three types of area based on gene expression profiles. First, goes to select informative features in order to reduce the dimensionality. Second, learning classifier that significant to construct a robust classifier which promisingly can achieves high accuracy. The third is knowledge discovery to understand the relationships among genes. This chapter focuses on the second one.

III. Cancer Classification

As a basis, classification of microarray data aims to construct efficient classifier (model) that derived from informative genes into one of the diagnostic categories, for example tumor/normal tissue, or benign/malignant tumor [15] to make prediction for an unknown patient [11,16]. As stated by [17], there are two types of classification: class discovery and class prediction. Usually, classification [18] in term of class prediction has been focused and acts as a promising tool to assign a label to a given set of features correctly [6, 19] that useful for biomarker identification and disease-related genes [20].

For such purpose, many techniques were proposed and applied such as Bayesian networks, Boolean networks, *k*-NN, neural networks, support vector machines [11, 21]. Although these methods can successfully perform the classification task but the results that precisely give the high accuracy in classification still becoming issues.

Based on study undergone by [7], they have summarized four issues that lead to problems in classification. First, goes to the unique nature of microarray gene expression data. Most existing classification methods were not capable to handle sparseness of data and high dimensionality [15]. These cause to overfitting problems and consequently affect the computational time of classification process. Second, regarding to noise which is consists of biological and technical noise. Biological noise here means genes that are not useful for identification of cancer class meanwhile technical noise are associated with the non-uniform genetic backgrounds of the samples that detected at data preparation stages as well as refers to misclassification of the samples. Third problem involves dealing with a numerous irrelevant genes compared to relevant genes. Fourth and finally, since the goal of bioinformatics is to elucidate the process inside a cell, it is necessity for researchers to take into account the application domain in cancer classification study.

IV. Approaches in Classification

In this study, since the classification using microarray gene expression data that poses a large number of records (genes) and less number of fields (samples) as described above, researchers are facing challenge in discover new robust method in creating a high likelihood of finding false positives due to chance for building predictive models and finding differentially expressed genes [4]. Thus, the techniques that can reduce the dimensions are needed to select informative genes because most of genes are irrelevant [22].

There are two approaches introduced to address this dimension reduction problem: feature selection [23] and feature extraction [24]. Feature extraction comes up with combination of information from identified small number of dimensions [15].

Meanwhile, feature selection that consisting three approaches was frequently used instead of feature extraction because it is depends on the problems to be solved. The approaches include filter approach, wrapper approach and hybrid approach [22, 25]. Actually, the selection of variable can be done through a subset with aggregate discriminative power or ranked for their individual relevance [26]. [15] mentioned two ranking approaches based on their popularity. First, the approach that uses univariate ranking according to their informativeness. Second, gene ranking based on Markov blanket filtering was commonly used.

Although they have been given promising results, these approaches known as single gene based method only focuses on expression level [27] due to limitation of data and computational techniques constraint. In fact, inspired from gene regulation in a cell that one gene has to interact each other to function, interaction of groups of genes suppose to be considered.

V. Network-based Approach for Classification

Rather than for class prediction, as clinicians demanding on interpretation of results for biological, the new era of classification considering biological function of each informative genes that residing in the microarray data in terms of drug targets. Towards establishment of system biology, analysis of microarray classification focuses on specific systems as an interaction such as genes, with respect to phenotype, such as particular cancer [28]. For this purpose, it is worth to extract the huge amount of information from genome-wide analysis but it is a challenge task [29].

Even for task such as clustering also take for granted the component interaction. With the availability of microarray gene expression data, it is can possibly predict gene-gene interaction accurately not only for clustering and classification but reconstruction of biological gene network as in [30-31].

With the existence of many computational methods that facilitate research in microarray data, gene networks that organized from components such as genes, proteins and other molecules [32] can be associated with it in producing better result especially for understanding process inside a cell. In cancer classification perspective, we are not intend to construct gene network even though supervised classification can works for it as in [19, 33, 34] but studying the information gained from interaction between two genes.

Previous studies had proven that precise of classification accuracy achieved by prior knowledge rather than single gene based. The significance of interaction genes towards classification compared to single gene-based is shown in Table I.

Emphasized again, network-based here means that interactions among genes is taken into account [35-37] for informative gene selection in term of building classifier. It is because disease phenotype cannot be determined by only single gene [20], but need at least interaction information from a pair of genes. The frameworks for network-based classification and single-gene based classification are illustrated respectively in Fig. 1.

Besides, biological prior knowledge has been attracted many researchers to integrate the gene expression data with interaction networks [38, 39] such as metabolic networks, protein-protein interaction networks and gene networks since these networks are relevantly available. This effort is benefit to perform classification in order to identify marker genes from pathway perspective that globally represents diseases in term of network interactions. Such examples that used pathway analysis in their research are [2, 38, 39]. In term of network-based classification, the network construction is built from expression of gene pairs [2, 40].

Many methods have been proposed in order to construct network using microarray data. One of them, Bayesian network is seen as a fruitful method for such purpose [41-44] as it can representing network in matrix and topology-based form. In term of cancer classification, construction of subnetwork (group of genes based on their interaction) can be defined as unsupervised and supervised approach [45] and summarized in Table II.

To prevent bias in proposed model, [37] never used a priori biological information since the available biological networks not fully completed [46]. Thus, phenotype distribution will be used for such purpose similar to what have ever done by [47]. Moreover, synergic network was used by considering correlation between their joint expression levels and cancer [40]. Besides, topological-based [2, 48] and dependence network [36] were applied instead of using biological network.

TABLE I NETWORK-BASED VS SINGLE GENE-BASED

Issues	Single gene-based	Network-based
Addressing high dimensionality problem	[5] [6] [8] [11] [25] [26] [27]	[2] [36] [37] [39] [47] [48] [49]
Improving classification accuracy	[5] [8] [11] [25] [26] [27]	[36] [37] [39] [47] [48]
Marker genes detection	[5]	[36] [39] [50]
Benefit in biological interpretation	-	[36] [37] [38]
Reproducibility across data sets	-	[39]
Extensible to pathway-based	-	[2] [38] [39]

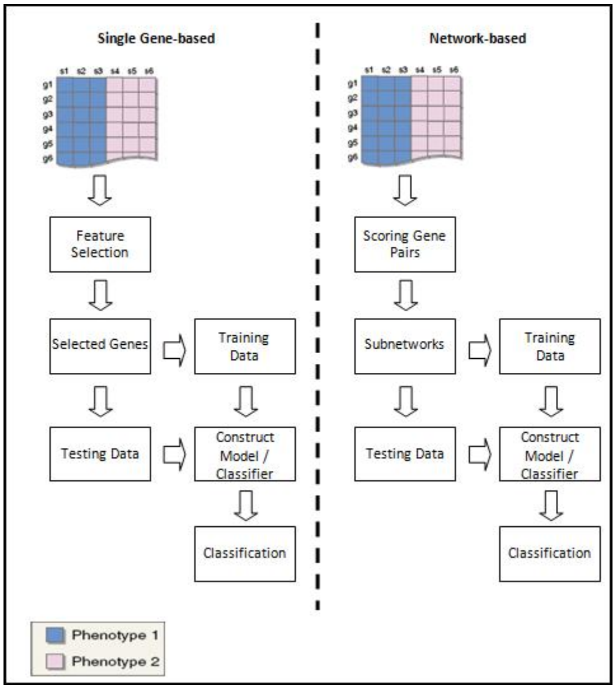

Figure 1 Framework for single-based and network-based classification.

TABLE II METHOD USED FOR SUBNETWORK CONSTRUCTION

Methods Used	Authors
Pearson Correlation	[2] [48]
Spectral Decomposition	[38]
Principal Component Analysis	[38]
Bayesian Network	[37]
Mutual Information	[39] [47]
Dependence Network	[36]
Simulated Annealing	[50]
K-nearest neighbor	[49]

VI. DISCUSSION AND CONCLUSION

Today, research in systems biology has shifted where focusing on data integration, network alignment, interactive visualization and ontological markup [51] towards pathway-related network. So, Instead of using classification approach in term cancer class prediction, classification can also be used in gene regulatory network by predicting the label of microarray data such in [52].

Similarly, established networks form interaction gene is significant towards molecular system biology. Although, many researcher interested to use biological prior knowledge such as protein-protein interaction (PPI), metabolic networks [46] and so on in term of prediction of phenotype [53], there is still make senses to study interaction genes in term of microarray data (DNA level) rather than proteomic and metabolic. Thus, the interactions in a small scope such as genes are significant towards analyzing interaction networks for systems biology as mentioned by [54].

As described above, subnetwork was regarded as training classifier and defined for four matters [55]. First, the groups of genes subject to the constraints of the molecular interaction network. Second, subnetworks are scored over only a subset of conditions. Third, only the significance of change for group genes will be considered. Fourth, some genes unaffiliated with any subnetwork will be left. In fact, network construction from microarray data has been faced with the problems of sparse data. So, the promising tools need to be applied.

Actually, not many researchers focusing on the statistical information that the comparison of different sample types contributes rather than just look for differently expressed genes to build their model [37]. Because of it, the methods that can deal with this matter is desired not only for precise accuracy but computation time as well.

Acknowledgment

This work was supported by Minister of Science, Technology and Innovation through grant Vot 79228. Also thank to our project leader, Prof. Dr. Safaai Deris for his full support.

References

[1] Keedwell, E. and Ajit, N. Intelligence Bioinformatics: The Application of Artificial Intelligence Techniques to Bioinformatics Problems. West Sussex, England: Wiley. 2005.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Liu, C.-C., Chen, W.-S. E., Lin, C.-C., Liu, H.-C, Chen, H.-Y., Yang, P.-C., Chang, P.-C., and Chen, J. J.W., "Topology-based Cancer Classification and Related Pathway Mining Using Microarray Data", Nucleic Acids Research, 2006, 34(14): 4069-4080.

[3] Ucar, D., Neuhaus, I., Ross-MacDonald, P., Tilford, C., Parthasarathy, S., Siemers, N. and Ji, R.-R., "Construction of a Reference Gene Association Network from Multiple Profiling Data: Application to Data Analysis", Bioinformatics, 2007, 23(20): 2716-2724.

[4] Piatetsky-Shapiro, G., and Tamayo, P., "Microarray Data Mining: Facing the Challenges", ACM SIGKDD Explorations, 2003, 1-5.

[5] Brown, M., P., S., Grundy, W., N., Lin, D., Cristianini, N., Sugnet, C., W., Furey, T., S., Ares, Jr., M., and Haussler, D., "Knowledge-based Analysis of Microarray Gene Expression Data by Using Support Vector Machines", Proceedings of the National Academy of Sciences of the United States of America, 2000, 97(1): 262-267.

[6] Dettling, M. and Bühlmann, P., "Boosting for Tumor Classification with Gene Expression Data", Bioinformatics, 2003, 19(9): 1061-1069.

[7] Lu, Y. and Han, J., "Cancer Classification Using Gene Expression Data", Information Systems, 2003, 28(4): 243-268.

[8] Qiu, P., Wang, J. and Liu, K.J.R., "Ensemble Dependence Model for Classification and Prediction of Cancer and Normal Gene Expression Data", Bioinformatics, 2005, 21(14): 3114-3121.

[9] Varadan, V. and Anastassiou, D., "Inference of Disease-Related Molecular Logic from Systems-Based Microarray Analysis", PLoS Computational Biology, 2006, 2(6): 0585-0597.

[10] Chen, X.-W., Anantha, G. and Wang, X., "An Effective Structure Learning Method for Constructing Gene Networks", Bioinformatics, 2006, 22(11): 1367-1374.

[11] Hanczar, B. and Dougherty, E.R., "Classification with Reject Option in Gene Expression Data", Bioinformatics, 2008, 24(17): 1889-1895.

[12] Bellazzi, R. and Zupan, B., "Towards Knowledge-based Gene Expression Data Mining", Journal of Biomedical Informatics, 2007, 40: 787-802.

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

[14] Fogel, G., Corne, D., W., and Pan, Y, Computational Intelligence in Bioinformatic, Piscataway, N.J.: IEEE Press Series on Computational Intelligence. 2008.

[15] Lai, C., "Supervised Classification and Spatial Dependency Analysis in Human Cancer Using High Throughput Data", 2008.

[16] Kuramochi, M. and Karypis, G., "Gene Classification using Expression Profiles: A Feasibility Study", International Journal on Artificial Intelligence Tools, 2005, 14(4): 641-660.

[17] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., and Lander, E.S., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", Science, 1999, 286(5439): 531-527.

[18] Perlich, C. and Merugu, S., "Gene Classification: Issues and Challenges for Relational Learning. International Conference on Knowledge Discovery and Data Mining", Chicago, Illinois: ACM, 2005, 61-67.

[19] Soinov, L.A., "Supervised Classification for Gene Network Reconstruction", Biochemical Society Transactions, 2003, 31(6): 1497-1502.

[20] Xu, M., Kao, M.-C. J., Nunez-Iglesias, J., Nevins, J. R., West, M. and Zhou, X. J., "An Integrative Approach to Characterize Disease-specific Pathways and Their Coordination: A Case Study in Cancer", BMC Genomics, 2008, 9(Supp 1):S12.

[21] Wang, L., Zhu, J. and Zou, H., "Hybrid Huberized Support Vector Machines for Microarray Classification and Gene Selection", Bioinformatics, 2008, 24(3): 412-419.

[22] Saeys, Y., Inza, I. and Larranaga, P., "A Review of Feature Selection Techniques in Bioinformatics", Bioinformatics, 2007, 23(19): 2507-2517.

[23] Yang, E., Maguire, T., Yarmush, M.L. and Androulakis, I.P., "Informative Gene Selection and Design of Regulatory Networks Using Integer Optimization", Computers and Chemical Engineering, 2008, 32: 633-649.

[24] Liu, Y., "Detect Key Gene Information in Classification", Eurasip Journal on Advances in Signal Processing, 2008, art. no. 612397.

[25] Sivagaminathan, R. K. and Ramakrishnan, S., "A Hybrid Approach for Feature Subset Selection Using Neural Networks and Ant Colony Optimization", Expert Systems with Appications, 2007, 33(2007): 49-60.

[26] Filippone, M., Masulli, F., and Rovetta, S., "Supervised Classification and Gene Selection Using Simulated Annealing", International Joint Conference on Neural Networks, July 16-21, Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada, 2006, 3566-3571.

[27] Van De Vijver, He, Y.D., Van □T Veer, L.J., Dai, H., Hart, A. A.M., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., Van Der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H. and Bernards R., "A Gene-Expression Signature as a Predictor of Survival in Breast Cancer", New England Journal of Medicine, 2002, 347(25): 1999-2009.

[28] Anastassiou, D., "Computational Analysis of the Synergy among Multiple Interacting Genes", Molecular Systems Biology, 2007, 3(art. no. 83).

[29] Cavalieri, D. and De Filippo, C., "Bioinformatic Methods for Integrating Whole-genome Expression Results into Cellular Networks", Drug Discovery Today: Biosilico, 2005, 10(10): 727-734.

[30] Nacu, S., Critchley-Thorne, R., Lee, P., and Holmes, S., "Gene Expression Network Analysis and Applications to Immunology", Bioinformatics, 2007, 23(7): 850-858.

[31] Liang, K-C., and Wang, X., "Gene Regulatory Network Reconstruction Using Conditional Mutual Information", Eurasip Journal on Bioinformatics and Systems Biology, 2008, art. no. 253894.

[32] Zainudin, S. and Deris, S., "Towards Evaluation of Inferred Gene Network", Proceedings of Fifth International Conference on Computational Science and Applications. IEEE. 2007, 57-62.

[33] Middendorf, M., Kundaje, A., Wiggins, C., Freund, Y. and Leslie., "Predicting Genetic Regulatory response Using Classification", Bioinformatics, 2004, 20(Supp. 1): i232-i240.

[34] Cawley, G.C. and Talbot, N.L.C., "Gene Selection in Cancer Classification Using Sparse Logistic Regression with Bayesian Regularization", Bioinformatics, 2006, 22(19): 2348-2355.

[35] Vert, J.P. and Kanehisa, M., "Extracting active pathways from gene expression data", Bioinformtics, 2003, 19(Supp. 2): ii238-ii244.

[36] Qiu, P., Wang, Z.J., Liu, K.J.R., Hu, Z.-Z., and Wu, C.H., "Dependence Network Modeling for Biomarker Identification", Bioinformatics, 2007, 23(2): 198-206.

[37] Armananzas, R., Inza, I., and Larranaga, P., "Detecting Reliable Gene Interactions by a Hierarchy of Bayesian Network Classifiers", Computer Methods and Programs in Biomedicine, 2008, 91: 110-121.

[38] Rapaport, F., Zinovyev, A., Dutreix, M., Barillot, E. and Vert, J. P., "Classification of Microarray Data Using Gene Networks", BMC Bioinformatics, 2007, 8(35): 1-15.

[39] Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D., and Ideker, T., "Network-based Classification of Breast Cancer Metastasis", Molecular Systems Biology, 2007, 3(140).

[40] Watkinson, J., Wang, X., Zheng, T. and Anastassiou, D., "Identification of Gene Interactions Associated with Disease from Gene Expression Data Using Synergy Networks", BMC Systems Biology, 2008, 2(10).

[41] Wang, M., Chen, Z. and Cloutier, S., "A Hybrid Bayesian Network Learning Method for Constructing Gene Networks", Computational Biology and Chemistry, 2007, 31: 361-372.

[42] Polanski, A., Polanska, J., Jarzab, M., Wiench, M. and Jarzab, B., "Application of Bayesian Networks for Inferring Cause-effect Relations from Gene Expression Profiles of Cancer Versus Normal Cells", Mathematical Biosciences, 2007, 209: 528-546.

[43] Chen, X., Chen, M. and Ning, K., "BNArray: An R Package for Constructing Gene Regulatory Networks from Microarray Data by Using Bayesian Network", Bioinformatics, 2006, 22(23): 2952-2954.

[44] Imoto, S., Goto, T. and Miyano, S., "Estimation of Genetic Networks and Functional Structures Between Genes by Using Bayesian Networks and Nonparametric Regression", Pacific Symposium on Biocomputing, 2002, 175-186.

[45] Ma, S. and Huang, J., "Penalized Feature Selection and Classification in Bioinformatics", Briefings in Bioinformatics, 2008, 9(5): 392-403.

[46] Ma, H. and Goryanin, I., "Human Metabolic Network Reconstruction and Its Impact on Drug Discovery and Development", Drug Discovery Today, 2008.

[47] Hanczar, B., Zucker, J. D., Henegar, C., and Saitta, L., "Feature Construction From Synergic Pairs to Improve Microarray-based Classification", Bioinformatics, 2007, 23(21): 2866-2872.

[48] Liu, C.-C., Chen, W.-S.E., Chang, P.-C., and Chen, J.J.W., "Topological-based Classification Using Artificaial Gene Networks", Proceeding of the Fourth IEEE International Conference on Cognitive Informative (ICCI 2005), August 8-10, University of California, Irvine, USA: IEEE, 2005, 49-56.

[49] Lin, Y.-C., Yeh, H.-Y., Cheng, S.-W., and Soo, V.-W., "Comparing Cancer and Normal Gene Regulatory Networks Based on Microarray Data and Transcription Factor Analysis", Proceedings of the 7th IEEE International Conference, October 14-17, IEEE, 2007, 151-157.

[50] Guimera, R., Sales-Pardo, M., and Amaral, L.A.N., "A Network-based method for Target Selection in Metabolic Networks", Bioinformatics, 2007, 23(13): 1616-1622.

[51] Balaji, S. S., Nigam, H. S., Jason, A. F., Eduardo, A., Antal, F. N., and Serafim B., "Current Progress in Network Research: Toward Reference Networks for Key Model organisms", Briefings in Bioinformatics, 2007, 8(5): 318-332.

[52] Kundaje, A., Middendorf, M., Shah, M., Wiggins, C., H., Freund, Y., and Leslie, C., "A Classification-based Framework for Predicting and Analyzing Gene Regulatory Response", BMC Bionformatics, 2006, 7(Supp. 1).

[53] McGary, K.L., Lee, Insuk and Marcotte, E.M., "Broad Network-based Predictability of Saccharomyces Cerevisiae Gene Loss-of-function Phenotypes" Genome Biology, 2007, 8(12).

[54] Bader, S., Kühner, S., and Gavin, A.-C., "Interaction Networks for Systems Biology", Federation of European Biochemical Societies Letters, 2008, 582(8): 1220-1224.

[55] Ideker, T., Ozier, O., Schwikowski, B. and Siegel, A.F., "Construction of Genetic Network Using Evolutionary Algorithm and Combined Fitness Function", Bioinformatics, 2002, 18(Supp. 1): S233-S240.