Proceedings of the 2004 IEEE
Conference on Cybernetics and Intelligent Systems
Singapore, 1-3 December, 2004

Experimental Condition Selection in Whole-Genome Functional Classification

Zexuan Zhu, Y. S. Ong, K. W. Wong, K. T. Seow
School of Computer Engineering
Nanyang Technological University
Blk N4, Nanyang Avenue, Singapore 639798
Email: zhuzexuan@pmail.ntu.edu.sg, asysong@ntu.edu.sg, askwwong@ntu.edu.sg, asktseow@ntu.edu.sg

Abstract— **Microarray technologies enable the quantitative simultaneously monitoring of expression levels for thousands of genes under various experimental conditions. This is new technology has provided a new way of learning gene functional classes on a genome-wide. Previously, lots of unsupervised clustering methods and supervised classification have shown power in assigning functional annotations based on gene co-expression. However, due to the noisy and highly dimensional nature of microarray data and the inherent heterogeneity of gene functional classes, the whole-genome learning of gene functional classes from microarray data has remained a great challenge for scientists. Currently, most of the methods do not discriminate the different attribution of experimental conditions in the learning process, which impaired the ability of learning functional classes and prevented these methods from discovering the links between the experimental conditions and gene functional classes. In this study, we perform a selection of experiment conditions during the systematically learning of ~100 functional classes categorized in MIPS's comprehensive yeast genome database. In particular, a hybridization of genetic algorithm and k-nearest neighbors classifier has been adopted here. Through a comparison of the results with other previous methods our studies indicate promising improvements in learning performance. Further, by identifying the critical experimental conditions, significant links between the experiments and the functional classes were uncovered.**

I. Introduction

Microarray technology has attracted increasing interest in many academic communities and industries over the recent years. This new breakthrough promises a new insight into the mechanisms of living systems by providing a way to simultaneously measure the activities and interaction of thousands of genes. It is supposed genes that co-regulate and exhibit similar patterns of expressions when exposed to identical experimental conditions are likely to be involved in similar biological functional classes. By analyzing the large volume of gene expression data, it is possible to discover the underlying functional groupings of genes that are involved in a particular pathway, or respond to a common environmental stimulus. Through these analyses, the function of undetermined gene products can then be systematically identified through guilt-by-association principle [1].

Over the recent years, a number of clustering algorithms (e.g. K-means [11], SOM [12,13], hierarchical clustering [14], graph theoretic approaches [15,16], Fuzzy C-means [17], etc) and classification algorithms (e.g. SVM [18], ANN [19], etc) have been employed for gene functional analysis using microarray data. Through these analyses a significant amount of new discoveries have been made and new understandings of the living systems were generated. However, the learning of gene functional classes from microarray expression data has remained a great challenge to computer scientists. In particular, the difficulties mainly lie in the nature of genome expression data. Microarray data is inherently noisy and highly dimensional. The natural biological fluctuations are likely to import measurement variations and bring implications to microarray analysis. In addition, the microarray experiment involves complex scientific procedures during which errors are commonly introduced due to the imperfections of instruments, impurity of materials and negligence of scientist. Microarray data is also high dimension with thousands of genes and hundreds of samples (arrays). This makes learning from microarray data an arduous task under the effect of curse of dimensionality. The biological heterogeneity is another factor that deters the successful analysis of the data. The gene functional classes exhibit great intra heterogeneity due to the difference in derivation organisms and complex regulation systems. Genes are categorized into the same class based on their similar transcript in different biochemistry experiments. However, this is does not warrant their co-transcription when exposed to all the given experiments in microarray.

To alleviate these problems, feature selection was often used to identify the subset of relevant features and eliminate irrelevant ones. Generally, feature selection tends to fall under two categories, namely, filter methods and wrapper methods. The former selects features according to criteria that are independent of the learning machine. Signal-to-noise [3,4], T-statistics [5], entropy-base [6] and χ^2-statistics [7] are some of the commonly employed criteria used for ranking genes in cancer data. In the later, the machine-learning algorithm employed affects

0-7803-8643-4/04/$20.00 © 2004 IEEE

the search of important feature subset directly. Evolutionary algorithms are used to search for this subset of features due to its well-known ability to produce high quality solution at tractable time even on complex problems. In [8-10], Genetic Algorithm (GA) has been shown to effectively identify discriminative genes in multi-class tumor medical diagnostic tests.

In this study, our focus is to systematically classify ~100 gene functional classes categorized in MIPS's CYGD [2] (Comprehensive Yeast Genome Database) on a genome-wide scale. By conducting feature selection on the experimental conditions, we aim to identify and select the most significant subset of conditions given a particular functional class of genes, so that the classification becomes less affected by the noisy experiments. At the same time, the learning performance of the classifier may be improved. In addition, the identification of relevant conditions further reveals the underlying links among the specific sets of genes and pathway. In this study, a genetic algorithm/k-nearest neighbors wrapper method is adopted to perform the feature selection.

II. System and Methodologies

A. Dataset and Data Preprocessing

In the present study, we use the genomic expression data of wild-type *S. cerevisiae* selected by Gasch et al [17], which is achieved under the experiments described in [20-23]. The 6153 genes and 93 experiments are represented by a gene expression matrix of dimension 6153×93. The magnitude of the element value indicates the expression level of the gene in the corresponding experimental condition. The missing value (i, j) in the data set was estimated with a weighted average of values in experiment j of k other genes that have a value present in experiment i and expression most similar to j in all experiment other than i [24].

Among the 6153 genes, only 3700 were chosen for this study based on the availability of accurate functional annotations in the MIPS's CYGD reference database [2]. This database contains several hundred functional classes, whose definitions come from biochemical and genetic studies of gene function. Because it is arranged in a hierarchical scheme, we normalize the reference functional annotations by using functional class labels up to level 2. There ~100 functional classes containing more than three genes are used in the training process. The database is available at http://mips.gsf.de/proj/yeast/catalogues/funcat.

B. Methodology

In this multi-class classification problem, we applied a one-versus-all strategy. For each functional class, the data set is split into two parts: genes belonging to the given function (positive instances) and genes not in this function (negative instances). A binary GA/KNN classifier is built for each class. The GA/KNN classifier consists of two main components: (1) a GA-based experimental condition selector and (2) a KNN gene classifier.

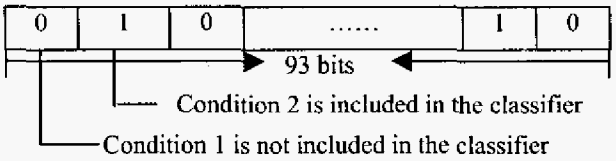

Fig 1. A representation of chromosome as a binary bit string

It is not practical to use "brute force" to select a subset of experiments conditions from a total of 93 to jointly discriminate between the different classes of genes. Here GA is used as it has been shown to be effective and efficiency in searching for the global optimal of complex high-dimensional problems [27]. There are three major design decisions to consider when implement a GA to solve a particular problem. In the GA, a representation for candidate solutions must be chosen and encoded as a chromosome. Here, a chromosome is denoted as a 93 binary bit string, which coincides with the 93 experimental conditions considered. Hence a '1' at location bit 2 implies the inclusion of condition 2 in the classification process, and vice versa (Fig. 1).

A GA population size of 20 chromosomes is initialized randomly. Linear ranking selection is used for selection, with two-point crossover and mutation operators at probabilities 0.9 and 0.05, respectively, and a stopping criterion of 50 generations. The objective fitness function considered in this study is based on the F-measure [28] given in (1):

$$F(c) = \frac{(b^2+1)PR}{b^2P+R} \tag{1}$$

BEGIN

For each non-trivial gene functional class

Split the positive and negative samples

Begin GA

Initialize: generate an initial GA population

While (generation < 50)

Evaluate: for each chromosome, run 5 times 3-fold cross-validation of KNN (k=3) with conditions included in the chromosome. Return the average F-measure (1) as fitness.

GA Operators: apply roulette wheel selection, two-point crossover (r_c=0.9) and mutation (r_m=0.05) to generate new population.

End While

End GA

End For

END

Fig 2: The framework of the GA/KNN method.

where P and R denote the precision and recall of KNN classifier respectively. A three-fold stratified cross-validation is used to explore the precision and recall. It is worth noting that due to the random nature on the partition of the data set, any two different runs of cross-validation often results in different outcomes, hence, we repeat the three-fold cross-validation five times and the average precision and recall is used. Parameter b adjusts the weighting for precision and recall. In this study, precision and recall are weighted equally, i.e. b is set to one in (1). The summary of the GA/KNN functional classification process is outlined in Fig 2.

III. Empirical Results and Discussion

The results obtained from our empirical studies on GA/KNN functional classification of whole-genome data are presented in Table 1. The top 25 functional classes of highest average F-Measure are listed in the descending order. The references of functional annotations in the second level of CYGD functional categorization are tabulated in the first column. The second column indicates the number of genes found in each functional class. The entries in the third and fourth columns are the means and standard variances for precision and recall over five runs using three-fold stratified cross-validation scheme. The fifth column shows the F-Measure derived based on (1). The numbers of experimental conditions selected are listed in the last column. On the average, 36 conditions were selected among the top 25 classes.

These results indicate that most of the functional classes were learned at a precision above 0.5, however the recalls were relatively low in comparison, with only 5 functional classes having a value greater than 0.4. The effective performance for each class is evaluated based on F-Measure, which balances both precision and recall. From a biological perspective, the obtained results match well with earlier studies [19] where the 'organelle-specific protein' function (e.g. organization of cytoplasm, mitochondrial organization, nuclear organization, organization of chromosome structure, etc) was exposed to be more learnable, and the "ribosomal proteins" function being the most learnable class among all others.

A. *Comparison with other methods*

Over the recent years, many machine-learning methods have been considered for functional classification. Among them, Support Vector Machine (SVM) and Artificial Neural Network (ANN) were regarded as better learners of gene functional classes [18,19]. In the present empirical study, we consider four state of the art learning methods for gene functional classification limited to the top 25 functions. These include SVM with radial basis kernel, back-propagation multi-layer perceptron network, K-nearest neighbour (KNN) and GA+KNN. The model

TABLE I. The top 25 functional classes with highest average F-Measure learned by GA/KNN method.

Functional Class	Size	Precision	Recall	F-Measure	Features
ribosomal proteins	202	0.89±0.01	0.78±0.00	0.84±0.01	41
glycolysis and gluconeogenesis	35	1.00±0.00	0.40±0.04	0.57±0.04	37
tRNA-synthetases	37	0.83±0.03	0.41±0.03	0.55±0.05	41
respiration	74	0.72±0.03	0.42±0.01	0.53±0.00	40
organization of cytoplasm	554	0.68±0.01	0.40±0.00	0.50±0.01	31
mitochondrial organization	337	0.65±0.01	0.39±0.01	0.48±0.01	33
tricarboxylic-acid pathway	23	0.75±0.09	0.26±0.05	0.39±0.07	37
amino-acid transporters	25	0.86±0.08	0.24±0.00	0.38±0.01	31
amino-acid metabolism	203	0.75±0.02	0.25±0.01	0.38±0.01	31
organization of endoplasmatic reticulum	154	0.50±0.02	0.30±0.01	0.38±0.01	36
fermentation	34	0.89±0.06	0.24±0.00	0.37±0.01	42
nuclear organization	758	0.40±0.01	0.34±0.01	0.37±0.01	44
nitrogen and sulphur metabolism	73	0.63±0.04	0.26±0.01	0.37±0.02	34
rRNA transcription	104	0.48±0.03	0.28±0.03	0.35±0.03	27
proteolysis	152	0.60±0.03	0.23±0.01	0.33±0.01	32
translation	62	0.54±0.05	0.24±0.01	0.33±0.02	30
organization of chromosome structure	41	0.98±0.00	0.20±0.00	0.33±0.00	44
transport ATPases	38	0.48±0.03	0.24±0.01	0.32±0.01	34
mRNA transcription	559	0.35±0.00	0.28±0.01	0.31±0.01	38
intracellular transport vesicles	42	0.53±0.04	0.21±0.02	0.31±0.02	41
peroxisomal organization	39	0.64±0.04	0.18±0.01	0.28±0.01	32
C-compound and carbohydrate metabolism	413	0.47±0.01	0.19±0.00	0.27±0.00	40
mitochondrial transport	72	0.69±0.04	0.15±0.01	0.25±0.01	42
nucleotide metabolism	142	0.70±0.03	0.15±0.01	0.24±0.02	34
phosphate metabolism	32	1.00±0.00	0.13±0.02	0.22±0.02	40

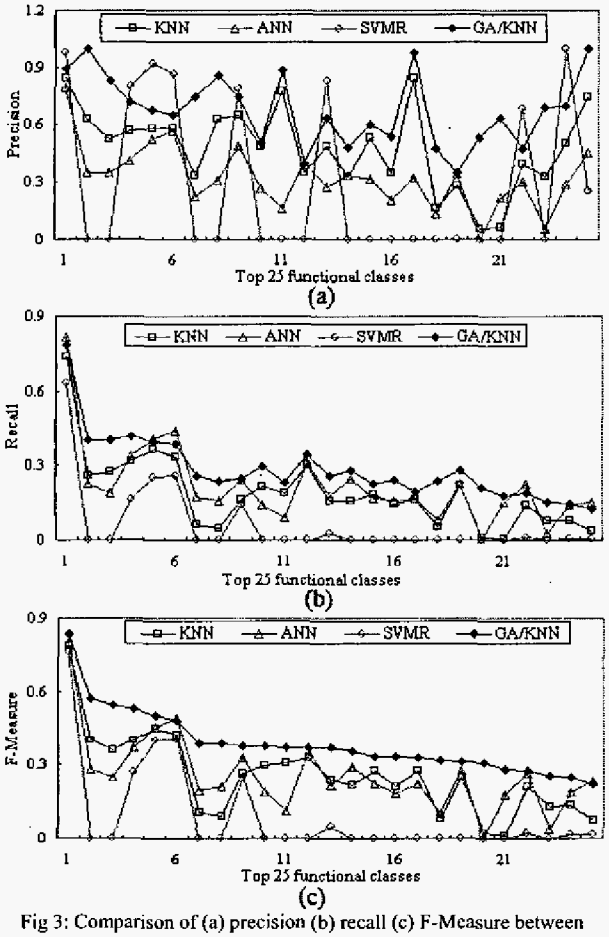

Fig 3: Comparison of (a) precision (b) recall (c) F-Measure between KNN, ANN, SVM and GA/KNN on the top 25 functional classes.

parameters of the machine-learning methods were configured according to [18,19] or the defaults. On each class, we conducted 5 runs of three-fold cross-validation using each method and the average of precision, recall and F-measure are plotted in Fig. 3. Because most of the standard deviations are small relative to the mean, they are omitted for the sake of brevity.

From Fig. 3(a), it may be observed that GA/KNN exhibits better result than other methods in most of the functional classes. SVM displays great inconsistency with the best classification performance in several classes whereas zero precision on the others. Moreover, SVM seems to favor a higher precision in those classes at the cost of a lower recall. ANN exhibits a consistent level of precisions on all the classes. Noticeably, the non-parametric method, KNN performs better than other sophisticated methods, this is mainly due to the noise and the imbalance in the number of positive and negative examples. The relatively large proportion of negative training examples easily outweighs the small number of positive examples in each class. Further, the erroneous examples in the training set often mislead learning in SVM and ANN. As a result, SVM and ANN are not correctly trained to separate the classes. In contrast, KNN is a lazy learner that predicts target sample based on the local information, which make it much less susceptible to be affected by the noisy data and erroneous training examples.

The average recalls for the various machine learning techniques are plotted in Fig. 3(b). These methods are generally competitive except for SVM, which fails to learn many of the classes (i.e. 0 recall rates or at best $\leq$ 0.4). Such performances are generally caused by the natural heterogeneity in functional classes. Because different complexes are elicited under different conditions, the genes are unlikely to be expressed in a coordinated fashion under different conditions. Accordingly, genes in the same functional class are not necessary to express coherent pattern (shown in Fig. 4) that is assumed by machine learning.

The effective performances of the methods are also summarized in Fig. 3(c) using F-measure. Based on F-measure, KNN and ANN display competitive performance. SVM is overwhelmed by other methods ascribing to its incapability of learning many classes. While, GA/KNN displays competitive to the other methods in terms of recall, it emerges as superior in F-measure due to the evidently good precision. This indicates that GA/KNN is capable of efficiently filtering out the false positive samples without significant change in true positive.

B. Reproducibility and Heterogeneity

To assess the reproducibility of the selected conditions, we independently perform the GA/KNN algorithm 50 times on functional class "glycolysis and gluconeogenesis". The conditions were selected with frequency > 0.5 and diagrammed in the top section of Fig. 4. Genes belonging to class "glycolysis and gluconeogenesis" are hierarchically clustered and shown in the Eisen plot [14], bottom section of Fig. 4, where they are clustered into subclass A and B.

In the top section of Fig. 4, 15 conditions are selected with frequency > 0.7. Among them, condition 4, 55, 64, 69 and 88 were most frequently selected. We can find that these conditions show most coherent expression in the subclass A. Hence, subclass A is best characterized with highly induction in response to phosphate limitation (condition 4) [21] and diamine treatment (condition 55), while strong repression in response to other environmental changes including amino acid starvation (condition 64), nitrogen starvation (condition 69) and stationary phase (condition 88) [23]. These relationships between the experimental conditions and functional classes are essential for revealing the underlying mechanism of biological process. For example, the induction of class "glycolysis and gluconeogenesis" in the condition 4 may implicate the importance of phosphate in Glycolysis / Gluconeogenesis metabolism.

The repeated selection of identical conditions indicates the ability of GA/KNN to cream off the discriminative

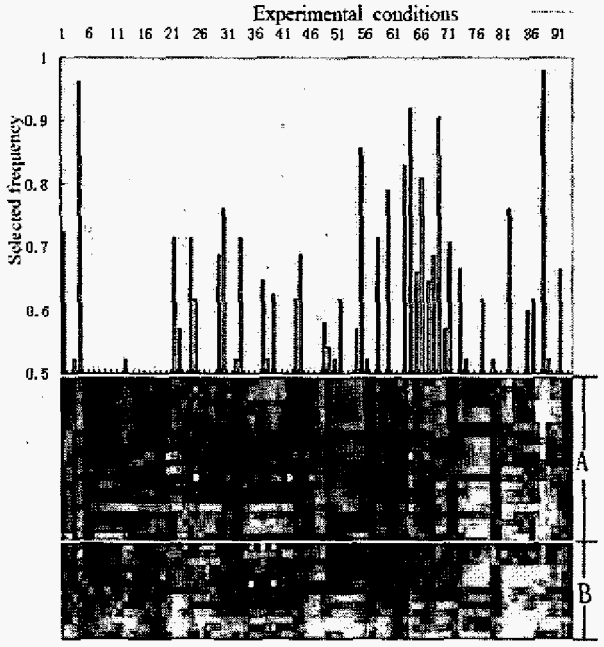

Fig 4: Top section: conditions selected with frequency >0.5 in the 50 runs of GA/KNN on class "glycolysis and gluconeogenesis". The x-axis presents the 93 experimental conditions and y-axis represents the selected frequency. Bottom section: Eisen plot of expression profile of class "glycolysis and gluconeogenesis", where genes (rows) are organized based on their similarities. Color red represents a gene is induced in the corresponding experiment and green represent repression of the gene. The color saturation represents the magnitude of the expression ratio. Black indicates no detectable difference in expression levels. In this combinogram, two sections are aligned with experimental conditions. For each condition, its selected frequency and express profile is diagram in the same column.

features as well as filter out the noisy and redundant ones. In the bottom section of Fig. 4, conditions 80-93 show similar expression profiles, which compose the whole progression of stationary phase, but only the best one, condition 88 was selected all the times. As such, condition 55 and 69 were selected much more frequently than other diamide treatments (conditions 49-56) and nitrogen starvation (conditions 69-78).

This expression profile in Fig. 4 also shows the evidential heterogeneity in this class. All the genes were grouped into 2 subclasses with subclass A containing the learnable genes discovered in this study. Most genes in subclass A are repressed during all the experiments. They are found to be involved in the Glycolysis / Gluconeogenesis pathway in KEGG's PATHWAY database [29]. However, genes in subclass B were highly induced in most of the experiments. The discovery of common transcription factor biding sites (MSN2/4) within 800 bp upstream of their open reading frames [17] indicates their co-regulation corresponding to "environmental stress response" [23]. Unfortunately, Msn2/Msn4p also regulated other classes (e.g. respiration, tricarboxylic-acid) at the same time. Therefore, genes in subclass B are much more likely to be assigned to other classes which make it difficult to be learned by classifier even when feature selection is attempted. One of the solutions for this heterogeneity problem is nonetheless to collect more experimental data.

IV. Conclusions

In this paper, we have conducted a study on using hybrid genetic algorithm/k-nearest neighbors classifier for genome-wide yeast gene expression data. In particular, we employ the GA to identify the critical experimental conditions and evaluate the goodness of candidate solutions using the k-nearest neighbors classifier. The hybrid method is shown to be capable of improving the performance of functional class annotation when compared to existing state of art machine learning methods. Further, our studies also expose the relationships between the experiment conditions and specific functional classes. These revelations will enhance new developments in pathway and regulatory system analysis.

References

[1] M. G. Walker, W. Volkmuth, E. Sprinzak, et al. Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome Research*, 9(12): 1198-203, (1999).

[2] H. W. Mewes, D. Frishman, U. Guldener, et al. Mips: a database for genomes and protein sequences. *Nucleic Acids Res*, 30(1): 31-4, (2002).

[3] T. R. Golub, D. K. Slonim, P. Tamayo, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439): 531-7, (1999).

[4] A. A. Scott, E. S. Jane, et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, volume 30, (2002).

[5] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(25): 436-442, (2002).

[6] J. Li and L. Wong. Emerging patterns and gene expression data. *Genome Informatics*, 12:3-13, (2001).

[7] H. Liu, J. Li and L. Wong. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics* 13, 51-60, (2002)

[8] L. Li, C. R. Weinberg, T. A. Darden and L. G. Pedersen. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, 17: 1131-42, (2001)

[9] C. H. Ooi and P. Tan. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics*, 19: 37-44, (2003)

[10] S. Peng, Q. Xu, X. B. Ling, et al. Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. *FEBS*, 555: 358-63, (2003)

[11] S. Tavazoie, J. D. Hughes, et al. Systematic determination of genetic network architecture. *Nature genetics*, 22, 281-5, (1999)

[12] P. Tamayo, D. Slonim, J. Mesirov, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA*, 99(6): 2907-12, (1999)

[13] P. Toronen, M. Kolehmainen, G. Wong and E. Gastren. Analysis of gene expression data using self-organizing maps. *FEBS lett*, 451(2), 142-6, (1999)

[14] M. B. Eisen, P. T. Spellman, P. O. Brown & D. Botstein, 'Cluster analysis and display of genome-wide expression patterns', *Proc Natl Acad Sci USA*, 95(25): 14863-8, (1998)

[15] E. Hartuv, A. Schmitt, et al. An Algorithm for Clustering cDNAs for Gene Expression Analysis, *RECOMB*, (1999).

[16] R. Shamir and R. Sharan. Click: A clustering algorithm for gene expression analysis. *ISMB 00*, (2000).

[17] A. P. Gasch and M. B. Eisen. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology*, 3(11): 1-22, (2002)

[18] M. P. S. Brown, W. N. Grundy, D. Lin, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. *PNAS*, 97(1): 262-7, (2000)

[19] A. Mateos, J. Dopazo, R. Jansen, et al. Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. *Genome Research*, 12: 1703-15, (2002)

[20] T. J. Lyons, A. P. Gasch, L. A. Gaither, et al. Genome-wide characterization of the Zap1p zinc-responsive regulon in yeast. *Proc Natl Acad Sci USA*, 97:7957-7962, (2000)

[21] N. Ogawa, J. DeRisi, P. O. Brown. New components of a system for phosphate accumulation and polyphosphate metabolism in Saccharomyces cerevisiae revealed by genomic expression analysis. *Mol Biol Cell*, 11:4309-4321, (2000)

[22] A. P. Gasch, M. Huang, S. Metzner, D. Botstein, S. J. Elledge, P. O. Brown. Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Mol Biol Cell*, 12:2987-3003, (2001)

[23] A. P. Gasch, P. T. Spellman, C. M. Kao, et al. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, 11:4241-4257, (2000)

[24] O. Troyanskaya, M. Cantor, G. Sherlock, et al. Mission value estimation method for DNA microarrays. *Bioinformatics*, 17(6): 520-5. (2001)

[25] S. A. Armstrong, J. E. Staunton, et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, vol 30, (2002)

[26] D. Singh, P. G. Febbo, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, vol 1, (2002)

[27] J. Holland. Adaptation in natural and artificial systems. 2nd edition, *MIT Press, Cambridge, Massachusetts*, (1992)

[28] I. H. Witten and E. Frank. Data mining: practical machine learning tools and techniques with java implementations. *Academic Press*, (2000)

[29] M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 28, 27-30 (2000)