Application of Emerging Patterns for Multi-source Bio-Data Classification and Analysis*

Hye-Sung Yoon[1], Sang-Ho Lee[1], and Ju Han Kim[2]

[1] Ewha Womans University, Department of Computer Science and Engineering, Seoul 120-750, Korea
comet@ewhain.net, shlee@ewha.ac.kr
[2] Seoul National University Biomedical Informatics (SNUBI), Seoul National University College of Medicine, Seoul 110-799, Korea
juhan@snu.ac.kr

Abstract. Emerging patterns (EP) represent a class of interaction structures and have recently been proposed as a tool for data mining. Especially, EP have been applied to the production of new types of classifiers during classification in data mining. Traditional clustering and pattern mining algorithms are inadequate for handling the analysis of high dimensional gene expression data or the analysis of multi-source data based on the same variables (e.g. genes), and the experimental results are not easy to understand. In this paper, a simple scheme for using EP to improve the performance of classification procedures in multi-source data is proposed. Also, patterns that make multi-source data easy to understand are obtained as experimental results. A new method for producing EP based on observations (e.g. samples in microarray data) in the search of classification patterns and the use of detected patterns for the classification of variables in multi-source data are presented.

1 Introduction

Microarray experiments have brought innovative technological development to the classification of biological types. But more powerful and efficient analytical strategies need to be developed to carry out complex biological tasks and to classify data sets with various types of information such as mining disease related genes and building genetic networks.

The analytical strategy of bio-data can be classified into two categories according to the form of learning algorithm. First, as an unsupervised learning method such as typical clustering algorithms, the analytical method deals directly with genes while ignoring the biological attributes (labels) when handling DNA data (instance). Supervised learning is a target-driven process in that a suitable induction algorithm is employed to identify the genes that contribute the most toward a specific target, such as the classification of biological types, gene

* This work was supported by the Brain Korea 21 Project in 2004.

L. Wang, K. Chen, and Y.S. Ong (Eds.): ICNC 2005, LNCS 3610, pp. 965–974, 2005.

mining or data-driven gene networking[8]. Among supervised learning methods, rule based approach can be said that this partitions the sample and feature gene space simultaneously and it is especially an efficient method for classification of multi-source data. In statistics, the analysis of gene expression profiles is related to the application of particular supervised learning schemes. The structure of gene expression profiles must be suited for the typical data situation with a small number of patients n (=observations) and a large number of genes p (=variables), the so-called 'small n large p' paradigm in gene expression analysis[2][12].

In this paper, we develop a new rule-based ensemble method using EP for the classification of multi-source data. EP are those whose support changes significantly from one data set to another[19]. EP are among the simplest examples used to understand interaction structures, and are not only highly discriminative in classification problems[19], but can also capture the biologically significant information from the data. However, a very large volume of EP is generated for high dimensional gene expression data[7]. In this paper, we apply concise EP for multi-source data classification based on observations from each individual data set. When dealing with classification methods, microarray data is generally used, but only a few number of approaches are designed to consider explicitly the interaction among the genes being investigated. Interaction is well understood as (co-)expression genes in a cell governing a complicated network of regulatory controls. Hence, the interdependencies of all genes must be taken into consideration in order to achieve optimal classification. We propose a new method that can handle all variables in an appropriate way. It must be noted that the goal of the analyses presented in this paper is not to present correlated interaction genes for multi-source data but rather to illustrate our proposed classification method using EP.

The remainder of this paper is organized as follows. The application of multi-source data and classification methods in bioinformatics, and the analysis of EP are reviewed in section 2. A method for extracting EP from multi-source data sets and their applications are explained in section 3. Furthermore, significant experimental results by applying the proposed method and its details are described in section 4. Finally, concluding remarks and future works are presented in section 5.

2 Related Works and Background

In section 2, multi-source data, classification algorithm applications in bioinformatics and EP for multi-source data classification are reviewed as related works and background.

2.1 Multi-source Data

Bioinformatics not only deals with raw DNA sequences, but also with other various types of data, such as protein sequences, macromolecular structure data, genomes data and gene expression data[19]. The various types of data provide

researchers with the opportunity to predict phenoma that were formerly considered unpredictable, and most of these data can be accessed freely on the internet.

We assume that the analysis of combined biological data sets leads to more understandable direction than experimental results derived from a single data set. The purpose for combining and analyzing different types of data is to identify with more accuracy and to provide more correlations using diverse independent attributes in gene classification, clustering and regulatory networks and so on. Among the features of bio-data, one is that the same variables can be used to make various types of multi-source data through a variety of different experiments and under several different experimental conditions. These multi-source data are useful in understanding cellular function at the molecular level and also provide further insight into their biological relatedness by use of information from disparate types of genomic data. In [14], the problem of inferring gene functional classification from a heterogeneous data set consisting of DNA microarray expression measurements and phylogenetic profiles from whole-genome sequence comparisons is considered. As a result, it is proposed that more important information can be extracted by using disparate types of data.

2.2 Classification Problem in Bioinformatics

Classification problems aim at building an efficient and effective model for predicting class membership of data. Initial analysis of multi-source data focused on clustering algorithms, such as hierarchical clustering[9] and self-organizing maps[16]. In these unsupervised learning algorithms, genes that share similar expression patterns form clusters of genes that may show similarities in function[14]. But, because clustering methods ignore biological attributes (labels), they have limitations in the search of attributes or the discovery of rules in observations.

In [10], supervised learning techniques were applied to microarray expression data from yeast genes. It was verified through this application that an algorithm known as support vector machine (SVM)[3][4][13] provides excellent improvement in classification performance compared to a number of other methods, including Parzen windows and Fisher's linear discriminant[10]. Also, the methods used in [10] have been successfully applied to disease genes classification with machine learning approaches such as support vector machines (SVM), artificial neural network(ANN), k-nearest neighbors (kNN), and self-organizing map (SOM). In recent studies on the application of classification methods, supervised learning methods are aiming at showing the existence or nonexistence of disease by searching for disease genes[19].

2.3 Analysis of Emerging Patterns

A wide variety of gene patterns can be found for each data set. In [2] and [19], gene expression profiles were used to individually apply CART algorithm, a supervised learning method, and clustering method, an unsupervised learning

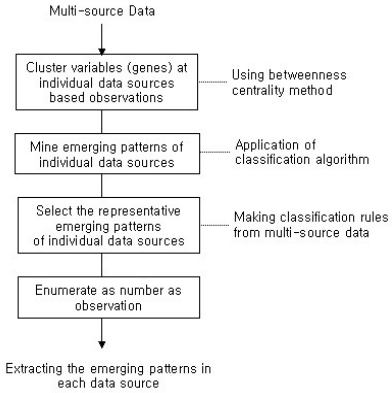


Fig. 1. Flowchart of the experimental method

method, to detect all types of early cancer development. To improve accuracy in classification, EP were applied to express the interaction between cancer-causing genes. Pattern association and clustering are both data mining techniques that are frequently applied in the fields of cancer diagnosis and correlation studies of gene expression[20]. But the results from these methods do not meet our requirements because multi-source data was not considered. EP were first introduced in [5], and they were defined as the item set that significantly increases support in each data sets D_1 and D_2 using the appropriate cut-off value of the growth rate. Unlike frequent patterns in common association analysis, EP are applied to classification problems to provide high discrimination, and are proved to be more useful. Also, EP are easy to understand because they are the collections of attributes in a dataset, and this property is especially important in bioinformatics application problems.

Thus, this paper proposes an efficient classification method using EP that is efficient when using analysis based on the smaller number of observational attributes rather than the very large number of variable attributes.

3 Methods

In this section, the experimental data and experimental methods applied in this paper are explained in detail. The overall framework is illustrated in Figure 1 and it will be explained in order.

3.1 Data

In this paper, two types of genomic data were used as multi-source data for the application of the proposed method. The first data set was derived from a collection of DNA microarray hybridization experiments. Each data point in the microarray data represents the logarithm of the ratio of expression levels of a particular gene under two different experimental conditions. The data consists of a

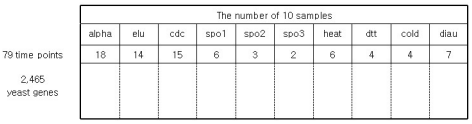


Fig. 2. Data structure of microarray data

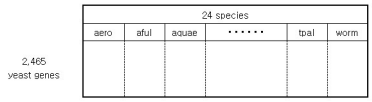


Fig. 3. Data structure of phylogenetic profile

set of 79-element gene expression vectors across time points for 2,465 yeast genes. These genes were selected by [9] based on accurate functional annotations. The data were collected at various time points during the diauxic shift[6], the mitotic cell division cycle[15], sporulation[17], and temperature and reducing shocks, and are available on the stanford website (http://www-genome.stanford.edu)[14].

In addition to the microarray expression data, we applied data characterized by 24 phylogenetic profiles[11] to each of the 2,465 yeast genes. In this data set, a phylogenetic profile is a bit string, in which the boolean value of each bit reflects whether the gene of interest has a close homolog in the corresponding genome. The profiles employed in this paper contain, at each position, the negative logarithm of the lowest E-value reported by BLAST version 2.0[18] in a search against a complete genome, with negative values truncated to 0. The profiles were constructed using 24 complete genomes, collected from the Institute for Genomic Research website (http://www.tigr.org/tdb) and from the Sanger Centre website (http://www.sanger.au.uk). Prior to learning, the gene expression and phylogenetic profile vectors were adjusted to have a mean of 0 and a variance of 1. The description of each data set, composed of the microarray data and phylogenetic profile data about the 2,465 yeast genes, are as shown in figure 2 and figure 3.

In the experiments of this paper, the betweenness centrality values based on 10 samples (=observations) that have 79-element time points values in the first microarray data set were as shown in figure 2. Thus, gene clusters were formed by extracting the most closely related genes in the order of high betweenness centrality value first.

As a result, a total of 10 clusters were formed (all genes were included in at least one sample in this experiment). And also for the second phylogenetic profile data set, 25 clusters (one additional cluster was formed with the genes that were not included in any of the 24 species) were formed by extracting the most closely related genes in order of high betweenness centrality value first.

3.2 Application of Betweenness Centrality Based on Observation

Bio-data is characterized by having a small number of observations compared to the number of variables. This characteristic found in bio-data can also be

observed in microarray data, and this is well reflected in the data where the number of columns corresponding to observations are outnumbered by the number of rows corresponding to variables (=genes). The exclusion of some variables can lead to significant differences in experimental results because a characteristic of bio-data is that interaction among the data is highly dependent. Therefore, in this paper, the characteristic EP are represented by considering all variables in each data set and the results are applied for the classification of multi-source data. Also, since EP are easy to understand and represent, they are useful for judging the features other types of data.

The following explains in order the proposed method of forming EP with a single data set.

1. First, based on the observations, clusters are formed with genes that contribute the most toward these observations, since bio-data sets have a smaller number of observations than variables. Then, the betweenness centrality method used in social network analysis to extract the variables that are closely related to the observations is applied. Social network analysis is a theory in Sociology, and it is the mapping and measuring of relationships and flows between people, groups, organizations, animals, computers or other information/knowledge processing entities. The nodes in the network represent people and groups, and this means that the most active people have the most relationships (=links) with many other people[14]. That is, the entire network is closely related to this node.

   In this experiment, the betweenness centrality value of each observation was computed, then the observation with the highest value was found and genes that were the most closely related were extracted.
2. From the previous experiment, the observation with the highest betweenness centrality value and the genes that were the most closely related to the observation were set aside, and the betweenness centrality value for the remaining observations and genes are computed again. From the resulting values, the observation with the highest betweenness centrality value and the genes that were the most closely related to the observation are clustered.
3. In the same manner, the betweenness centrality value is computed repeatedly as many times as the number of observations, in order to form clusters according to the relations between observations and variables.
4. And finally, one cluster is formed with variables that are not included in any other observation.

The methods mentioned above are shown in figure 4, when applied to a phylogenetic profile data set. In the case of phylogenetic profiles data, 24 species (=observations) and 2,465 genes (=variables) are formed, and the method is repeated to extract the genes that are the most closely related in the order of the observation with the highest betweenness centrality value. And finally, one cluster is formed with genes that are not included in any other species.

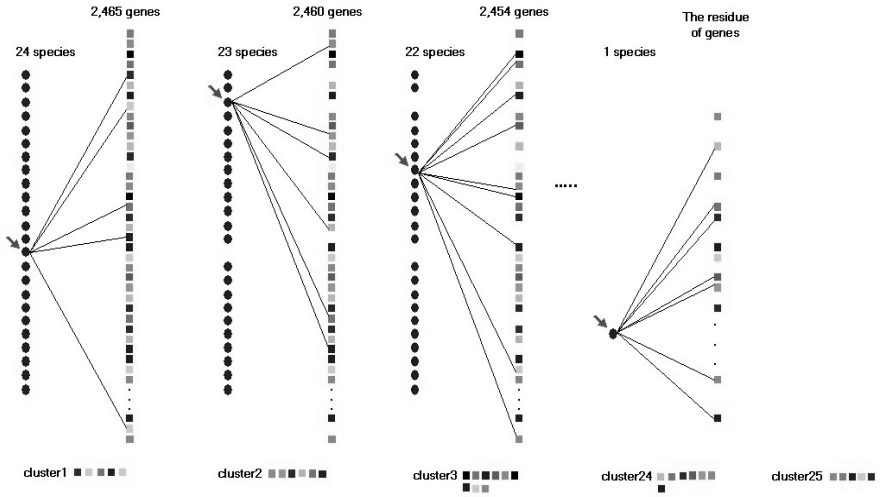


Fig. 4. Application method of betweenness centrality based on observation

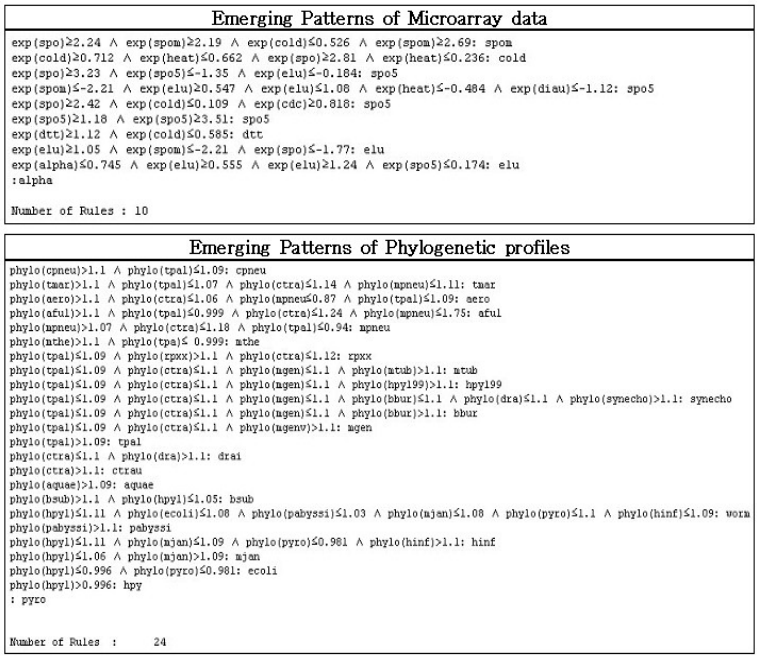
3.3 Forming Emerging Patterns Between Individual Datasets

As shown in figure 4, the betweenness centrality value of each observation is computed for the experimental data in section 3.1. Then clusters are formed by extracting variables (=genes) that could explain the observations with the highest betweenness centrality value. As a result, the microarray data of yeast applied to the first experiment are clustered by reducing 79-elements time points to 10 observations (the clusters were formed from 10 samples composed of 79-elements time points, while the experiment handles observations according to sample number and not time point. See figure 2). In the second phylogenetic profiles experimental data, the 24 species corresponding to the columns are regarded as observations, and clusters are formed in as many number as observations. In this paper, EP formed in microarray data sets and phylogenetic profile data are represented in the following way. EP in microarray data sets and phylogenetic profiles are expressed in the form of $exp(X_1) > a_1 \bigwedge exp(X_2) < a_2$ and $phylo(Y_1) > b_1 \bigwedge phylo(Y_2) > b_2$, respectively. In each representation of EP, X_i is the measured expression level of observations in microarray data sets and Y_j is the sequence similarity of observations in phylogenetic profiles data. The a_i and b_j in the representations are boundary constants that can be inferred from each data set, and they represent the threshold value of the expression level in microarray data and the sequence similarity in phylogenetic profiles data.

4 Experimental Results

In this paper, R package was used to compute betweenness centrality and Weka algorithm was applied to make classification rules. The results are shown in figure 5, where 10 rules are made for the microarray data set and 24 rules are

made for the phylogenetic profiles. We can confirm that 6 out of 10 samples are applied in making EP of the microarray data and all 24 observations are applied to the classification rules for making EP of the phylogenetic profiles data.



Emerging Patterns of Microarray data

```
exp(spo)≥2.24 ∧ exp(spom)≥2.19 ∧ exp(cold)≤0.526 ∧ exp(spom)≥2.69: spom
exp(cold)≥0.712 ∧ exp(heat)≤0.662 ∧ exp(spo)≥2.81 ∧ exp(heat)≤0.236: cold
exp(spo)≥3.23 ∧ exp(spo5)≤-1.35 ∧ exp(elu)≤-0.184: spo5
exp(spom)≤-2.21 ∧ exp(elu)≥0.547 ∧ exp(elu)≤1.08 ∧ exp(heat)≤-0.484 ∧ exp(diau)≤-1.12: spo5
exp(spo)≥2.42 ∧ exp(cold)≤0.109 ∧ exp(cdc)≥0.818: spo5
exp(spo5)≥1.18 ∧ exp(spo5)≥3.51: spo5
exp(dtt)≥1.12 ∧ exp(cold)≤0.585: dtt
exp(elu)≥1.05 ∧ exp(spom)≤-2.21 ∧ exp(spo)≤-1.77: elu
exp(alpha)≤0.745 ∧ exp(elu)≥0.555 ∧ exp(elu)≥1.24 ∧ exp(spo5)≤0.174: elu
:alpha

Number of Rules : 10
```

Emerging Patterns of Phylogenetic profiles

```
phylo(cpneu)>1.1 ∧ phylo(tpal)≤1.09: cpneu
phylo(tmar)>1.1 ∧ phylo(tpal)≤1.07 ∧ phylo(ctra)≤1.14 ∧ phylo(mpneu)≤1.11: tmar
phylo(aero)>1.1 ∧ phylo(ctra)≤1.06 ∧ phylo(mpneu≤0.87 ∧ phylo(tpal)≤1.09: aero
phylo(aful)>1.1 ∧ phylo(tpal)≤0.999 ∧ phylo(ctra)≤1.24 ∧ phylo(mpneu)≤1.75: aful
phylo(mpneu)>1.07 ∧ phylo(ctra)≤1.18 ∧ phylo(tpal)≤0.94: mpneu
phylo(mthe)>1.1 ∧ phylo(tpa)≤ 0.999: mthe
phylo(tpal)≤1.09 ∧ phylo(rpxx)>1.1 ∧ phylo(ctra)≤1.12: rpxx
phylo(tpal)≤1.09 ∧ phylo(ctra)≤1.1 ∧ phylo(mgen)≤1.1 ∧ phylo(mtub)>1.1: mtub
phylo(tpal)≤1.09 ∧ phylo(ctra)≤1.1 ∧ phylo(mgen)≤1.1 ∧ phylo(hpyl99)>1.1: hpyl99
phylo(tpal)≤1.09 ∧ phylo(ctra)≤1.1 ∧ phylo(mgen)≤1.1 ∧ phylo(bbur)≤1.1 ∧ phylo(dra)≤1.1 ∧ phylo(synecho)>1.1: synecho
phylo(tpal)≤1.09 ∧ phylo(ctra)≤1.1 ∧ phylo(mgen)≤1.1 ∧ phylo(bbur)>1.1: bbur
phylo(tpal)≤1.09 ∧ phylo(ctra)≤1.1 ∧ phylo(mgenv)>1.1: mgen
phylo(tpal)>1.09: tpal
phylo(ctra)≤1.1 ∧ phylo(dra)>1.1: drai
phylo(ctra)>1.1: ctrau
phylo(aquae)>1.09: aquae
phylo(bsub)>1.1 ∧ phylo(hpyl)≤1.05: bsub
phylo(hpyl)≤1.11 ∧ phylo(ecoli)≤1.08 ∧ phylo(pabyssi)≤1.03 ∧ phylo(mjan)≤1.08 ∧ phylo(pyro)≤1.1 ∧ phylo(hinf)≤1.09: worm
phylo(pabyssi)>1.1: pabyssi
phylo(hpyl)≤1.11 ∧ phylo(mjan)≤1.09 ∧ phylo(pyro)≤0.981 ∧ phylo(hinf)>1.1: hinf
phylo(hpyl)≤1.06 ∧ phylo(mjan)>1.09: mjan
phylo(hpyl)≤0.996 ∧ phylo(pyro)≤0.981: ecoli
phylo(hpyl)>0.996: hpy
: pyro


Number of Rules  :     24
```

Fig. 5. Emerging patterns of microarray data and phylogenetic profiles

The results in figure 5 can be interpreted as follows: The EP in the 7th line are in the form $exp(dtt) \geq 1.12 \bigwedge exp(cold) \leq 0.585$, and this means that the variables with *dtt* gene expression levels greater than 1.12 and cold values less than 0.585 for $'dtt'$ observation in the entire microarray data set can classify the $'dtt'$ observations in the entire microarray data set. Also, these EP can be considered as classifiers that can be classified among other observations in the microarray data set. The *alpha* in the last line of the EP of the microarray data shows that the genes that can explain the $'alpha'$ observation are those that do not correspond to any of the above rules. The results of the phylogenetic profile can be interpreted in the same way, where the EP of the first line is in the form of $phylo(cpneu) > 1.1 \bigwedge phylo(tpal) \leq 1.09$, and this becomes the classifier that can classify the $'cpneu'$ observation in the phylogenetic profile data set.

Validation results of the EP, as to how accurately they can classify the two types of data sets, show that accuracy is 86.76% and 97.79% for microarray data and phylogenetic profile data, respectively. The relatively low accuracy for the microarray data set could be explained by the reduction of 79 time points to 10 observations before the start of the experiment.

5 Conclusions and Future Works

Typical bio-data analysis methods deal directly with genes while ignoring biological attributes, but since the interaction among genes plays an important role in bio-data analysis, new methods must be developed. Also, multi-source data classification and analysis problems are much more complex and have more factors to be considered than single-source data problems. When handling bio-data, disparate types of multi-source data can be made based on the same variables, and we are in need of classifiers that can classify the data sets and methods to easily understand the features of the data sets. Therefore, this paper proposes a new method that considers the characteristics of bio-data, and while existing methods ignore biological attributes and analyze only the genes, the proposed method provides an analysis method based on observations using all variables from each data set. This method makes EP that take into account the relations between genes in the data set and the results are applied to the multi-source data classification. An existing paper introduced a method to map variables to gene function categories by applying the SVM method using the same data set in this paper[14]. But the method introduced in the existing paper differs from the proposed method, which considers both observations and variables, in that the existing method has no regard of the interaction structure between genes in the analysis stage, that it is not easy to interpret and that the analysis is done after variables are removed first by some threshold value in the preprocessing stage.

The experimental methods introduced in this paper suggest several avenues that can be taken for future research. One direction would be to find a better classifier of multi-source data in bio-data. Another direction would be, since only two biological data types were used for multi-source data classification, to include multiple biological data types for discovering EP and for extending the proposed method in multi-source data classification. Also, another important task would be to come up with a theoretically and experimentally justified verification of disparate data.

References

1. Barabasi, A.L.: Link, Penguin,(2003)
2. Boulesteix, A.L., Tutz, G., Strimmer, K.: A CART-based approach to discover emerging patterns in microarray data. Bioinformatics, **19** (2003) 2465–2472
3. Boser, B.E., Guyon, I.M., Vapnik, V.: A training algorithm for optimal margin classifiers. Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, **5** (1992) 144–152
4. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, **2** (1998) 121–167
5. Dong, G., Li, J.: Efficient mining of emerging patterns: discovering trends and differences. Proceedings of the SIGKDD (5th ACM International Conference on Knowledge Discovery and Data Mining), **5** (1999) 43–52
6. DeRisi, J., Iyer, V., Brown, P.: Exploring the metabolic and genetic control of gene expression on a genomic scale. Science, **278** (1997) 680–686

7. Li, J., Wong, L.: Emerging patterns and gene expression data. Proceedings of 12th Workshop on Genome Informatics, **12** (2001) 3–13
8. Xia, L., Shaoqi, R., Yadong, W., Binsheng, G.: Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling. Nucleic Acids Research, **32** (2004) 2685–2694
9. Eisen, M., Spellman, P., Brown, P., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America, **95** (1998) 14863–14868
10. Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C., Furey, T., Ares, J.M., Haussler, D.: Knowledge-base analysis of microarray gene expression data using support vector machines. Proceedings of the National Academy of Science of the United States of America, **97** (2000) 262–267
11. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O.: Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proceedings of the National Academy of Sciences of the United States of America, **96** (1999) 4285–4288
12. West, M., Nevins, J.R., Spang, R., Zuzan, H.: Bayesian regression analysis in the 'large p, small n' paradigm with application in DNA microarray studies. Technical Report 15, Institute of Statistics and Decision Sciences, Duke University,(2000)
13. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines. Cambridge UP,(2000)
14. Pavlidis, P., Weston, J., Cai, J., Grundy, W.N.: Learning gene functional classifications from multiple data types. Journal of Computational Biology, **9** (2002) 401–411
15. Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.Q., Botstein, D., Futcher, B.: Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Molecular Biology of the Cell, **9** (1998) 3273–3297
16. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E., Golub, T.: Interpreting patterns of gene expression with self-organizing maps. Proceedings of the National Academy of Sciences of the United States of America, **96** (1999) 2907–2912
17. Chu, S., DeiRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P., Herskowitz, I.: The transcriptional program of sporulation in budding yeast. Science, **282** (1998) 699–705
18. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research, **25** (1997) 3389–3402
19. Larray, T. H. Yu., Fu-lai, C., Stephen, C.F.: Using Emerging Pattern Based Projected Clustering and Gene Exptression Data for Cancer Detection. Proceedings of the Asia-Pacific Bioinformatics Conference, **29** (2004) 75–87
20. Yuhang, W., Filla M, M.: Application of Relief-F Feature Filtering Algorithm to Selecting Informative Genes for Cancer Classification Using Microarray Data. International IEEE Computer Society Computational Systems Bioinformatics Conference, **3** (2004) 497–498