A linear programming model for identifying non-redundant biomarkers based on gene expression profiles

Xianwen Ren*, Yong Wang†, Luonan Chen‡ and Xiang-Sun Zhang†
*State Key Laboratory for Molecular Virology and Genetic Engineering,
Institute of Pathogen Biology,
Chinese Academy Medical Sciences and Peking Union Medical College, Beijing, 100730, China
Email: renxwise@gmail.com
†Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, China
Email: ywang@amss.ac.cn, zxs@amt.ac.cn
‡Key Laboratory of Systems Biology, SIBS-Novo Nordisk Translational Research Centre for PreDiabetes,
Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200233,China;
Collaborative Research Center for Innovative Mathematical Modelling, Institute of Industrial Science,
University of Tokyo, Tokyo 153-8505, Japan
Email: lnchen@sibs.ac.cn

***Abstract*—With the development of high-throughput technologies, e.g. microarrays and the second generation sequencing technologies, gene expression profiles have been applied widely to characterize the functional states of various samples at different conditions. This is especially important for clinical biomarker identification that is vital to the understanding of the pathogenesis of a certain disease and the subsequent therapies. Because of the complexity of multi-gene disorders, a single biomarker or a set of separate biomarkers often fails to discriminate the samples correctly. Moreover, biomarker identification and class assignment of diseases are intrinsically linked while the current solutions to these two tasks are generally separated. Motivated by these issues, we give out a novel model based on linear programming in this study to simultaneously identify the most meaningful biomarkers and classify accurately the disease types for patients. Results on a few real data sets suggest the effectiveness and advantages of our method.**

I. Introduction

Identification of biomarkers that can indicate a specific biological state of samples is an important topic in biomedical research because it can provide insightful clues into the pathogenesis of a certain disease and important indices for accurate diagnosis. With the development of high-throughput technologies, e.g. microarrays and the second generation sequencing technologies, thousands of genes can be measured simultaneously. How to select the most meaningful biomarkers from the large number of genes forms a common question that clinicians often come across.

The most straightforward method for identifying biomarkers is calculating the fold-changes of gene expressions in different classes of samples, given that the gene expression data is used to characterize the biological states. The larger the fold-change is, the more likely the gene is a biomarker. However, this method does not consider the variations among samples of the same classes. So the methods based on or similar to the student t-test or Wilcoxon rank-sum test are introduced to eliminate the irrelevant or noisy features [1], [14]. Due to the multiple testing issues, methods such as SAM that provide fine false discovery rate (FDR) control are invented [16]. Generally, all these methods can generate many biomarkers that are redundant. Peng et. al. propose a criterion based on mutual information to find a set of biomarkers that have the maximal relevancy to the class labels but minimal redundancy within themselves [11].

In nature, biomarker identification is intrinsically linked to class assignment to samples [3], [10], [13]. From a machine learning view, biomarker identification is a feature selection problem given the biological states (e.g. disease or normal) of samples. The aim of feature selection is to find a set of features that can maximize the prediction of accuracy of a classifier [8]. With different classifiers, the identified biomarkers may be different. Many supervised or semi-supervised machine learning method, such as support vector machine and Bayesian networks, can be exploited as the classifiers to guide the identification of biomarkers [1], [4], [5], [7], [12], [19]. However, biomarker identification is not explicitly embedded in these methods. A model for simultaneous biomarker identification, especially non-redundant biomarker identification, and classification is needed to explicitly model the properties of biological states.

In this study, we explicitly consider the properties of normal and disease biological states and propose a novel model based on linear programming. The model can simultaneously identify biomarkers from thousands of candidates and classify samples based on the identified biomarkers. Different from the general biomarker identification approaches, it produces a set of non-redundant but complementary biomarkers that maintain the maximal classifying power. It is flexible. It

2011 IEEE International Conference on Systems Biology (ISB)
978-1-4577-1666-9/11/$26.00 ©2011 IEEE

Zhuhai, China, September 2–4, 2011

can deal with two classes, multiple classes with order and multiple classes without order. We tested it and compared it to the classic biomarker identification method and classifying method on real gene expression data sets. The results highlight its effectiveness.

II. METHODS

A. The heuristics we used to build our model

We argue that a stable biological state is characterized by the expressions of a set of genes rather than all genes. The expressions of this set of genes form the basis that the biological state is different from others. Assuming the expressions of all genes form a high-dimensional space, one biological state may correspond to a local compact region in the space (1). We use an ellipsoid to model the compact region and try to make sure that samples with that biological state localize in the inner part of the ellipsoid whereas other samples without that biological state localize in the outer part of the ellipsoid. We seek those ellipsoids that have low dimensions and discriminate those samples with different biological states.

B. The formulations

Based on the arguments above, we proposed the following model to model the biological states explicitly based on their gene expressions. Let us consider the two-class cases first.

Given two biological states (denoted by a and b), we assume that there are totally m genes to describe their status. There are n_a and n_b samples with states a and b, respectively. We try to find a minimal gene set that can discriminate the samples with different biological states maximally. It can be formulated as follows:

$$\min \sum_{i=1}^{n} w_i + \alpha * (z_1^a - z_2^a + z_1^b - z_2^b) \tag{1}$$

Subject to

$$\sum_{i=1}^{m} w_i (x_{ij} - x_i^a)^2 \le z_1^a \quad \text{for} \quad j \in I^a \tag{2}$$

$$\sum_{i=1}^{m} w_i (x_{ij} - x_i^a)^2 \ge z_2^a \quad \text{for} \quad j \in I^b \tag{3}$$

$$\sum_{i=1}^{m} w_i (x_{ij} - x_i^b)^2 \le z_1^b \quad \text{for} \quad j \in I^b \tag{4}$$

$$\sum_{i=1}^{m} w_i (x_{ij} - x_i^b)^2 \ge z_2^a \quad \text{for} \quad j \in I^a \tag{5}$$

$$0 \le z_1^a \le z_2^a \tag{6}$$

$$0 \le z_1^b \le z_2^b \tag{7}$$

$$0 \le w_i \le 1 \quad \text{for} \quad i \in \{1, \cdots, m\} \tag{8}$$

Where $w_i, i \in \{1, \cdots, m\}$, z_1^a, z_2^a, z_1^b and z_2^b are variables, α is a parameter and x_{ij} is the ith feature of the jth sample in I^a or I^b. x_i^a and x_i^b are the mean values of the ith features of class a and b, respectively. Constraint (2) requires the samples of class a belong to a compact ellipsoid of radius $\sqrt{z_1^a}$. Constraint (3) describes that the samples of class b localize outside the ellipsoid of class a with a distance at least $\sqrt{z_2^a}$. So do the constraints (4) and (5). Constraints (6) and (7) define the inner and outer parts of the ellipsoids. Constraint (8) confines $w_i, i \in \{1, \cdots, m\}$ in the range of zero and one. The objective (1) tries to find a minimal set of genes that can maximize the distances between the two ellipsoids and minimize their volume.

If the samples of two classes are non-separable, a formulism like the soft margin support vector machine (SVM) is proposed to tolerate the training errors as follows:

$$\min \sum_{i=1}^{n} w_i + \alpha * (z_1^a - z_2^a + z_1^b - z_2^b) + C \sum_{j=1}^{n} (\xi_j^1 + \xi_j^2) \tag{9}$$

Subject to

$$\sum_{i=1}^{m} w_i (x_{ij} - x_i^a)^2 \le z_1^a + \xi_j^1 \quad \text{for} \quad j \in I^a \tag{10}$$

$$\sum_{i=1}^{m} w_i (x_{ij} - x_i^a)^2 \ge z_2^a - \xi_j^2 \quad \text{for} \quad j \in I^b \tag{11}$$

$$\sum_{i=1}^{m} w_i (x_{ij} - x_i^b)^2 \le z_1^b + \xi_j^1 \quad \text{for} \quad j \in I^b \tag{12}$$

$$\sum_{i=1}^{m} w_i (x_{ij} - x_i^b)^2 \ge z_2^a - \xi_j^2 \quad \text{for} \quad j \in I^a \tag{13}$$

$$0 \le z_1^a \le z_2^a \tag{14}$$

$$0 \le z_1^b \le z_2^b \tag{15}$$

$$0 \le w_i \le 1 \quad \text{for} \quad i \in \{1, \cdots, m\} \tag{16}$$

$$0 \le \xi_j^1 \quad \text{for} \quad j \in \{1, \cdots, n\} \tag{17}$$

$$0 \le \xi_j^2 \quad \text{for} \quad i \in \{1, \cdots, n\} \tag{18}$$

Where $\xi_j^1, j \in \{1, \cdots, n\}$ represents the bias of sample j from the inner part of the ellipsoid of its class and $\xi_j^2, j \in \{1, \cdots, n\}$ denotes the bias of sample j from the outer part of the ellipsoid of the class it does not belong to. C is a parameter to tune how the training errors are punished.

Here we model each of the biological state as an ellipsoid in one linear programming framework. In fact, like twin-SVM [9], if only one biological state is modeled, the model above can still work for biomarker identification and class assignment. That is, model that is composed of (1), (2), (3), (6) and (8) still works. This is especially useful when one biological state is very heterogeneous.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

C. Adaptations to multiple classes

Because each biological state is modeled by an ellipsoid, it is very easy to extend the model to those cases where multiple classes are available. If the multiple classes are independent, the extension can be implemented by adding one ellipsoid for each state. If the multiple classes are ordered, ellipsoids can be added for each binary partition of the multiple classes. For example, given three classes in order, Class One, Class Two and Class Three, we can set one ellipsoid for Class One, one ellipsoid for Class One and Class Two and one ellipsoid for Class Three such that all classes can be discriminated from each other. Below we give out the formulations for these two cases.

Given n samples of c classes without order that are described by m features, the biomarker identification and classification framework based on ellipsoids are described as follows:

$$\min \sum_{i=1}^{m} w_i + \alpha \sum_{i=1}^{c} (z_1^i - z_2^i) + C \sum_{i=1}^{n} (\xi_i^1 + \xi_i^2) \tag{19}$$

Subject to

$$\sum_{i=1}^{m} w_i (x_{ij} - x_i^a)^2 \le z_1^a + \xi_j^1 \quad \text{for} \quad j \in I^a, a \in \{1, \cdots, c\} \tag{20}$$

$$\sum_{i=1}^{m} w_i (x_{ij} - x_i^a)^2 \ge z_2^a - \xi_j^2 \quad \text{for} \quad j \notin I^a, a \in \{1, \cdots, c\} \tag{21}$$

$$0 \le z_1^a \le z_2^a \quad \text{for} \quad a \in \{1, \cdots, c\} \tag{22}$$

$$0 \le w_i \le 1 \quad \text{for} \quad i \in \{1, \cdots, m\} \tag{23}$$

$$\xi_i^1 \ge 0 \quad \text{for} \quad i \in \{1, \cdots, n\} \tag{24}$$

$$\xi_i^2 \ge 0 \quad \text{for} \quad i \in \{1, \cdots, n\} \tag{25}$$

Where z_1^a and z_2^a defines the inner and outer radius of the ellipsoid representing class a. ξ_i^1 and ξ_i^2 are slack variables to tolerate the training errors.

Given n samples of c classes with order that are described by m features and assuming that the c classes have been ordered from 1 to c, the ellipsoid-based framework for biomarker identification and sample classification is given as follows:

$$\min \sum_{i=1}^{m} w_i + \alpha \sum_{i=1}^{c} (z_1^i - z_2^i) + C \sum_{i=1}^{n} (\xi_i^1 + \xi_i^2) \tag{26}$$

Subject to

$$\sum_{i=1}^{m} w_i (x_{ij} - x_i^a)^2 \le z_1^a + \xi_j^1 \quad \text{for} \quad j \in J^a \tag{27}$$

$$\sum_{i=1}^{m} w_i (x_{ij} - x_i^a)^2 \ge z_2^a - \xi_j^2 \quad \text{for} \quad j \notin J^a \tag{28}$$

$$\sum_{i=1}^{m} w_i (x_{ij} - x_i^a)^2 \le z_1^a + \xi_j^1 \quad \text{for} \quad j \in I^c \tag{29}$$

$$\sum_{i=1}^{m} w_i (x_{ij} - x_i^a)^2 \ge z_2^a - \xi_j^2 \quad \text{for} \quad j \notin I^c \tag{30}$$

$$0 \le z_1^a \le z_2^a \quad \text{for} \quad a \in \{1, \cdots, c\} \tag{31}$$

$$0 \le w_i \le 1 \quad \text{for} \quad i \in \{1, \cdots, m\} \tag{32}$$

$$\xi_i^1 \ge 0 \quad \text{for} \quad i \in \{1, \cdots, n\} \tag{33}$$

$$\xi_i^2 \ge 0 \quad \text{for} \quad i \in \{1, \cdots, n\} \tag{34}$$

Where $J^a = \bigcup_{i=1}^{a} I^i, a \in \{1, \cdots, c-1\}$.

D. Data and Evaluation methods

Here we use two real data sets to demonstrate the effectiveness and features of our models in both biomarker identification and sample classification. One data set is the classical Fisher's Irish data which is used frequently in classification and clustering [15]. There are three classes in this data set. Class One can be separated linearly from Class Two and Class Three whereas Class Two and Class Three can not be classified by a hyperplane in the input space. We use it to show that our model has the same accuracy as the state-of-the-art machine learning algorithms such as support vector machines [2], [17], [18]. The second data set is about seventy-two acute leukemia patients in which there are two types of samples [6]. One type is acute myeloid leukemia (AML) and the other is acute lymphoblastic leukemia (ALL). ALLs can be further classified into T cell ALLs and B cell ALLs. The gene expressions of these samples were profiled by microarrays. So thousands of genes are measured and identifying biomarkers is necessary to reduce the complexity of the future measurement and to deepen the understanding of the pathogenesis of these diseases. We use this data set to show how a non-redundant set of biomarkers is identified whereas the power for sample classification is maintained. We use five-fold cross-validation to evaluate our model and compare it to the state-of-the-art classifiers and biomarker-identification method.

III. Results

A. On Iris data

We compared our method with support vector machines coupled with the Gaussian kernel which has been shown dominant performances in either linearly separable or nonlinearly separable applications. By setting the gamma parameter of the support vector machines with the Gaussian kernel as one, the classifier can classify the samples with 99% mean accuracy (100 five-fold cross validations) in which Class One is the positive class and Class Two and Class Three are the negative class. The mean accuracy of our method in 100 five-fold cross validations also reaches 99%. This high accuracy is because Class One can be separated linearly from Class Two and Class Three. The mean accuracy of support vector machines to classify Class Two from Class Three is about 94% (100 five-fold cross validations) where gamma is set to one. The accuracy of our method also reaches 94%, suggesting the competitive performance of our method to the state-of-the-art classifiers in sample classification. Because there are only

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

four features in Irish data, the performance in feature selection (biomarker identification) is not evaluated.

B. On AML and ALL data

Because of the inherent noise in the microarray data, we first preprocessed the raw gene expression data of 6817 genes by filtering those non-informative genes with $\max - \min \leq 500$ or $\max / \min \leq 5$ where $\max$ and $\min$ mean the maximum and minimum gene expression values in all the samples, respectively. Finally 1751 informative genes were retained to do the subsequent analysis.

We first tested the performance of support vector machines with the Gaussian kernels to distinguish AMLs from ALLs. We used the svmtrain and svmclassify functions in Matlab and found that the mean accuracy of support vector machines with the Gaussian kernels (gamma=50, selected by grid search) was about 91%. The accuracy of linear support vector machines is about 98.06%. Then we tested the accuracy of our method to discriminate AMLs from ALLs and the result suggested that the mean accuracy of our method can reach 99.14%. This suggests again that our method have competitive performance to the state-of-the-art classifiers.

Further, we compared the performances of sample classification of both methods on the data subset of T cell ALLs and B cell ALLs. The accuracy of support vector machines with Gaussian kernels (gamma=50, selected by grid search) is about 81.68%. The accuracy of linear support vector machines is 96.68% whereas our methods can reach the mean accuracy of 97.11%, suggesting again that our method is competitive to the state-of-the-art methods when classification is the task.

Given the competitive classification power of our method, we will show its performance for identifying non-redundant biomarkers. To discriminate AMLs from ALLs, our method identified nineteen probes as a candidate biomarker set. Upon this set of biomarkers, we applied the hierarchical clustering method to group the samples (clustergram in Matlab R2010b). The result suggests that the AMLs can all be distinguished from ALLs accurately (2). The mean accuracy of 100 five-fold cross-validations upon these biomarkers reaches 99.93%.

If t-test is used to identify biomarkers distinguishing AMLs from ALLs, eighty-three probes were selected (alpha: 0.05, Bonferroni correction). Upon this set of biomarkers, hierarchical clustering (clustergram in Matlab R2010b) grouped ALLs from all except one AML (3). The mean accuracy of 100 five-fold cross-validations upon these biomarkers is 98.47%, lower than the biomarker set identified by our method significantly (p-value $<$ 3.35e-71, student t-test).

The same comparison was conducted to identify biomarkers that discriminate T cell ALLs from B cell ALLs. When t-test is used, twenty-three probes were selected (alpha:0.05, Bonferroni correction for the multiple testing problem). Upon this set of biomarkers, hierarchical clustering clustered all except one T cell ALLs correctly from the B cell ALLs (5). The mean accuracy of 100 five-fold cross-validations upon these biomarkers approaches 96.04%. Our method identified twenty-one probes as a biomarker set. Hierarchical clustering

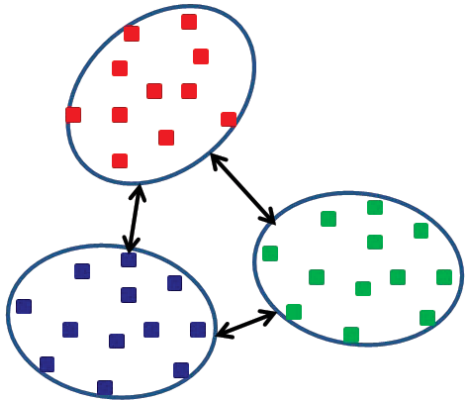

Fig. 1. The idea behind our method is to minimize the volumes of each ellipsoid and to maximize the distances among ellipsoids.

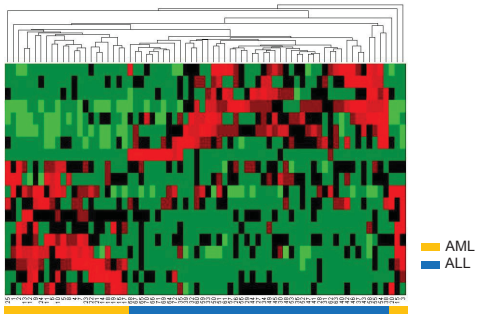

Fig. 2. Hierarchical clustering of AMLs and ALLs based on biomarkers identified by our method.

based on this probe set discriminate all T cell ALLs correctly from B cell ALLs (4). The mean accuracy of 100 five-fold cross-validations upon these biomarkers reaches 96.30%.

We compared the biomarker sets identified by t-test and our method. For AMLs and ALLs, there are nine probes overlapped between the two biomarker sets. The mean correlation coefficient of our biomarker set is significantly less than that of biomarkers identified by t-test (6). For T cell ALLs and B cell ALLs, there are six probes shared in the two biomarker sets. The expression profiles of probes identified by our method are less correlated significantly to each other compared to those probes identified by t-test (7).

We further applied our method to identify biomarkers that can discriminate ALL, T cell ALL and B cell ALL simultaneously. A total of twenty-four probes were selected and hierarchical clustering (clustergram in Matlab R2010b) upon this set of biomarkers suggests their discriminative effectiveness (8).

IV. Discussions and Conclusion

Biomarker identification and sample classification is important in current biomedical researches and clinical practice. Although there are many methods available for biomarker identification and sample classification, few can simultaneously identify biomarkers and classify samples. Further, non-redundant biomarkers have not been paid much attention, limiting the inclusion of those biomarkers that have weak

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

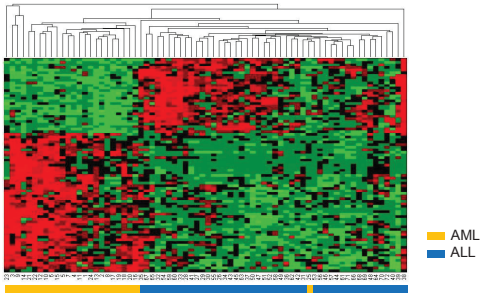

Fig. 3. Hierarchical clustering of AMLs and ALLs based on biomarkers identified by t-test.

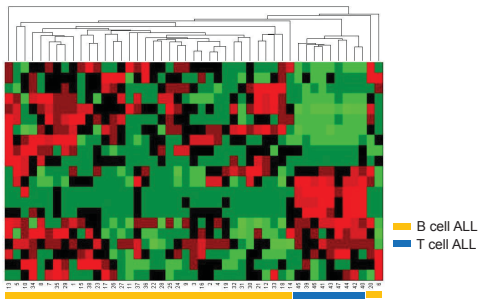

Fig. 4. Hierarchical clustering of B cell ALLs and T cell ALLs based on biomarkers identified by our method.

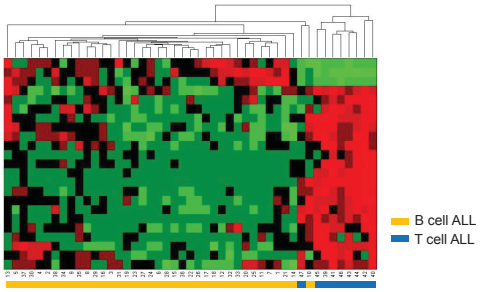

Fig. 5. Hierarchical clustering of B cell ALLs and T cell ALLs based on biomarkers identified by t-test.

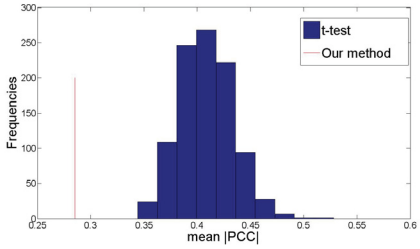

Fig. 6. Biomarkers for AMLs and ALLs identified by our method are less correlated than those identified by t-test.

correlations with a specific biological state. In this study, we use ellipsoids to model biological states and try to identify non-redundant complementary biomarkers and to classify samples simultaneously. The idea is formulated as a linear programming problem that can be solved easily and be applied

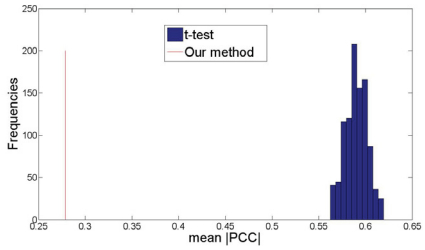

Fig. 7. Biomarkers for B cell ALLs and T cell ALLs identified by our method are less correlated than those identified by t-test.

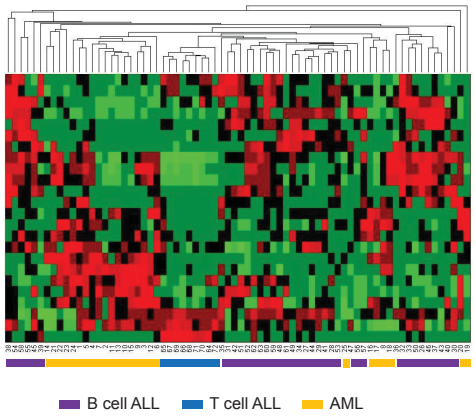

Fig. 8. Hierarchical clustering of AMLs, B cell ALLs and T cell ALLs based on biomarkers identified by our method.

to very large scale data sets. The Iris data set suggests that our method has competitive performance with the state-of-the-art classifiers (support vector machines with Gaussian kernels) no matter the samples are separated linearly or not. The leukemia gene expression data set suggests not only the dominant performance of our method in classification but also the power in identifying biomarkers. The biomarkers identified by our method are less redundant and more predictive than those identified by the classical methods.

ACKNOWLEDGMENT

The authors would like to thank members of Zhangroup for their valuable discussions.

REFERENCES

[1] Baldi, P. and Long, A. D. (2001). A bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics*, **17**(6), 509–519.

[2] Boser, B., Guyon, I., and Vapnik, V. (1992) A training algorithm for optimal margin classifiers. *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*,144–152.

[3] Buturovic, L. J. (2006). Pcp: a program for supervised classification of gene expression profiles. *Bioinformatics*, **22**(2), 245–247.

[4] Fox, R. J. and Dimmic, M. W. (2006). A two-sample bayesian t-test for microarray data. *BMC Bioinformatics*, **7**, 126–126.

[5] Gevaert, O., Smet, F. D., Timmerman, D., Moreau, Y., and Moor, B. D. (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics*, **22**(14), e184–e190.

[6] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., *et al*. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**(5439), 531–537.
[7] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**(1), 389–422.
[8] Hwang, T., Sicotte, H., Tian, Z., Wu, B., Kocher, J.-P., *et al*. (2008). Robust and efficient identification of biomarkers by classifying features on graphs. *Bioinformatics*, **24**(18), 2023–2029.
[9] Jayadeva, Khemchandani, R., and Chandra, S. (2007). Twin support vector machines for pattern classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **29**(5), 905–910.
[10] Li, T., Zhang, C., and Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, **20**(15), 2429–2437.
[11] Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, **27**, 1226–1238.
[12] Ramon, D.-U. and de Andres Sara, A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**, 3–3.
[13] Saeys, Y., Inza, I., and Larra?aga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**(19), 2507–2517.
[14] Sandrine, D., Hwa, Y. Y., Matthew, C. J., and Terence, S. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Stat. Sin*, **12**, 29.
[15] Schwartz, I., Sajin, A., Fisher, I., Neeb, M., Shochina, M., *et al*. (2009). The effectiveness of locomotor therapy using robotic-assisted gait training in subacute stroke patients: a randomized controlled trial. *PM & R : the journal of injury, function, and rehabilitation*, **1**(6), 516–523.
[16] Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, **98**(9), 5116–5121.
[17] Vapnik, V. (1998). *Statistical Learning Theory*. Wiley-Interscience.
[18] Vapnik, V. (1999). *The Nature of Statistical Learning Theory (Information Science and Statistics)*. Springer.
[19] Zhang, H. H., Ahn, J., Lin, X., and Park, C. (2006). Gene selection using support vector machines with non-convex penalty. *Bioinformatics*, **22**(1), 88–95.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE