A strategy for identifying class-separating genes in drug-treatment microarray data

Sampsa Hautaniemi*, Päivikki Kauraniemi†, Pauli Rämö*,
Olli Yli-Harja*, Jaakko Astola*, Anne Kallioniemi†

Abstract

An important application of the microarray technology is to explore the effects of a drug to gene expression patterns on a genome-wide scale. Furthermore, it is important to identify a set of genes whose expressions separate predefined classes. As the time period from the administration of the drug to an effect in gene expression pattern is generally unknown, the drug-treatment data consist of time-series experiments. However, identification of discriminating genes is very difficult, if not impossible, to achieve by direct application of statistical and clustering methods to time-series data. In this study we present a strategy that identifies a set of discriminating genes between two predefined classes. In brief, the suggested strategy includes equalization transformation, regression analysis and a stepwise gene selection algorithm. We have applied the suggested strategy to an experiment where the goal was to evaluate the effect of Herceptin treatment on breast cancer cells and to identify a set of discriminating genes. Based on the preliminary results we conclude that the suggested strategy is applicable to real-life drug treatment studies, where the number of observations can be rather small and the data may contain missing values.

1 Introduction

DNA microarray technology, allowing a rapid large-scale analysis of gene expression, has provoked a real paradigm shift in biology. With DNA microarrays it is possible to measure expression ratios of tens of thousands of genes in different conditions simultaneously. A large amount of information enables various types of studies that concentrate on finding the mechanisms behind the biological phenomena, such as cancer development.

Recently several studies aiming to identify a set of genes whose expression patterns separate the samples into (predefined) classes have been published

*Institute of Signal Processing, Tampere University of Technology, PO Box 553, 33101, Tampere, Finland.

†Laboratory of Cancer Genetics, Institute of Medical Technology, University of Tampere and Tampere University Hospital, Finland.

[8], [11]. For example, identification of sets of genes that allow classification of tumor samples according to clinical characteristics, such as patient outcome, has been performed [12]. In these studies, the set of discriminating genes is found by a rather straightforward combination of statistical tests and clustering methods. It is anticipated that the set of discriminating genes would ultimately contain genes that have a crucial role in the disease pathogenesis and, therefore, would represent ideal targets for drug development.

Given that a drug targeting a specific gene product is developed, it is important to explore the effects of the drug on gene expression patterns on a genome-wide scale. The purpose is to identify additional genes or genetic pathways that could further deepen our understanding on the development of disease. One approach is to study the effects of the drug to gene expression patterns in predefined classes of samples and to identify a set of discriminating genes. In essence, the samples are chosen from (usually two) predefined classes, such as responders (those where the drug is showing the expected effect) and non-responders. Because the time period from the administration of the drug to an effect in gene expression pattern is generally unknown, there is a need for time-series experiments that should be designed to be sufficiently long for all possible genes to respond to the stimulus.

Time-series data accompanied with the purpose of identifying a set of discriminating genes between the classes present substantial problems to the data analysis. Direct application of clustering or statistical methods is not feasible because they will identify genes having similar expression levels across the time-series, which is uncalled for.

In this study we present a strategy that identifies a group of genes that separate two predefined classes based on expression patterns. In order to achieve this goal we have utilized an equalization transformation method, a regression analysis and a stepwise gene selection approach. We also provide some general guidelines for applying the suggested strategy to real-life experiments through a case study where we have identified a set of discriminating genes in breast cancer cell lines treated with Herceptin.

2 Drug-treatment data

We assume that there are two predetermined classes; A and B, consisting of n_A and n_B samples, respectively. Classes are chosen according to research hypothesis. For example, the first class (A) could include samples where the drug induces an expected effect, while the second group (B) could contain samples that do not respond in an expected fashion. From each sample there have been made time-series experiments, which are referred to as *cases*. In Figure 1 we have illustrated the hierarchy of the drug treatment data.

The number of samples in class A (n_A) may be different from the number of samples in class B (n_B). Usually the length of time-series is equivalent for

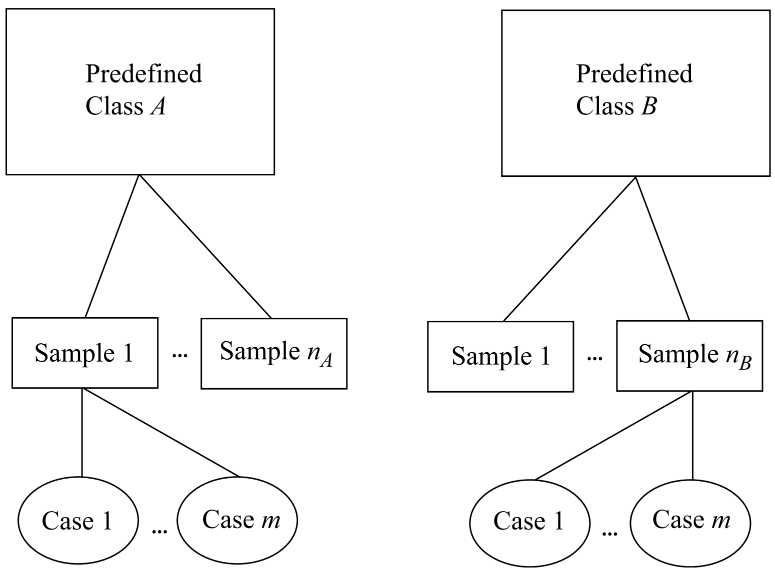

Figure 1: Schematics of the data in drug treatment experiment. Two predefined classes (A and B) contain n_B and n_B samples, respectively. Samples contain m time-series measurements, which are referred as cases.

each sample (m) so there are $N = n_A \cdot m + n_B \cdot m$ microarray hybridizations altogether. However, sometimes it may happen that a hybridization fails and some samples have less than m cases. A failure in microarray experiment may remain undetected until the data analysis phase, when it is too late to make a substituting experiment. Therefore it is essential that the strategy applied to the drug treatment data analysis is capable to deal with missing values.

3 Equalization transformation

Samples in a predetermined class are combined and compared to samples assigned to another class, so it is essential that at least the minimum and the maximum are equal across the cases. Further, it is often desired that the cases have equal mean, equal standard deviation and are normally distributed in order to utilize statistical methods in upstream data analysis. There are many attempts for fulfilling these requirements, for example, logarithmic transformation accompanied with additional data manipulations such as mean/median centering.

In this section we describe the equalization transformation, which is capable of fulfilling the requirements above without additional manipulations. A method similar to the equalization method has been used earlier for oligonucleotide data normalization [1], where it was found to outperform other two normalization methods.

The equalization transformation constructs a data set having a distribution as close as possible to the desired distribution. In other words, one must explicitly determine the desired distribution prior to transformation. In this study we assume that the desired distribution is $\mathcal{N}(0,1)$.

The general form of a transformation is the following.

$$x^{new} = T(x^{old}), \tag{1}$$

where T is a single valued and monotonically increasing function.

When the cumulative distribution function (CDF) of the data is used as the transformation function T_U, it generates the uniform distribution from 0 to 1.

$$T_U(x) = \frac{|\{x_i | x_i < x, i = 1 \dots n\}| + 1}{n+2}, \tag{2}$$

where n is the number of data values. In other words, the x^{old} which lies in the i:th position in the sorted list of data values becomes $x^{new} = (i+1)/(n+2)$. The constants 1 and 2 are added to prevent the transformed values from becoming exactly 0 or 1, which may yield infinite values in the following transformation. The uniform distribution can be further transformed by using the inverse CDF of the desired distribution. Usually the normal distribution is chosen. The inverse of the Gaussian CDF cannot be written analytically and therefore the transformation has to be done using numerical methods. The transformation function $T_G(x)$ which transforms data to the desired distribution is:

$$T_G(x) = T_{CDF^{-1}}(T_U(x)). \tag{3}$$

The idea of the equalization transformation is illustrated in Figure 2. The curve on the left hand side illustrates the CDF^{-1} of the desired distribution (the CDF^{-1} of the normal distribution). On the right hand side the CDF of the measured expression ratios is depicted. In the equalization transformation the transformed ratio depends solely on the corresponding position in the list of sorted expression ratios. The order of the genes remains the same, but the relative ratio values are scaled.

4 Regression based dimension reduction for transformed time-series data

In drug treatment experiments researchers are mainly interested in genes whose expressions increase or decrease as a function of time when the cells are treated with a drug. Therefore, our approach to identify such genes is to apply regression analysis for computing the quantity that reflects the increase or decrease in expressions.

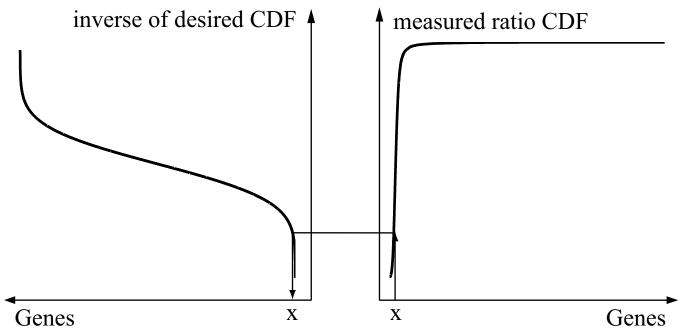

Figure 2: An illustration of the equalization transformation.

Time-series observations are converted to a single number as follows. Assume that we fit a first-order polynomial (line) to the data. Parameters needed for the line are a slope (β_1) and an interception (β_0), which are estimated from the data, for example, using least-square method [3]. Corresponding estimated values for the slope and the interception are $\hat{\beta}_1$ and $\hat{\beta}_0$, respectively. With the estimated parameters it is possible to measure how much an observed value differs from an estimated value, i.e. $r_i = (y_i - \hat{y}_i)^2$, where y_i is observed value and $\hat{y}_i$ is the estimated value. β_1 and r_i:s are combined to a single number for a gene g:

$$T_g = \frac{\hat{\beta}_1}{res + 1}, \tag{4}$$

where $\hat{\beta}_1$ is the estimated slope and $res = \sum_{i=1}^{m} r_i$, and m is the number of cases.

Adding one to denominator in Equation 4 prevents dividing by zero if $res = 0$. Sign of T_g is positive if expressions are increasing and negative if expressions are decreasing. The magnitude of T_g is large if expressions are either increasing or decreasing with a steep slope combined with a small error in fitting. Small magnitudes of T_g reflect genes whose expressions are not increasing or decreasing as function of time or who are unreliable due to aberrant observations in respect to a fitted model.

For example, in Figure 3 we have illustrated transformed expression ratios of gene *v-myb avian myeloblastosis viral oncogene homolog* (*myb*) across three time-series observations with linear regression with corresponding fitting errors and the slope.

It may be worthwhile to utilize other polynomial fittings than linear. Especially, if the first time point is a treated sample, it may be good to utilize zeroth order polynomial fitting. That is, to identify genes whose expressions have gone down instantaneously and stayed there. It is also possible to use other functions than polynomials in fitting. Therefore, so called template-based clustering [7] is possible to perform with the regression analysis as

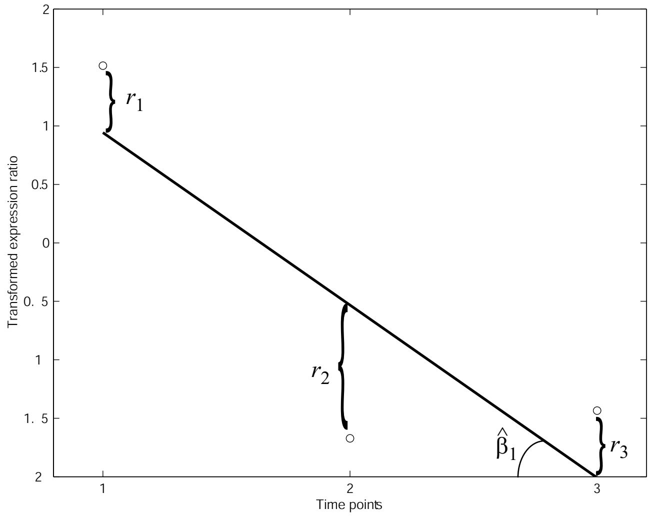

Figure 3: An application of the linear regression to three time-series measurements of gene *myb*. For each observation a fitting error (r_i) is computed. $\hat{\beta}_1$ is slope of the linear fitting and estimated from the data.

discussed in this section.

In the regression analysis one should always check that the residual variances are constant across the cases. This is the reason why we transformed the data with the equalization transform before fitting. If the target function in equalization transformation is Gaussian, resulting data are normally distributed with identical variance. Hence, the polynomial fitting procedure described in this section is justified. If logarithmic-transformation is used instead of the equalization transformation, one needs to carefully check that regression analysis is applicable and possibly additional data processing techniques are required.

5 Stepwise gene selection

After computing T-values as described in Section 4, each gene in each sample is characterized by a single quantity that reflects changes of the expressions across the cases. The next step is to find a set of genes that separates classes A and B. Our approach to identify such a set of genes is to utilize stepwise gene selection procedure. A similar kind of stepwise gene selection algorithm was used in tumor classification in [13] and was found to be both accurate and practical approach to classify tumors.

Let $D \in \mathbb{R}^{p \times N}$ be a data matrix containing T-values, where p is the number of genes and N is the number of the samples. Further, let I_A and I_B

IN: D, I_A, I_B, S.

Set $K \neq \emptyset$, deletion index $k = 0$ and allocate kernel data matrix (D^K).

If initialization is needed, set $K = \{g_I\}$, where g_I is set of genes chosen for initialization. Further, set D^K to be rows corresponding to g_I in D, delete the rows from D, and set $k = c(K)$, where $c(K)$ is cardinality of the set K.

1. Add gene $g_i, i = 1, ..., p - k$ to kernel: $K_i = \{K \cup g_i\}$, where g_i is i:th gene in D. Compute weight (W_i) using I_A, I_B, D^K and a row in D that correspond to g_i.
2. Let g_w be a gene having the biggest W_i. Set $K = \{K \cup g_w\}$, set $k = k + 1$, store row corresponding to g_w in D to D^K and delete row corresponding to g_w from D.
3. Repeat steps 2-3 until stopping criterion S is reached.
4. Compute statistical significance to genes in K.

Table 1: Pseudo-code for the stepwise gene selection procedure.

be indices to samples (columns of D) belonging to classes A and B, respectively. In short, our approach is to add one gene to a set of discriminating genes, referred to as *kernel*, until stopping criterion (S) is reached. The pseudo-code for the method is given in Table 1.

5.1 Initialization

Sometimes it is necessary to initialize the kernel before executing the stepwise algorithm. This is especially the case when the number of samples in classes is small and standard deviation is used in weight computation. Initialization, if needed, is guided by the purpose of the experiment. For example, kernel could be initialized with a gene having greatest distance between mean of samples belonging to class A and mean of samples belonging to class B.

5.2 Weight computation

The core of the stepwise procedure is to compute a weight for each gene. There are several ways to compute the weight. In our case study we have utilized the Fisher linear discriminant [5]:

$$W_i = \frac{(m_A - m_B)^2}{(s_A^2 + s_B^2)}, \tag{5}$$

where m_A and m_B are sample means and s_A and s_B standard deviations for classes A and B, respectively. Remember that in weight computation the data matrix contains as many rows as there are genes in the K_i and as many columns as there are samples. In weight computation these data must be further separate to two classes by using I_A, I_B. For example, if there are three samples in both classes (indices to columns are $I_A = [1\ 2\ 3]$, $I_B = [4\ 5\ 6]$), and the kernel contains two genes, the kernel data matrix could be

$$D^K = \begin{bmatrix} 5.1 & 4.3 & 3.3 & -3.3 & -2.3 & -5.4 \\ 4.4 & 4.6 & 3.2 & -1.2 & -1.0 & -1.1 \end{bmatrix}.$$

The values in a class can be considered as a separate realizations from the class. Accordingly, we can compute means and standard deviations for the class. In this example $m_A = 4.15, m_B = -2.38, s_A = 0.75, s_B = 1.73$ and $W = 12.04$.

In the next round a gene is added to the kernel and the data from which the weight is computed is as follows.

$$D^K = \begin{bmatrix} 5.1 & 4.3 & 3.3 & -3.3 & -2.3 & -5.4 \\ 4.4 & 4.6 & 3.2 & -1.2 & -1.0 & -1.1 \\ 3.6 & 3.9 & 3.4 & -0.5 & -0.3 & -1.0 \end{bmatrix}.$$

giving new weight $W = 10.64$.

5.3 Stopping criteria and statistical significance

As in all stepwise algorithms, it may be sometimes difficult to define an appropriate stopping criterion. One approach is to observe how the weight is decaying. Usually in the beginning of a stepwise procedure, the magnitude of weight decays exponentially, but after some iterations the decay is linear rather than exponential. The point where decay is turning to linear could be a stopping criterion. Another way to obtain a stopping criterion from the weights is to observe a difference between two consequent weights and if the difference is small enough, stop the iteration.

A third way to find a stopping criterion is by computing the statistical significance of the discriminating set of genes. As adding the genes one by one to the kernel reduces the magnitude of the weight, there is a point where it should be also possible to choose randomly a set of genes that separates the classes as well as the genes in the kernel. For example, one could choose $p < 0.05$ and stop the iteration when the kernel contains genes for which $p > 0.05$. The benefit of this approach is that the statistical significance must be computed anyway, and finding a stopping criterion by using p-value gives statistical significance as a by-product of the algorithm. The drawback is that usually the computation of the statistical significance tends to be computationally very expensive.

6 Case study

We have applied the strategy discussed in this paper to evaluate the effect of Herceptin treatment on breast cancer cells. Herceptin is a humanized antibody that targets the protein coded by the *v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2 gene* (*ERBB2*). *ERBB2* is found to be amplified and overexpressed in around 25% of all breast cancers and Herceptin has been shown to be effective only in tumors with *ERBB2* overexpression.[10]

The multitude of cellular and molecular effects caused by the Herceptin treatment are largely unknown and therefore the evaluation of the consequences of Herceptin treatment on gene expression patterns in breast cancer cells is very important. The goal of the drug treatment experiment was to identify a set of genes that separates the samples having amplification/overexpression of *ERBB2* from those with no amplification. Manuscript concentrating on the biological implications of this study is in progress [6]. In this section we explain how the stepwise procedure was applied to the data and briefly discuss the preliminary results.

The data set contained originally six breast cancer cell lines, from which three had *ERBB2* amplification (BT474, SKBR3, ZR7530), two did not have amplification (MDA436, MCF7) and one had low-level amplification (MDA453). Each cell line was exposed to Herceptin and samples were obtained after one, two, and three days of treatment. Treated samples were subjected to a microarray analysis using an untreated sample as a reference. The microarray used contained 14,380 cDNA clones. Materials and protocols are explained in greater details in [6].

The data were preprocessed as follows. First, we treated all ratios, whose intensity in the treated or untreated channels were smaller than 100 fluorescent units or whose area was less than 50 pixels, as missing values. After filtering, the data set consisted of 11,311 genes. These data were then normalized using LOcal WEighted Scatter plot Smoother (LOWESS) [2] for each print-tip group as described in [14]. After calibration we applied the equalization transformation and computed the T-values using Equation 4.

We chose BT474, SKBR3, ZR7530 to comprehend group A, while MDA436, MCF7 were assigned to group B. MDA453 was excluded from the stepwise procedure because it does not fit into either categories. Before applying the stepwise procedure we transformed the T-values again with the equalization transformation because the minima and the maxima of the T-values were

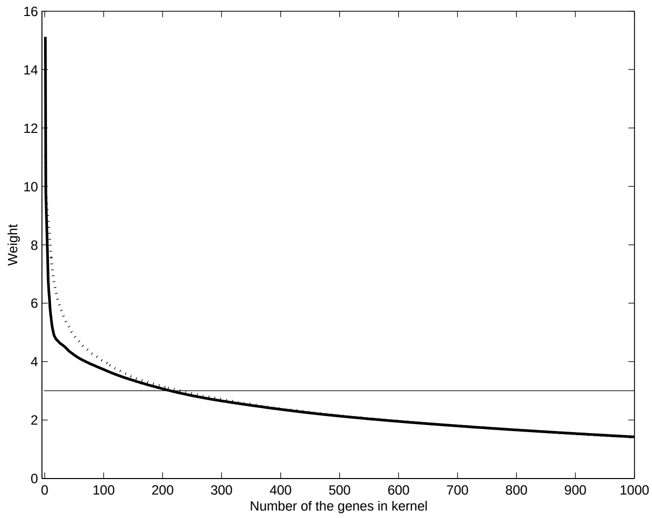

Figure 4: An illustration of the weight decay. Dotted curve denotes genes whose expression levels are increasing in A and decreasing in B. Solid curve denotes genes whose expression levels are decreasing in A and increasing in B. Horizontal line denotes stopping criterion $W = 3$.

not the same across the samples. Further, for each gene we computed a median across the T-values in class A and B, denoted by med_A and med_B, respectively. Then we discarded all genes for which $|med_A - med_B| < 1$, leaving 2990 genes to the stepwise analysis.

As we were interested in both genes whose expression values increase in group A but decrease in group B and genes whose expression values decrease in A but increase in B, we ran the stepwise algorithm two times. First, we searched for genes that are increasing in A but decreasing in B. Thus, we initialized the kernel by a gene whose T-values were positive in A and mean was the furthest away from the mean computed using B. Second, we searched for genes that are decreasing in A but increasing in B and initialized the kernel to contain a gene for which T-values were positive in B and mean was the furthest away from the mean computed using A.

The stopping criterion, $W = 3$, was chosen so that the weight decay was not exponential but linear as illustrated in Figure 4.

The stepwise procedure resulted in two set of genes (results shown in [6]); the first set contained 229 genes (increasing expressions in A and decreasing in B) and the second set contained 213 genes (decreasing expressions in A and increasing in B). Then we computed a statistical significance for these sets of genes as follows. First, we initialized the kernel by a gene whose T-values were positive in A and mean was the furthest away from the mean computed using B. Second, we randomly chose 228 genes from the set of 2989 genes (initialization gene excluded) and compared the weight to the one obtained from stepwise method. This procedure was repeated one million times and applied to the set of 213 genes in an identical fashion. We obtained $p < 10^{-9}$ for both sets.

These two sets of genes included several genes involved in apoptosis, cell proliferation, transcriptional control, and RNA/protein synthesis, implicating that these key cellular events are affected by Herceptin treatment. An in-depth analysis of the data and discussion on the possible relevance of these genes in breast cancer is given in [6].

7 Discussion

In this study we have described a strategy that is designed to identify a set of discriminating genes between two classes in drug treatment experiments. However, many drug treatment experiments contain more than two classes. The strategy described in this paper can also be modified for multivariate discrimination. The difficulty in multivariate discrimination is that each gene should be assigned a weight that is scalar, which may be problematic to compute in a robust manner if the number of classes is large. However, the extension of the Fisher linear discriminant to multivariate one is straightforward [5] and multivariate standard deviations can be converted to scalar ones by computing generalized or total variance. In-depth discussion on the multivariate extension is out of the scope of this study but would make a very interesting topic for further study.

In practical application, the genes whose expression levels either increase or decrease across the cases are of interest. The regression analysis based method to compute T-values is a feasible way to compress the time-series data to a single number that reflects the magnitude and direction of the expression change. Combining the slope and fitting error also allows a sample to have missing cases, providing that the number of cases is sufficiently large. As we have utilized regression analysis, the data must have constant variance across the cases. This requirement is fulfilled by the use of the equalization transformation. However, when utilizing the equalization transformation it is very important to utilize a quality filtering for discarding unreliable genes prior to transformation because after transformation it is impossible to determine whether a gene is an outlier or not.

After computing the T-values for each sample, we utilized a stepwise algorithm for identifying a set of discriminating genes. The stepwise algorithm is flexible and thus it can be applied to several types of microarray studies. For example, instead of the Fisher linear discriminant one may utilize so called noise-to-statistic measure [4], which has been used extensively in cancer classification studies (for a detailed list, see [9]).

In order to demonstrate the utility of the suggested strategy, we have applied it to a breast cancer drug treatment study. In-depth analysis of the results is beyond the scope of this study and will be reported elsewhere [6]. However, preliminary analysis of the results from the case study suggest that the suggested strategy is applicable to real-life drug treatment experiments where the number of samples and cases can be rather small.

References

[1] Bolstad, B., Irizarry, R., Astrand, M., Speed, T. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19: 185–193, 2003.

[2] Cleveland, W. Robust locally weighted regression and smothing scatterplots. *J. of the American Statistical Association*, 74: 829–836, 1979.

[3] Draper, N., Smith, H. *Applied regression analysis.* 3rd edition, John Wiley & Sons 1998.

[4] Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., Lander, E. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286: 531–537, 1999.

[5] Johnson, R., Wichern, D. *Applied multivariate statistical analysis.* 4th edition, Prentice-Hall, 1998.

[6] Kauraniemi, P., Hautaniemi, S., Autio, R., Astola, J., Monni, O., Elkahloun, A., Kallioniemi, A. Effects of a recombinant humanized anti-HER2 antibody (Herceptin) treatment in HER2 amplified and non-amplified breast cancer cell lines. *Manuscript in progress.*

[7] Mousses, S., Wagner, U., Chen, Y., Kim, J., Bubendorf, L., Bittner, M., Pretlow, Y., Elkahloun, A., Trepel, J., Kallioniemi, OP. Failure of hormone therapy in prostate cancer involves systematic restoration of androgen responsive genes and activation of rapamycin sensitive signaling. *Oncogene*, 20(46): 6718–6723, 2001.

[8] Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J., Poggio, T., Gerald, W., Loda, M., Lander, E. and Golub, T. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A*, 98(26): 15149-15154, 2001.

[9] Ringnér, M., Peterson, C., Khan, J. Analyzing array data using supervised methods. *Pharmacogenomics*, 3: 403–415, 2002.

[10] Slamon, D., Godolphin, W., Jones, L., Holt, J., Wong, S., Keith, D., Levin, W., Stuart, S., Udove, J., Ullrich, A., Press, M. Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. *Science*, 244: 707–712, 1989.

[11] van de Vijver, M., He, Y., van't Veer, L., Dai, H., Hart, A., Voskuil, D., Schreiber, G., Peterse, J., Roberts, C., Marton, M., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E., Friend, S., Bernards, R. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*, 347(25): 1999–2009, 2002.

[12] West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson,J. Jr, Marks, J., Nevins, J. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci U S A*, 98: 11462-11457, 2001.

[13] Xiong, M., Li, W., Zhao, J., Jin, L., Boerwinkle, E. Feature (gene) selection in gene expression-based tumor classification. *Mol. Genet. Metab.* 73: 239–247, 2001.

[14] Yang, Y., Dudoit, S., Luu, P., Speed, T. Normalization for cDNA microarray data. *Proceedings of SPIE BiOS*, 2001.