Statistical methods for selecting differentially expressed genes from microarray experiments

Margus Lukk
Margus.Lukk@cs.helsinki.fi

Tiedon louhinta biomolekyyliaineistoista
Helsingin yliopisto, tietojenkäsittelytieteen laitos
Raportti C–2003–52, s. 111–121, marraskuu 2003

Abstract

DNA microarrays provide fast and systematic way to detect genes with cell type specific expression. One of the crucial steps in the analysis of expression data is to identify differentially expressed genes in two or more user defined groups of microarrays. Past couple of years have resulted significant number of techniques that deal this problem. The fact that many of the methods in literature are used only in particular publication has raised lots of confusion. Researchers new to subject often don't know which method to prefer and what these methods actually do with the data.

The purpose of this review is to familiarize the reader with the problem and process of finding good class separating genes between groups of microarrays. Some of the most prominent statistical methods, their strengths and weaknesses will be discussed in detail.
KEYWORDS: microarrays, class separation, between group analysis, t-test, Wilcoxon rank sum, ANOVA, permutation test, S2N, Bonferroni adjustment.

1 Introduction to microarrays

The ultimate goal of biomedical researchers has been obtaining full knowledge about life, thus about living cells. Although many of the cell components and their ways of functioning are well studied, the general picture how all the pieces fit together is still a mystery. In the middle of last decade good ground for learning more about gene regulation was established by introduction of microarray technology. The central dogma of molecular biology says that DNA makes RNA makes protein. Or, in other words, the whole information needed for "running" a cell is stored in DNA which acts like central warehouse of information, pieces of that information are "written out" to mRNAs, which are then used as manuals to manufacture proteins. Proteins themselves are components that hold up a cell structure, catalyze different chemical reactions, transport signals between cell parts as well as between cells, maintain DNA and switch certain parts of it on and off etc. The microarray technology makes it possible to measure the expression of all mRNAs in a cell in particular time point. Knowing the the presence or absence of each mRNA thus shows which proteins are in production and what processes in the cell are active.

In 1975 Edwin Southern introduced a method, today known as Southern blot, for detecting presence of DNA probe in unknown DNA sample[E92]. The technique is based on

5' ATAGGAGTCGCTGATG 3'
3' TATCCTCAGCGACTAC 5'

Figure 1: Two complementar DNA chains in DNA double helix. Two DNA chains are complementar if positions of A and C nucleotides in one of the chains correspond to T and G in the other.

the property that single stranded DNA pairs always with other single stranded DNA, complementar to it, to form a DNA double helix (see Figure 1.). In Southern blot, radioactive labeled sample DNA is hybridized to the probe fixed on the surface (e.g on nylon filter). Hybridization reaction is a process where two complementary strands of nucleic acids are joint to double stranded helix. Reaction is carried out on higher temperature where original double stranded probe and sample take single stranded form. Hybridization reaction will take place only if the radioactively labeled sample contains same DNA as probe. In case of success the sample will stick on the probe and will be later detectable, as non hybridized radioactive DNA from the sample will be washed away. The higher the concentration of probe DNA in the sample, the stronger the radioactive signal on the spot on the filter where the DNA chains of the probe are fixed.

On microarrays, one performs thousands of Southern blot kind of reactions at once. Big number of known DNA fragments corresponding to mRNA sequences of particular genes are fixed on the miniature surface (e.g. 1cm x 1cm) and hybridized with labeled mRNA extract from the cells. The surface material is commonly glass or plastic and the sample is marked with fluorescent labels instead of radioactive ones. The rest of the section will be organized as follows. Subsections 1 and 2 describe two main types of microarrays, cDNA microarrays and oligonucleotide arrays respectively, and subsection 3 will introduce formats of raw microarray data.

1.1 cDNA microarrays

cDNA microarrays have got their name from the molecules fixed on their surface. cDNA is DNA synthesized by reverse transcriptase from an mRNA template. For the hybridization, two pools of mRNAs samples (test and control) are prepared from two different kind of cells and labeled with different fluorescent dyes. Test and control mRNAs or cDNAs are hybridized to the microarray at the same time making them to compete for the same probes on the surface of microarray. If the amount of the mRNA molecule in one of the samples is higher then also the intensity of particular fluorescent dye in the spot where the probe is fixed is stronger. When the concentrations are equal or mRNA molecules corresponding the probe are missing then so is it with the intensities of fluorescent dyes. cDNA microarrays are thus used for capturing the relative changes of each mRNA molecule in transcriptome (see Figure 2a).

1.2 Oligonucleotide microarrays

The main weakness of cDNA microarrays comes from relatively long length of the probes. On 1953 Watson and Cric discovered that two complementary chains in DNA double helix are hold together by the weak hydrogen bonds. Hydrogen bonds can be formed between

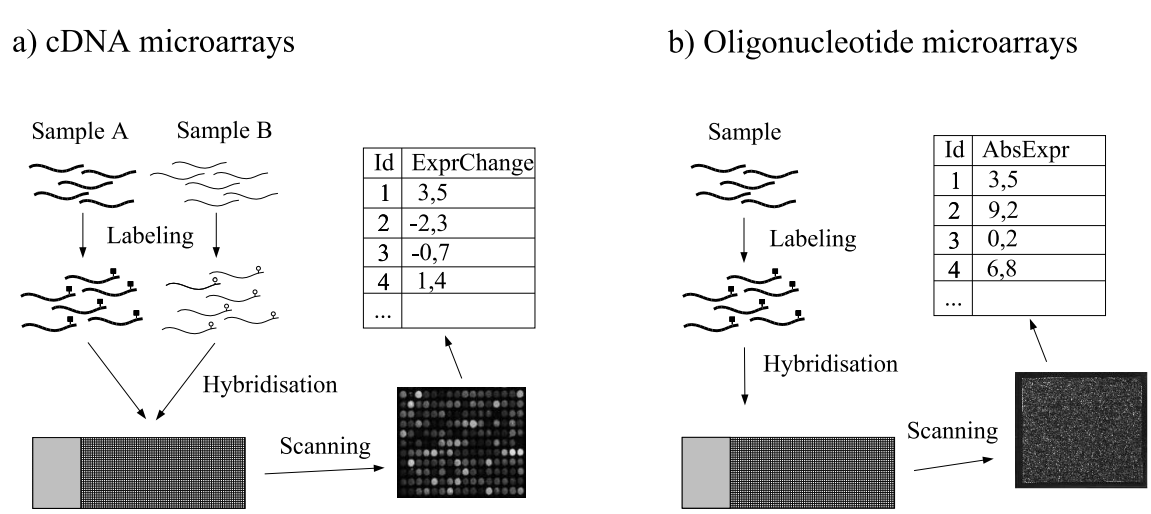

Figure 2: Simplified microarray experiments. **a)** cDNA microarray: sample cDNAs are labeled and hybridized to the chip. Microarray is then scanned and pseudo-colour image from two different dye intensities is constructed. Finally dots in the image are converted to gene names and numerical expression changes depending on their location and color on the image. **b)** Oligonucleotide microarray: sample cDNA is prepared and hybridized to the chip. Chip is scanned and dots are converted to gene names and their absolute expression values depending on their location and color in image.

nucleotides of Adenine (A) and Thymine (T), or Cytosine (C) and Guanine (G) (see again Figure 1). Longer DNA double helix means thus more hydrogen bonds, which means greater force to keep two stands of DNA together. Starting from some length of the DNA, the force between chains becomes so strong that the helix stays together even if double chain contains some nucleotide pairs that do not form hydrogen bonds. On the other words, with the length of the probes rises also the probability for unspecific hybridizations i.e. for hybridizations where all pairs of nucleotides are not complementar. It can thus easily happen that cDNA probe on the chip gives signal even when the exact copy of it in the sample is not present.

In oligonucleotide microarrays the specificity of hybridizations is much higher due to shorter (e.g. 25 bases) probes on the chip. In addition, the usage of several unique probes corresponding to one mRNA allows to compute mean expression which value is hoped to reflect better the true expression. The higher specificity in one hand and the use of only one sample instead of two in hybridization reaction allows to measure absolute expression of every mRNA (or cDNA or cRNA) in the sample on the scale from non-expressed to highly expressed (Figure 2b). The greatest disadvantage for using oligonucleotide arrays is their relatively high price.

1.3 Microarray data formats

Raw expression data has different formats depending on the software which was used to scan the array after hybridization. The data from the array is often stored in text tabulated format where each row contains data about one gene (gene, clone, mRNA, cDNA are used henceforth synonymously) on the microarray. Each row has an identifier (e.g. the name of

GeneId	Array 1	Array 2	Array 3	Array 4	Array 5	Array 6	Array 7	...
203629at	67,5	56,7	83,7	74,8	63,6	49,1	64,8	...
203630at	168	165,9	148,6	134,4	151,4	212,7	199,9	...
203632at	130,9	137,2	82,8	96,9	145,5	216,9	163,6	...
203633at	202,5	218,8	211,6	261,7	328,9	156,1	186,6	...
203635at	92,8	114,9	95,8	91,6	122,3	84,2	65,8	...
203640at	204,6	202,8	239,1	193,2	208,6	438,6	530,5	...
203641at	102,9	85,9	94,2	85,3	137,1	81,5	78,1	...
203642at	87	67	105	76,5	141,2	71,7	109,2	...
203645at	244,5	278,7	829,1	293,7	500,2	241	225,9	...
203647at	267,9	158,8	136,5	199,9	344,1	176,3	140,8	...
203648at	237,6	244,8	241,4	249,6	367,6	192,5	167,4	...
203656at	121,2	151,6	119,8	105,5	247,9	97,4	102,3	...
...	...	...	...	...	...	...	...	...

Figure 3: Example of expression data matrix. Clone 203640at expression value in array 4 is 193,2.

the clone), expression value of that clone, annotation of the clone and some other parameters which may or may not have significance. However, the most important columns in the data file are the ones of gene identifier and its expression value (which sometimes may be log transformed). For further analysis the data from different arrays is usually merged to one matrix of expression values where columns represent arrays and rows correspond to genes (Figure 3).

2 Fashion of profiles

In recent years prices of high resolution microarrays have dropped sufficiently to allow researchers to perform tens of hybridizations in one study. As the living cells are never exactly the same the repeated measurements simply help to catch normal biological variance of gene expression. Cancer cells may vary even in a scale where traditional diagnostic techniques fail to catch their exact type and status, whereas expression values for all genes maintain this power. Lately several new cancer subtypes have been discovered simply by hierarchical clustering of the expression data [Ram01][Bha01][van02]. Even though arrays allow extreme diagnostic accuracy they are far too expensive to become part of routine testing machineries in the hospitals. Often measuring the expression levels of only few genes may provide same answers to our questions with same degree of accuracy as testing them all on the array. The question is simply which genes carry the information we are interested in. In last couple of years several methods have been proposed for finding such genes, genes which expression allows to differentiate between cell types or even same cells under specific conditions. In practice this means several repeated expression measurements in all groups of cells and search of genes which expression differs between groups of microarrays while it is stable inside the groups. In addition to the diagnostic value that found genes may have, they are also interesting as genes that are most probably behind the status of the cells.

The focus of this review is on the statistical methods used for finding group differenciating genes from microarray data. The rest of the review is divided as follwos: section

3 discusses common statistical methods used for comparing groups of expression data and picking up informative genes, section 4 describes the usage of permutation test for computing the statistical strength of selected genes, section 5 gives glimpse of other statistical and non-statistical methods not covered here, section 6 discusses some weakensses comming from microarray technology that affect the analysis procedure, finally section 7 summarizes the whole review.

3 Group separation with common statistical methods

The crucial problem in building profiles for groups of expression data is to find genes that best separate particular dataset from some other group of expression data. Good separator genes should have as different expression between groups as possible. For that we need a distance measure that could show how far two groups are from each other. The most primitive one could be the absolute distance between the expression means of a gene in the groups. The bigger the difference the better the separation. However, the measure gets seriously biased if the groups are not homogeneous i.e. contain extreme outliers. Instead, the distance between two nearest elements in two groups could be used. Yet another solution is to correct the biased mean difference with the variances of the groups. There are several test statistics that combine in them means and variances of groups.

3.1 t-test

Two-sampled t-test is currently the most used test for selecting genes that can separate two groups of microarrays. The test statistic is computed as

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \tag{1}$$

where μ_1, μ_2 are mean values and σ_1, $sigma_2$ correspond to standard deviations of two samples. The main weakness of particular distance measure is the assumption of equal variances in both samples, which may not always be the case. Slight modifications in the denominator, namely addition of n_1 and n_2 that correspond to the number of elements in groups one and two, give the formula of two sampled t-test for unequal variances [GS01]

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}. \tag{2}$$

Regardless of its wide use, t-test has weaknesses which are often not paid enough attention. First, t-test assumes the normality of the data i.e. that the data inside the groups follows normal distribution. Often the number of arrays is too small for such assumption. Another problem of t-test as well as most of other tests reviewed henceforth comes from multiple testing. Let it be reminded that each gene from the data is tested separately and for some studies the total number of tests may easily exceed 10000. Usually statistical tests are carried out on significance level $\alpha = 0.05$ which shows the probability of getting sample as extreme (or worse) just by chance. Thus the probability of extreme results and also false positives (so called Type I errors) starts growing with the number of tests. The intuitive

Value	Group	Rank
2.3	A	1
2.9	A	2
3.2	A	3
3.4	A	4
3.6	A	5.5
3.6	B	5.5
3.9	B	7
4.1	B	8
4.2	B	9
4.9	A	10
5.3	B	11
5.7	B	12

Figure 4: Wilcoxon rank test wight groups A={2.3, 3.4, 3.2, 4.9, 2.9, 3.6}, B={4.1, 3.6, 5.7, 4.2, 3.9, 5.3}. The test statistic $R = 25.5$ (as $25.5 < 52.5$). Given $n_1 = n_2 = 6$ and $\alpha = 0.05$ the critical value for two-tailed test is 26. Thus the groups are significantly different on 0.05 significance level.

solution would be to correct α level with the number of tests, feasible by division of α with the number of tested hypothesis. In α/n the type I error rate for n tests stays the same as provided by α while testing only 1 hypothesis. Such way of correcting α is known as **Bonferroni method** [OJ02]. α correction with more complicated **Turkey's method** would give the same effect[Ric94].

3.2 Wilcoxon rank sum test

Genes we want to test may not always have underlying normal distribution, often due to small number of measurements in the groups. An alternative test without any distributional assumptions (non-parametric tests) is therefore needed. Common non-parametric alternative for t-test is Wilcoxon rank sum test. Many resources refer to Wilcoxon rank sum test and Mann-Whintey test. In fact they are almost the same, developed around the same time and follow the same logic and principles. Even the results from these tests are comparable.

To do the Wilcoxon rank sum test, the variables from both samples get merged and ordered by their values. Each variable obtains rank depending on its position in the list. For example value in position 10 will get rank 10. If two or more variables have equal value their rank will be mean of positions they share, e.g. if values on positions 11, 12 and 13 have same value, their rank will be 12 as $12 = (11 + 12 + 13)/3$ (Figure 2). Next, the sum of ranks is computed for both groups and smaller from the values is selected as a test statistic. Critical value for the test statistic given α is found from the table or with statistical software given the sample sizes, n_1 n_2. If the value of the test statistic is more extreme than critical value, two groups are said to have signifficant separation. Due to multiple testing Bonferroni adjustment is compulsory[Tro02].[Ric94]

3.3 ANOVA

In most of the cases only two groups of expression data are compared, however the number of groups may be bigger. For example, researchers might be interested in finding separator genes between samples of normal, tumorigenic and metastatic tissues. If such genes exist, their expression values can in theory be used for cancer diagnostics or even more important, they might have role in development and progression of cancer. The strength of each gene as a separator in more than 2 samples is tested by one-way analysis of variance (ANOVA). Non-parametric alternative to one-way ANOVA is Kruskal-Walls test. But due to its seldom usage on microarray data it will be not covered here. ANOVA can be seen as extension of t-test for more than 2 groups of samples, or vice versa, the t-test can be thought of as generalization of ANOVA for only two groups of samples. Similarily to t-test, also the test statistic of ANOVA, conventionally known as F, is ratio between values that measure variance between the groups (MS_{effect}) and within the groups (MS_{error}).

$$F = \frac{MS_{effect}}{MS_{error}} \tag{3}$$

Lets first focus on the numerator and denote the number of classes with letter G. In t-test, $G = 2$ and the distance between two means is subtraction of one from the other. In ANOVA $G > 2$, so there exist more than one distance and these have to get combined somehow to one value. For that the mean of all groups M_{tot} is computed and the the sum of the distances between M_{tot} and all group means are summed in the form of "sum of squared deviates", i.e. $\sum_{g=1}^{G}(M_g - M_{tot})^2$. As the groups may contain different number of elements the actual sum of squared deviates (SS_{bg}) in ANOVA has slightly modified form

$$SS_{bg} = \sum_{g=1}^{G} N_g(M_g - M_{tot})^2 \tag{4}$$

where N_g represents the number of values in group g. N_g can be seen as a weight that corrects the differences in groups sizes.

The denominator for F is the sum of variances inside the groups just like in t-test.

$$SS_{wg} = \sum_{g=1}^{G}\sum_{i=1}^{N_g}(X_{gi} - M_g)^2 \tag{5}$$

where X_{gi} is variable i in group g.

SS_{bg} and SS_{wg} look very similar as they both are measures of variance. One of them captures the variance between centers of the groups, and the other variance inside the groups. However these two measures can not be used together in one formula if the number of number of items, from which they are built of, is different. For SS_{bg} the number of items is G while for SS_{wg} it's $sum_{g=1}^{G}N_g$, and corresponding degrees of freedom are thus $df_{bg} = (G - 1)$ and $df_{wg} = sum_{g=1}^{G}(N_g - 1)$. Due to the differences in degrees of freedom the final equation for F takes form of

$$F = \frac{SS_{bg}/df_{bg}}{SS_{wg}/df_{wg}} \tag{6}$$

Perhaps the main disadvantage of using ANOVA is that it does not tell which groups are different from which groups, it just says if there is an overall difference. The test gives positive result even if some of the groups overlap completely while some other groups are

clearly separated. The relations between individual groups can be investigated only by pairwise comparisons. However it is not meaningful to start with pairwise tests as the pairwise tests are unlikely to be sensible if the result from ANOVA is non-significant.[R03]

4 Permutation test

In 1999, a group from MIT published paper about acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL)[Gol99]. The article contained not only the largest set of leukemia expression data at the time but also several new methods for analyzing such data. Among other novelties was usage of permutation test with distance measure similar to t-test statistic.

The distance measure for gene between groups groups, referred as "signal to noise" ratio has form of

$$S2N = \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2} \tag{7}$$

where μ_1, μ_2 and σ_1, σ_2 refer to expression means and standard deviations in groups. The measure looks like robust form of t-test statistic and has advantage in cases where the variance in one of the groups may be higher than in the other. Also S2N value may be positive or negative depending in which group the gene has higher expression.

The statistical significance of S2N scores were evaluated with permutation test. In permutation test, each microarray experiment gets class label depending on the group it belonged to. The class labels are then switched randomly many times (permutated) and after each permutation new S2N scores for each gene is computed. Genes are ranked according to their positions in sorted list of S2N values (e.g. gene with highest S2N score gets rank 1). Histograms of S2N scores for each rank are built, e.g histogram for all genes that obtained rank 1 over all permutations, histogram for genes with rank 2 etc. After permutations, S2N value of a gene computed under correct class labels is compared with the value of 1% percentile in the histogram that has the same rank. If the value is more extreme than in histogram, the gene is considered to have significant power to separate two groups or classes.

4.1 SAM

Several slightly modified variants of the original method of Golub and Slonim et al.[Gol99] have been recently published. Due to the limited space only Significance method of Analysis of Microarrays (SAM)[Tus01] and "ideal discriminator method" [Tro02] were selected to be referred here while other methods, including Mixture Model Method by Pan et al [Pan03], are left out.

Differences between SAM and other permutation test based methods lie mainly in the distance measure and how the permutations are used. The distance measure in SAM is

$$d(i) = \frac{x_1(i) - x_2(i)}{s(i) + s_0} \tag{8}$$

where $x_1(i)$ and $x_2(i)$ are average expressions of gene (i) in groups 1 and 2. The "gene-specific scatter" $s(i)$ is

$$s(i) = \sqrt{a * \{\sum_m [x_m(i) - x_1(i)]^2 + \sum_n [x_n(i) - x_2(i)]^2\}} \tag{9}$$

where $\sum_m$ and $\sum_n$ are sums of expression measurements in groups 1 and 2 while $a = (1/n_1 + 1/n_2)/(n_1 + n_2 - 2)$, $n1$ and $n2$ are the numbers of measurements in corresponding groups. s_0 is small positive constant computed from the dataset. The list of $d_p(i)$ values from each permutation test is received, $d_p(1)$ has the highest value $d_p(2)$ second highest etc. Expected relative difference $d_E(i)$ for each i is average value of $d_p(i)$ over all permutations i.e. $d_E(i) = \sum_p d_p(i)/b$ where b is the number of permutations performed for computation. In most of the cases $d(i) \cong d_E(i)$. Genes, which have expected and observed relative difference bigger than δ are called significant. The bigger the δ the smaller the number of genes and also the number of false positives[Tus01][SAM].

4.2 Ideal discriminator method

The distance measure for "ideal discriminator method" is Pearson correlation coefficient [Tro02]. Ideal discriminator is a gene with artificial expression values so that in one of the groups expression is minimum and in the other its maximum. Pearson correlation between ideal gene and all other genes are computed. Again class labels are permuted and p-value of each gene is computed as

$$p_{j_{perm}} = \frac{count(max_i(p_{i_{perm}}) > p_{j_{obs}})}{count(permutations)} \tag{10}$$

The method is significant mainly because of the different thinking its built on. In t-test and in its relatives the thinking is horizontal, from left to right, and expression values of one gene are considered at the time. In ideal discriminator method the thinking is vertical. Expression values of ideal gene are compared with expression values of all other genes. The operation unit is not anymore single value but vector.

4.3 Turning parametric test to non-parametric

So far we have seen how two different distance measures somewhat similar to t-test have been used in permutation test. This however does not mean that t-test or any other parametric test statistic could not be used in permutation test framework. The core idea of permutation test is to compare the test statistic with the same data after randomization. Therefore no information about the distribution of the data are used neither needed. Thus permutation test can be used to change parametric test to non-parametric [Tro02]. The p-value for gene j is then

$$P_j = \frac{count(t_{j_{perm}} > t_{j_{obs}})}{count(permutations)} \tag{11}$$

where $t_{j_{obs}}$ are t values from the data using original class labels and $t_{j_{perm}}$ are t values for the same gene after every permutation. Still p-values have to be corrected for multiple testing by using Bonferroni correction: $P_{j_{Bonferroni}} = min(m * p_j, 1)$ where m=number of genes [Tro02].

5 Glimpse of other methods

Last two sections discussed only most common statistical methods for group comparison. Many other statistical as well as non statistical methods have been developed from which some will be mentioned in following sentences. Methods published in statistical framework

include linear discriminant analysis, regression modeling, between group analysis (BGA) and Bayesian statistics. Machine learning is applied in maximum difference subset (MDSS) algorithm, prediction by collective likelihoods (PCL). Also principal component analysis (PCA) has been used for finding best genes for group separation.

6 Discussion

Regardless of all the methods referred and many other non-statistical approaches for reducing dimensionality to end up with small number informative genes, there seems to be no well established technique that would fully satisfy biological researchers as well as statisticians. One of the reasons for growing number of different methods comes from the specialty of the data. The number of samples in groups is often still too small for proper parametric tests. At the same time the number of experiments is ill-proportionally high. From the biomedical scientist point of view the problem is lack of knowledge of what is proper statistics. There are also serious limitations in availability of tools. Statistical packages need prior knowledge for working with them and implementations used in method publications are often not available. Different methods produce lists of genes that vary in length and may not overlap etc. However, majority of the methods still capture most from the relevant information.

The only exception where statistical tests for group separation may completely fail is finding good diagnostic marker genes. Good marker gene is here defined as one expressed only in one of the two groups. The main reason why methods described here and also elsewhere miss such markers are technical weaknesses that microarrays still have. Namely, even state of the art array technology fails to draw border between expressed and unexpressed genes. For example, on oligonucleotide arrays a gene may be not expressed even if its expression value is as high as 100 while another gene on the same microarray may be truly present with expression of 30. The scale of expression values may be 0 to 6000 while most of the values fall between 0 and 1000. Typical examples of statistically significant separator genes are 200 and 600, or 40 and 120, or 1800 and 2800. All such genes are interesting for further study but they are completely useless as diagnostic markers. Thus the best strategy for selecting good marker genes seems still to be manual selection of genes from lists produced by any statistical or non statistical method.

7 Summary

The invention of DNA microarrays has opened many new doors for biomedical scientists toward full understanding of life. In recent years the number of experiments where expression of genes in two or more types of cells are measured by using DNA microarrays is increased. Such wet lab experiments have triggered the adaption of old and invention of new statistical and non-statistical methods that can point genes which expression in these groups differs most. This review covered principles, strength and weaknesses of most used and/or peculiar statistical methods that deal this problem.

References

[Bha01] A. Bhattacharjee et al. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *PNAS*, 98:13790, 2001.

[E92] Southern E. Detection of specific sequences among dna fragments separated by gel electrophoresis. reprint of original article from 1975. *Biotechnology*, 24:122–39, 1992.

[Gol99] T. R. Golub et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531, 1999.

[GS01] S. A. Glanz and B. K. Slinker. Applied regression & analysis of variance. *McGraw-Hill Inc.*, second edition, 2001.

[OJ02] A. Olshen and A. N. Jain. Deriving quantitative concusions from microarray expression data. *Bioinformatics*, 18:961–970, 2002.

[Pan03] W. Pan. On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression. *Bioinformatics*, 19:1333–1340, 2003.

[R03] Lowry R. Concepts and applications of inferential statistics. *Interactive textbook at* `http://faculty.vassar.edu/lowry/webtext.html`, 1998-2003.

[Ram01] S Ramaswamy et al. Multiclass cancer diagnosis using tumor gene expression signatures. *PNAS*, 98:15149, 2001.

[Ric94] J. A. Rice. Mathematical statistics and data analysis. *Duxbury Press*, second edition, 1994.

[SAM] SAM. Sam: Signifficance analysis of microarrays. technical manual. `http://www-stat.stanford.edu/~tibs/SAM/`.

[Tro02] O. G. Troyanskaya et al. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, 18:1454–1461, 2002.

[Tus01] V. Tusher et al. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, 98:5116–5121, 2001.

[van02] L. J. van't Veer et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530, 2002.